Current Sleep Medicine Reports (2023) 9:140–151
https://doi.org/10.1007/s40675-023-00253-w

Error Propagation from Sleep Stage Classification to Derived Sleep Parameters in Machine Learning on Data from Wearables

Emil Hardarson[1,2] · Anna Sigridur Islind[1,2] · Erna Sif Arnardottir[2] · María Óskarsdóttir[1,2]

Accepted: 6 April 2023 / Published online: 30 June 2023

Abstract

Purpose of Review: Automatic analysis of sleep is an important and active area of research. Machine learning models are commonly developed to classify time segments into sleep stages. The sleep stages can be used to calculate various sleep parameters, such as sleep efficiency and total sleep time. The machine learning models are typically trained to minimize the sleep stage classification error, but little is known about how error propagates from sleep stages to derived sleep parameters.
Recent findings: We review recently published studies where machine learning was used to classify sleep stages using data from wearable devices. Using classification error statistics from these studies, we perform a Monte Carlo simulation to estimate sleep parameter error in a dataset of 197 hypnograms. This is, to our knowledge, the first attempt at evaluating how robust sleep parameter estimation is to misclassification of sleep stages.
Summary: Our analysis suggests that a machine learning model capable of 90% accurate sleep stage classification (surpassing current state-of-art in wearable sleep tracking) may perform worse than a random guess in estimating some sleep parameters. Our analysis also indicates that sleep stage classification may not be a relevant target variable for machine learning on wearable sleep data and that regression models may be better suited to estimating sleep parameters. Finally, we propose a baseline model to use as a reference for sleep stage estimation accuracy. When applied to a test set, the baseline model predicts 2-, 3-, 4- and 5-class sleep stages with an accuracy of 74%, 54%, 46% and 35%, respectively

Keywords Sleep · Sleep staging · Sleep parameters · Machine learning · Wearables

Introduction

Objective sleep measurements are a critical part of sleep research and diagnosis of sleep disorders [1]. Polysomnography (PSG) is the gold standard of objective sleep measurements, but analysis of PSG requires expert human annotation. In addition to being expensive to analyze, PSG is time-consuming to set up, requires specialized equipment and causes discomfort, making prolonged ambulatory sleep measurements infeasible. To mitigate these limitations, two avenues can be identified in both research and clinical measurements. On one hand, to ease the data collection for a broader part of the population, consumer-grade wearables have become a prominent tool [2], and on the other hand, machine learning models have been developed to automate analysis [3–5].

Machine learning models, and deep learning, in particular, are commonly used to classify sleep stages (sleep staging). Of the 185 Google Scholar article results where both “sleep" and “deep learning" were present in the title, approximately 40% were about sleep staging (similar findings for PubMed, author search October 20th, 2022). Sleep staging is an important step in the analysis and diagnosis of sleep because it establishes the macrostructure of sleep, and several sleep statistics are derived from the sleep stages [6]. These clinically relevant sleep parameters include, for example, sleep efficiency (SE), wake after sleep onset (WASO) and amount

✉ Emil Hardarson
emilh@ru.is

Anna Sigridur Islind
islind@ru.is

Erna Sif Arnardottir
ernasifa@ru.is

María Óskarsdóttir
mariaoskars@ru.is

[1] Department of Computer Science, Reykjavik University, Reykjavik, Iceland

[2] Reykjavik University Sleep Institute, School of Technology, Reykjavik University, Reykjavik, Iceland

of time spent in rapid eye movement (REM) sleep. Machine learning models typically follow the same protocol as human scorers and designate each 30-second non-overlapping time segment (epoch) one sleep stage. This results in a sequence of sleep stages called a hypnogram which is then compared to a hypnogram where an expert human scorer performed the sleep staging, to evaluate the accuracy of the machine learning model. In this paper, we refer to this manually scored hypnogram as the “ground-truth hypnogram".

Most published work describing machine learning models for classifying sleep stages only reports the sleep staging error of the model, with the implicit assumption that this would translate to a lower error in derived sleep parameters. However, it is unclear how robust sleep parameter estimation is to misclassifying sleep stages and to the best of our knowledge, the impact of sleep staging error on the derived sleep parameter error is poorly understood. For example, it is not known how well a machine learning model capable of 92% accurate sleep staging performs when estimating sleep efficiency. This paper focuses on the propagation of error from sleep stages to sleep parameters.

The mapping from a hypnogram to sleep parameters is highly non-linear and depends on multiple interacting factors, not least the structure of the underlying ground-truth hypnogram. Thus, a formulation of error propagation can not be derived analytically. Instead, here, we propose a Monte Carlo error propagation simulation framework tailored to the analysis of sleep data. We simulated how a machine learning model would estimate sleep stages for a dataset of ground-truth hypnograms, given the machine learning model’s confusion matrix.

Uncertainty in sleep staging can be attributed to several diverse sources, with recent developments in quantifying and describing this uncertainty [7, 8]. However, the literature is scarce regarding the error in sleep parameters derived from a hypnogram. Although not commonly applied in sleep science, Monte Carlo simulation is a well-established method for evaluating error propagation in other fields, such as physics, biology, chemistry and geology [9–11]. In this paper we present the results of a Monte Carlo simulation study showing the impact of error propagation relative to the error of sleep staging models.

Wearable sleep trackers, such as wrist-worn smartwatches, have been suggested as an alternative or complement to the resource-heavy PSG to measure sleep, e.g. for screening of sleep disorders [1]. The reason is two-fold: i) wearables provide light-weight measures at a relatively low cost compared to the PSG, and; ii) wearables can continuously capture data over a long time, which enables new types of analysis to be conducted, allowing measurements of so-called free-living sleep [12]. However, there is a gap in the literature regarding what type of data they can capture and how to analyze the data.

Most consumer-grade wearables currently in use are wrist worn [19, 20]. The technology is promising but requires validation, access to the algorithms applied to the aggregated data or access to the raw data. Furthermore, wearable sleep trackers are less accurate than PSG at classifying sleep stages [1, 2, 13–17]. Hence, understanding how staging error propagates to sleep parameters is especially important when working with data from wearable sleep trackers.

With this background, we opted to limit the scope of this simulation study to examine machine learning models which estimate sleep stages from wrist-worn wearables. This requires finding reported confusion matrices and other details of such machine learning models.

Our main contributions are three-fold:

1. A review of machine learning models using data from wrist-worn sleep trackers, demonstrating a gap in the literature when it comes to reporting error statistics.
2. A novel Monte Carlo simulation framework for estimating error propagation from sleep staging error to clinical sleep parameters, which demonstrates that sleep staging error is not representative of the error in derived sleep parameters.
3. Suggestions for future work developing machine learning models for estimating clinically relevant sleep parameters from wearable data.

Background

Previous work has reviewed the use of wearables to detect sleep stages [16, 21, 22] and the use of machine learning models to detect sleep stages [3, 4, 23, 24]. We searched the literature for papers about machine learning models performing sleep staging on data from wrist-worn wearables. Performing a Web of Science search for the query

```
(ALL=(wearable OR wearables OR wrist))
AND (ALL=(stages OR staging OR stage))
AND ALL=(classifier OR machine learning
OR deep learning OR probabilistic OR
logistic regression) AND (TI=sleep)
```

resulted in 39 papers, of which 30 were excluded because they did not fit the following eligibility criteria:

- Does the article present a machine learning model for automatic sleep staging?
- Does the machine learning model use data from a wrist-worn wearable device?
- Is the machine learning model’s accuracy reported?
- Does the paper describe an original machine learning model? Review articles are excluded.

An identical search on PubMed did not offer new results. The search was performed on August 5th, 2022. Table 1 summarizes the findings of the literature search regarding which sensors were used, how many sleep stages were considered, the machine learning model used, the staging accuracy and whether the paper reported the confusion matrix or information about sleep parameters. As can be seen in the table, most studies use photoplethysmography (PPG), acceleration (ACC) or electrodermal activity (EDA); their accuracy ranges from 59% to 93%; five out of nine papers show the confusion matrix and only two report information about sleep parameters. Eight of the nine papers used 30-second epochs; Gashi et al. classified 600-second segments [25]. Only Wulterkens et al. and Fonseca et al. reported the error in clinical sleep parameters [26, 27].

Epochs can be classified into various numbers of classes. Common designations are 2-class (wake or sleep), 3-class (wake, N1/N2/N3 or REM sleep), 4-class (wake, N1/N2, N3 or REM sleep) and 5-class (wake, N1, N2, N3 or REM sleep). When 4-class staging is performed, the N1/N2 class is commonly referred to as *light sleep* and the N3 class as *deep sleep*. Of the nine machine learning models, five performed 4-class sleep staging These five techniques form the basis of our simulation study. All five techniques use data from research devices, as opposed to consumer-grade wearables. Comparing the staging performance of machine learning models is problematic since the accuracy depends on multiple factors, such as the dataset used, which sensors were used, how lights out was defined and what validation method was used. For example, the machine learning model described by Anusha et al. has the highest reported 4-class staging accuracy in Table 1. However, it was validated against a home sleep apnea testing device with no electroencephalography data recorded. [28].

Sleep parameters

The American Academy of Sleep Medicine (AASM) Scoring Manual Version 2.6 lists the following recommended parameters for polysomnography which must be reported for clinical sleep studies [6]:

- *Total recording time (TRT)*. Time in minutes between the first epoch of any sleep and last epoch of any sleep. Since the dataset used in the simulation did not contain information about lights on and lights off, this parameter and those derived from it were defined differently from the AASM definition.
- *Total sleep time (TST)*. Time in minutes spent in light, deep and REM sleep.
- *REM latency*. Time in minutes between sleep onset to the first epoch of REM sleep.
- *Wake after sleep onset (WASO)*. TRT - TST.
- *Sleep efficiency (SE)*. Proportion of time spent sleeping: TST/TRT.
- *Time in minutes in each state*. The four included stages are wake, light, deep and REM slee
- *Proportion in each state*. Percentage of TST in each stage (time in each stage/TST)

The simulation study in this paper uses four sleep stages (wake, light, deep and REM), so time in each state and proportion in each sleep state leads to a total of six parameters.

Table 1 A summary of published papers using machine learning for sleep staging on data from wrist-worn wearables, in order of publicaton date

Reference	Sensors	Sleep Stages	Accuracy (kappa)	ML Method	C[1]	y[2]
Gashi et al. 2022 [25]	EDA, ST, ACC	2	92% (0.55)	GB		
Anusha et al. 2022 [28]	EDA, ST	4	93% (-)	RF	x	
Ko et al. 2022 [29]	HR, ACC	2	85% (-)	CK		
Radha et al. 2021 [18]	PPG	4	76.36% (0.65)	LSTM	x	
Wulterkens et al. 2021 [26]	PPG, ACC	4	76.4% (0.62)	LSTM	x	x
Li et al. 2021 [30]	PPG, ACC	4	69% (0.44)	CNN, SVM	x	
Sundararajan et al. 2021 [31]	ACC	2	79%[3] (0.5)	RF	x	
Walch et al. 2019 [32]	ACC, PPG	3	72.3% (0.27)	MLP		
Fonseca et al. 2017 [27]	PPG	4	59% (0.42)	LDC		x[4]

ACC: Acceleration, C: Confusion matrix, CK: Cole-Kripke, CNN: Convolutional neural network, EDA: Electrodermal activity, GB: Gradient boosting, HR: Heart rate, LDC: Linear discriminant classifier, LSTM: Long-short term memory, ML: Machine learning, MLP: Multi-layered perceptron, PPG: Photoplethysmograpy, RF: Random forest, ST: Skin temperature, SVM: Support vector machine, y: Sleep parameter

[1] Marked ‘x’ if a confusion matrix was reported

[2] Marked ‘x’ if information about estimated sleep parameters was reported

[3] Average precision

[4] Fonseca et al. reported estimated sleep parameters but only using a 2-class (sleep/wake) classifier

In addition to the AASM recommended parameters listed above, the number of awakenings (NA) is defined as the number of times at least two subsequent wake epochs occur after sleep onset [33]. As a result, we consider in total eleven sleep parameters.

Methods

Performance of machine learning sleep stage classifiers is most commonly reported in terms of staging accuracy, Cohen's kappa and sleep stage confusion matrix. Sleep parameter error is not commonly reported and can not be derived analytically from the staging error. To gain an understanding of how staging error propagates to sleep parameters, we performed a simulation study [34–36] where we resampled ground-truth hypnograms and monitored how the derived sleep parameters change. We note that in what follows, we used the error of sleep staging models instead of their accuracy, where error = 1 − accuracy. We refer to this as *staging error*.

The simulation study is based on the Sleep-EDF Database Expanded (Sleep-EDFx) dataset [37–39], which contains 197 PSGs scored by human experts according to the Rechtschaffen and Kales manual [40]. The data were obtained in two studies of mostly healthy adults. We refer to the original publications for more details on the composition of the subject group [37, 38]. We treated these 197 hypnograms, which we refer to as $x_1, ..., x_{197}$, as ground-truth hypnograms and use them to calculate ground-truth sleep parameters. Using a Monte-Carlo approach, we simulated the operation of a staging classifier with a given confusion matrix C by repeatedly resampling the hypnograms: We pseudo-randomly selected a new sleep stage for each epoch based on confusion matrix C. We then calculated the sleep parameters according to the simulated hypnogram and the residual error, i.e. the difference to the ground truth sleep parameter value. This was repeated n_{sim} times for every hypnogram. This gave us an estimated sleep parameter error for a given confusion matrix. The process is described in Algorithm 1 and schematically in Fig. 1.

Stated mathematically (following the notation in [34]), we considered the calculation of a sleep parameter y given a hypnogram x by computational model $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ where the random variable $X \in \mathcal{X}$ with probability density function $p_X(x)$ represents a hypnogram and $Y = \mathcal{M}(X)$ is a random variable representing the calculated sleep parameter.

Typically, uncertainty propagation is estimated by studying statistics such as the variance of the output variable, $Y \in \mathcal{Y}$. Since we treated the human expert scorer's classification as the ground truth, we instead considered the residual error, $e_y = y - \hat{y}$, that is the difference between the ground truth parameter value and the parameter value resulting from the simulation (sim), given a confusion matrix, C. The Monte Carlo estimation of sleep parameter error e_y is given by

$$\hat{e}_y = \frac{\sum_i^{n_{\mathrm{sim}}} (y - \hat{y}_i)}{n_{\mathrm{sim}}}$$

and the estimated root mean squared error (RMSE) and mean absolute error (MAE) over the dataset are given by

$$\mathrm{RMSE}_y = \frac{\sum_i^{n_{\mathrm{sim}}} \sqrt{\frac{\sum_j^N (y - \hat{y}_j)^2}{N}}}{n_{\mathrm{sim}}}$$

and

$$\mathrm{MAE}_y = \frac{\sum_i^{n_{\mathrm{sim}}} \frac{\sum_j^N |(y - \hat{y}_j)|}{N}}{n_{\mathrm{sim}}},$$

respectively, where N is the number of hypnograms in the dataset.

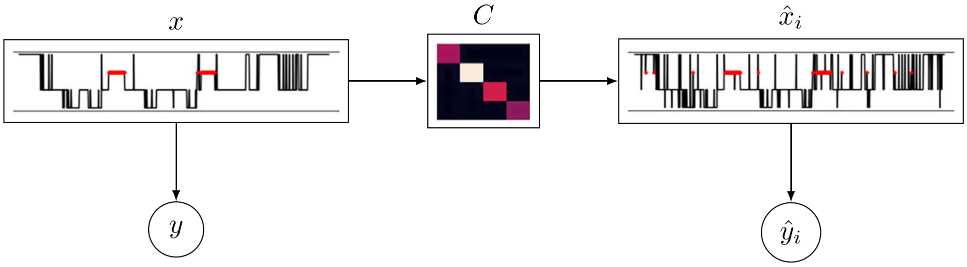

Fig. 1 A schematic of iteration i of the simulation process, which generates an estimated sleep parameter error, e_y. A manually scored hypnogram (ground truth) hypnogram, x, is selected from $\{x_1, ..., x_N\}$, a set of experimentally observed hypnograms. The resampled hypnogram, $\hat{x}_i$ is created by randomly reassigning each 30-second epoch in x a new sleep stage by making a random selection, weighted by row number s of confusion matrix, C, where s corresponds to the index of the true sleep stage. Sleep parameters y and $\hat{y}_i$ are calculated from x and $\hat{x}_i$, respectively and their residual, $e_y = y - \hat{y}_i$ is the sleep parameter error

Generation of Confusion Matrices

The Monte Carlo method described in Algorithm 1 simulates the resampling of a dataset of hypnograms as though they were estimated with confusion matrix C. We had a few examples of confusion matrices from the literature [18, 26, 28, 30], but in order to properly visualize the relationship between staging error and sleep parameter error, we also generated a multitude of hypothetical confusion matrices.

For a given target four-class staging error, $\tilde{e}_{\text{class}} \pm \epsilon$, where ϵ is a small value, we generated a hypothetical confusion matrix, $C_{\tilde{e}}$ as follows: $C_{\tilde{e}}$ is initialized as a 4x4 zero matrix. First, we assign the four elements on the diagonal of $C_{\tilde{e}}$ by sampling from a Beta distribution with a maximum at the target staging accuracy:

$$C_{\tilde{e}}(i,i) = M(i) \cdot k, \qquad k \sim \mathcal{B}(\alpha, \beta)$$

where $M(i)$ is the expected proportion of epochs of sleep stage $i \in \{1, 2, 3, 4\}$. We set

$$\alpha = \left(\frac{1-\mu}{\sigma^2} - \frac{1}{\mu}\right) \cdot \mu^2$$

and

$$\beta = \alpha \cdot \left(\frac{1}{\mu} - 1\right)$$

where μ is the target staging error and $\sigma^2 = 0.01$. The remaining cells of $C_{\tilde{e}(j)}$ were sampled uniformly from the range $\left[0; M(j) - \sum C_j\right]$ in random order. When there was only one cell left in row j it was assigned the value

$$M(j) \cdot \left(1 - \sum C_j\right)$$

This allowed us to visualize the relationship between staging error and sleep parameter error by generating thousands of hypothetical confusion matrices and applying Algorithm 1 to each one.

Table 2 Example of a generated confusion matrix of one thousand epochs. This confusion matrix corresponds to a classifier with an overall accuracy of 80% but only 42% REM sleep sensitivity

	Predicted			
	Wake	Light	Deep	REM
True				
Wake	69	15	4	12
Light	34	532	11	23
Deep	11	4	135	0
REM	6	2	79	63

Table 3 Example of a generated confusion matrix of one thousand epochs. This confusion matrix corresponds to a classifier with an overall accuracy of 80% but only 71% light sleep sensitivity

	Predicted			
	Wake	Light	Deep	REM
True				
Wake	89	1	6	4
Light	111	424	30	35
Deep	2	1	147	0
REM	3	4	5	138

Tables 2 and 3 show two examples out of twenty thousand generated confusion matrices, where $\mu = 0.2$ was used. These are examples of confusion matrices where sensitivity to specific sleep stages are low while overall accuracy is 80%. For the twenty thousand generated confusion matrices the overall accuracy was $(80 \pm 6)\%$ while the sensitivity to all individual sleep stages was $(80 \pm 10)\%$. Table 4 shows the average distribution of all twenty thousand confusion matrices.

Limitations

The simulation approach used has some shortcomings, which limit the scope of scenarios to which it is applicable. First of all, the resampling step in Algorithm 1 re-assigned each epoch a sleep stage independently of other epochs, so the method assumed that the staging error was random in time. This assumption holds for machine learning models where an epoch is scored without taking into account the epoch's temporal context, i.e., when $p(\hat{x}(k) \mid x(k))$ is independent of $x \setminus x(k)$. This is not the case for all of the machine learning models in Table 1. Secondly, the results of the simulation on the Sleep-EDFx dataset cannot be assumed to apply to other datasets unless the structure of participants' sleep is statistically similar to the Sleep-EDFx dataset. Finally, it should be acknowledged that all Monte Carlo methods are based on pseudo-random numbers, and uncertainty in the results is a

Table 4 The mean and standard deviation ($\bar{\mu} \pm$STD) of individual cells in twenty thousand generated confusion matrices with one thousand epochs each

	Predicted			
	Wake	Light	Deep	REM
True				
Wake	80 ± 10	7 ± 7	7 ± 7	7 ± 7
Light	40 ± 42	480 ± 59	40 ± 42	40 ± 42
Deep	10 ± 11	10 ± 11	119 ± 15	10 ± 11
REM	10 ± 11	10 ± 11	10 ± 11	119 ± 15

consequence not only of false assumptions but also of the method's inherent randomness [36].

Algorithm 1 Monte Carlo simulated sleep parameter error for sleep parameter y.

```
Require: Set of $N$ hypnograms $\{x_1, ...x_N\}$; confusion matrix, $C$; number of simulation repetitions, $n_{\text{sim}}$; formula for calculating sleep parameter, $f(x)$.
procedure SimulateParameterError($\{x_1, ...x_N\}, C$)
    for $i \in [1, n_{\text{sim}}]$ do
        for $j \in [1, N]$ do
            Select the $j^{\text{th}}$ hypnogram, $x_j$
            Calculate the ground-truth parameter value, $y_j = f(x_j)$.
            $M \leftarrow$ number of epochs in $x_j$
            for $k \in [1, M]$ do
                Set weights $w \leftarrow C_{x_j(k)}$, the $x_j(k)^{\text{th}}$ row of $C$.
                Generate $\hat{x}_j(k)$ by randomly selecting one of the sleep stages according to weights, $w$.
            end for
            Calculate estimated sleep parameter $\hat{y}_j = f(\hat{x}_j)$
            $e_{i,j} = y_j - \hat{y}_j$
        end for
    end for
    return $e_{i,j}$ or statistics of $e$ such as the mean of estimators $e_i$ across the $n_{\text{sim}}$ simulations, $\hat{e}$.
end procedure
```

Baseline Models

To put the simulated sleep parameter error into perspective, we need a naïve baseline model to compare against. To that end, we split the Sleep-EDFx dataset into a training and testing dataset with an approximately 50/50 split, ensuring that no individual participant's data was found in both sets. Using the training set, we calculated the average of sleep parameter y across every hypnogram,

$$\bar{y} = \frac{\sum_j^{N_{\text{train}}} y(x_j)}{N_{\text{train}}}$$

This was our baseline (BL) model for sleep parameter y, i.e. $\hat{y}^{\text{BL}} = \bar{y}$. The baseline sleep parameter model error was found by applying the baseline sleep parameter model to the test set:

$$\text{MAE}_y^{\text{BL}} = \frac{\sum_j^{N_{\text{test}}} \mid y_j - \bar{y} \mid}{N_{\text{test}}}$$

and

$$\text{RMSE}_y^{\text{BL}} = \sqrt{\frac{\sum_j^{N_{\text{test}}} (y_j - \bar{y})^2}{N_{\text{test}}}}$$

We also needed to decide on a range of staging errors for generating confusion matrices. To establish the highest reasonable staging error, we again created a naïve baseline model, this time for sleep staging. Using the test set, we split the dataset into half-hour time slots. We then counted how many epochs were assigned to each sleep stage per half-hour time slot. Figure 2 shows the result. The baseline staging model randomly assigned a sleep stage given the time of day, weighted by the proportional split of sleep stages in the corresponding half-hour time bin.

Results and Discussion

Algorithm 1 produced resampled hypnograms for a given staging error. Figure 3 shows examples of a hypnogram with 30-second epochs resampled by Algorithm 1, using confusion matrices reported in three of the papers in Table 1. We see that for a higher staging error, the hypnogram became more chaotic and less representative of actual sleep. However, the same may not necessarily be said about the derived sleep parameter error as reported in the literature. Of the machine learning models in Table 1, only Wulterkens et al. reported estimated sleep parameter statistics for four-class staging [26].

Figure 4 shows the distribution of residual sleep parameter error over the dataset using the same three confusion matrices. In blue, is the simulation of the staging in the paper by Wulterkens et al. (76% staging accuracy), and the dotted lines show the upper and lower limits of the error reported in the paper. When comparing the simulated error distribution to the error range reported by the authors (dotted lines in Fig. 4) we found that the simulation agreed reasonably well with their empirical results.

The simulated resampling of a low-error staging machine learning model resulted in a low-variance, low-bias sleep parameter estimation and as the staging error increased, so

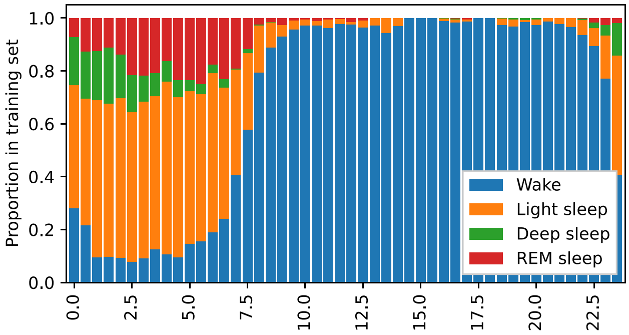

Fig. 2 Proportion of time spent in each sleep stage, every half hour from midnight, for the subjects in the training set. This information was used in a baseline staging model, which randomly assigned a sleep stage given the time of day to the subjects in the testing set. Abbreviations: REM: Rapid eye movement

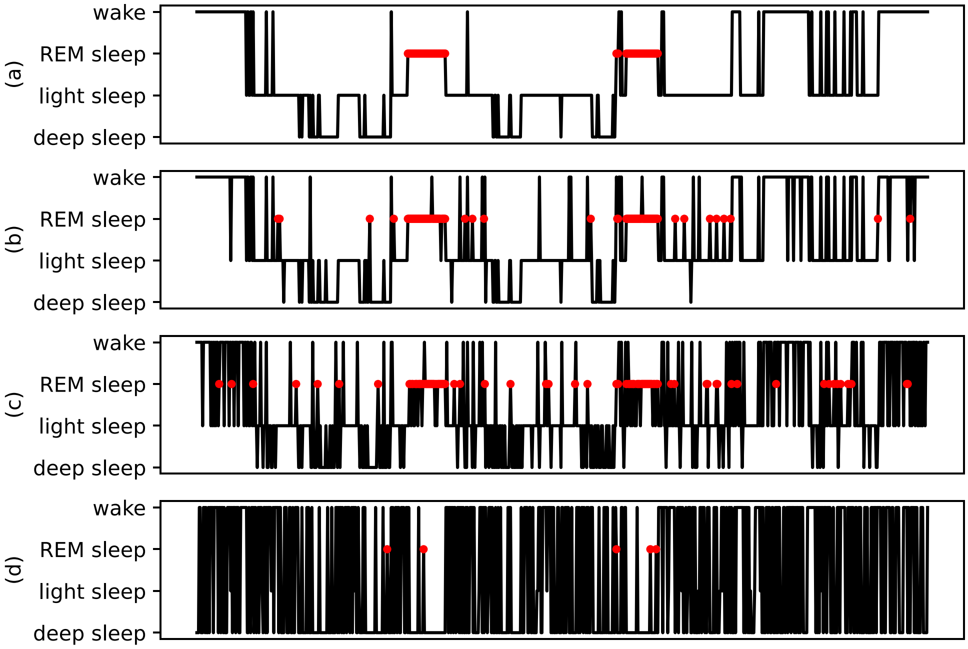

Fig. 3 The effect of resampling a hypnogram. a) An original manually scored hypnogram and the same hypnogram, resampled using the confusion matrix from b) Anusha et al. 2022 (93% staging accuracy), c) Wulterkens et al. (76% staging accuracy) and d) Li et al. 2019 (69% staging accuracy). Abbreviations: REM: Rapid eye movement

did the sleep parameter bias and variance (Fig. 4). Comparing the reported and simulated RMSE of the machine learning model from Wulterkens et al. listed in Table 5, we again found a reasonably good agreement.

Baseline values

We applied the baseline models for sleep staging and sleep parameters to test set. This gave us baseline error values to

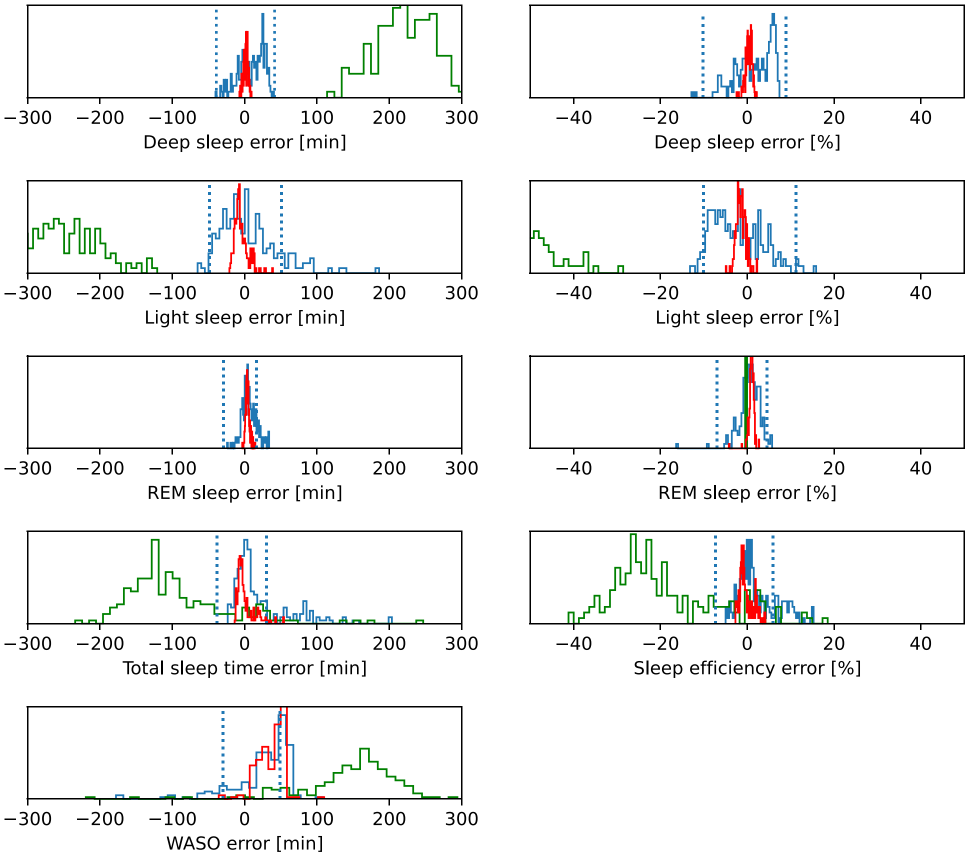

Fig. 4 Histograms showing the distribution of simulated sleep parameter error across the dataset for one iteration of the Monte Carlo method ($n_{\text{sim}} = 1$). The blue shows a simulation of the staging in the paper by Wulterkens et al. [26] (76% staging accuracy), and the dotted lines show the upper and lower limits of the error reported in the paper. The histogram in red shows a simulated error distribution for the staging in the paper by Anusha et al. [28] (93% staging accuracy), and the green shows a simulation of the machine learning model by Li et al. [41] (69% staging accuracy). Abbreviations: REM: Rapid eye movement, WASO: Wake after sleep onset

Table 5 RMSE mean and standard deviation for sleep parameter error residual from ground truth, as reported by Wulterkens et al. and the estimate from Monte Carlo simulation. The rows where reported values are missing correspond to parameters not reported by Wulterkens et al

Parameter	Error mean ± SD		RMSE	
	Reported	Simulated	Reported	Simulated
WASO [min]	9.8 ± 39.5	27.3 ± 46.8	40.6	54.2
TST [min]	−4.0 ± 34.1	22.2 ± 42.3	34.3	47.8
SE [%]	−0.7 ± 6.6	2.5 ± 4.6	6.7	5.3
Time in light sleep [min]	1.2 ± 50.0	6.4 ± 40.2	49.9	40.7
Proportion light sleep [%]	0.6 ± 10.7	−2.2 ± 5.9	10.7	6.3
Time in deep sleep [min]	1.2 ± 40.3	9.4 ± 17.9	40.2	20.2
Proportion deep sleep [%]	−0.6 ± 9.6	1.7 ± 4.3	9.6	4.7
Time in REM sleep [min]	−6.4 ± 22.9	6.4 ± 11.1	23.8	12.8
Proportion REM sleep [%]	−1.2 ± 5.8	0.6 ± 2.7	5.9	2.7
Number of awakenings		42.5 ± 47.1		63.4
REM latency [min]		−126.0 ± 117.2		172.1

Note: The uncertainty value in the simulated error column denotes the estimated error mean and estimated standard deviation, $\hat{\mu}(e)$ and $\hat{\sigma}^2(e)$, not the Monte Carlo standard error. Abbreviations: REM: Rapid eye movement, RMSE: Root mean squared error, SD: Standard deviation, SE: Sleep efficiency, TST: Total sleep time, WASO: Wake after sleep onset

compare the estimated sleep parameter error with. To be considered relevant, the models had to have an error value that was below this threshold. The resulting baseline values can be seen in Table 6. This set the lower baseline for 4-class sleep staging accuracy at 46% and a Cohen's kappa of 0.14. Table 6 also includes accuracy values for 2-, 3- and 5 class sleep staging baseline models for comparison. Furthermore, Table 6 lists baseline error values for sleep parameters in terms of MAE and RMSE. Not shown in Table 6 is the error in sleep parameters obtained by using hypnograms estimated using the baseline staging model. This consistently resulted in higher sleep parameter errors, than when using the average parameter values from the training set.

Clinical Relevance of Staging Error

Figure 5 shows the simulated sleep parameter RMSE over the whole dataset as a function of the four-class staging error for the twenty thousand generated confusion matrices, as well as confusion matrices reported in the literature (circles) [18, 26, 28, 41]. Also plotted are the staging error and sleep

Table 6 Performance of baseline models for staging and sleep parameter estimation. The staging model randomly assigns a sleep stage to an epoch according to the time of day. The parameter estimation model always assigns the average parameter value of the training set

Parameter	Baseline model performance
	Accuracy (Cohen's kappa)
Wake/sleep staging	74% (0.25)
Wake/NREM/REM staging	54% (0.15)
Wake/light/deep/REM staging	46% (0.14)
Wake/N1/N2/N3/REM staging	35% (0.10)
	MAE (RMSE)
WASO [min]	135.1 (197.0)
TST [min]	47.2 (64.0)
Number of awakenings [count]	6.5 (8.0)
SE [%]	14.1 (18.0)
REM Latency [min]	90.0 (142.5)
Time in light sleep [min]	52.7 (67.3)
Proportion light sleep [%]	9.9 (12.3)
Time in deep sleep [min]	32.2 (38.4)
Proportion deep sleep [%]	7.9 (9.3)
Time in REM sleep [min]	25.5 (33.7)
Proportion REM sleep [%]	5.2 (6.9)

Abbreviations: MAE: Mean absolute error, NREM: Non-rapid eye movement, REM: Rapid eye movement, RMSE: Root mean squared error, SE: Sleep efficiency, TST: Total sleep time, WASO: Wake after sleep onset

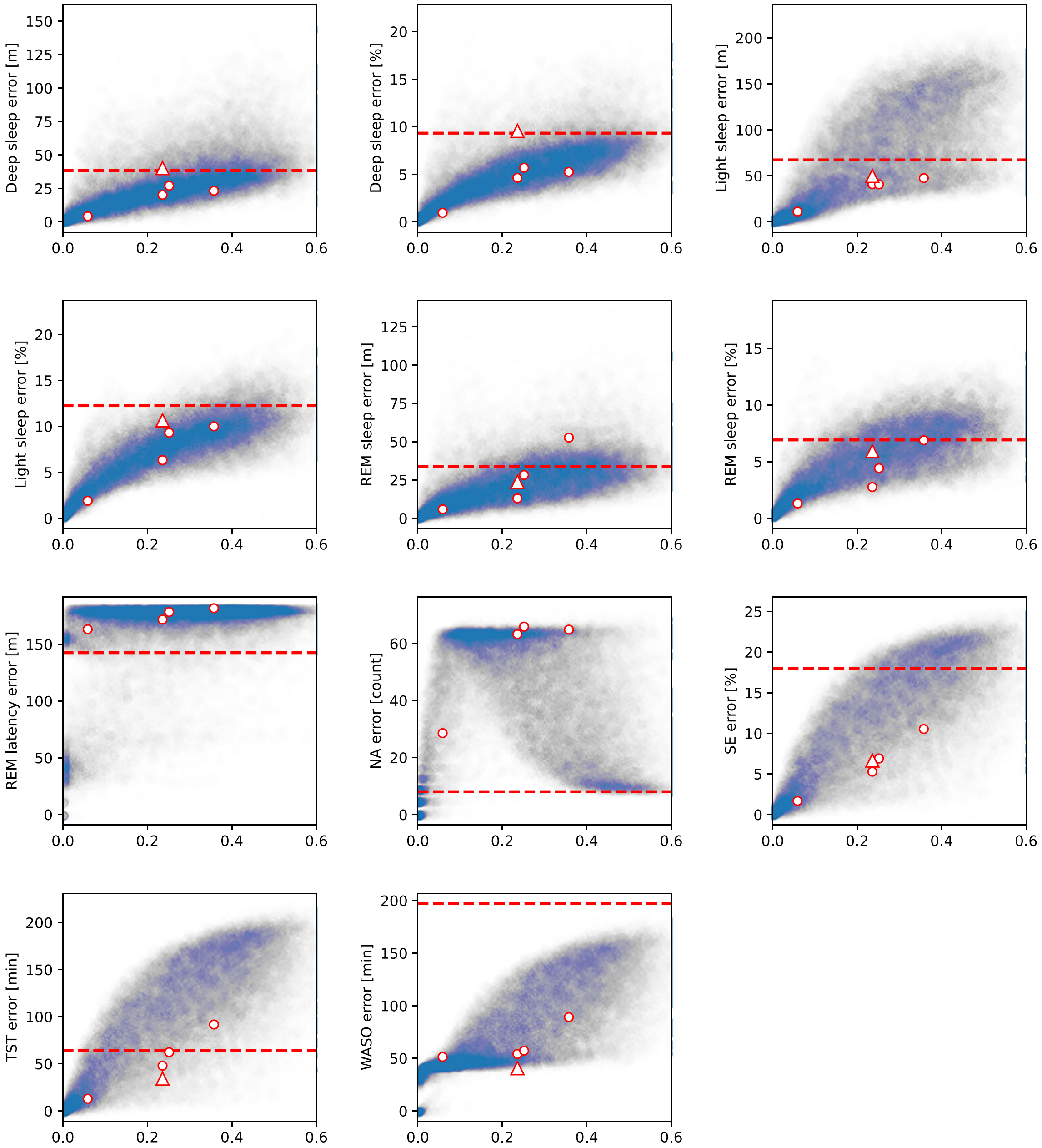

Fig. 5 The RMSE of different simulated sleep parameters as a function of the staging error of the confusion matrix used. The blue semi-transparent dots show simulations of twenty thousand generated confusion matrices, and the red circles show simulations of confusion matrices reported in the literature [18, 26, 28, 41]. The red triangles show the RMSE of reported sleep parameters for a given staging accuracy from Wulterkens et al. [26]. The dotted line shows the RMSE of the baseline sleep parameter estimation model. All data points which lie above the dotted line can be interpreted as less accurate than an educated guess. Abbreviations: NA: Number of awakening, REM: Rapid eye movement, RMSE: Root mean squared error, SE: Sleep efficiency, TST: Total sleep time, WASO: Wake after sleep onset

parameter RMSE reported in Wulterkens et al. (triangle) [26], again showing reasonable agreement between simulated and empirical results.[1]

Looking at Fig. 5, it is apparent that the relationship between staging error and sleep parameter error is non-linear. In general, the relationship seems to be logarithmic, although interestingly, error in the number of awakenings (NA error) does not strictly increase with staging error. This is an artefact of how NA is defined. A very inaccurate staging model will produce a hypnogram similar to the labeled (d) Fig. 3, with very few contiguous wake epochs. Since close to zero

[1] It should be noted that the relationship between kappa and staging error is approximately linear within the domain of four-class sleep staging on a homogeneous population. Thus, the relationship between sleep parameter error and Cohen's kappa of the staging machine learning model is similar to that seen in Fig. 5

contiguous wake epochs are found, the RMSE will be approximately equal to the true NA, on average around 13. This behaviour can be seen on the right side of the plot in Fig. 5. A more accurate staging model will correctly label more wake epochs, similar to the graph labeled (b) in Fig. 3. For prolonged wake bouts, made up of many epochs, some will likely be misclassified as sleep, splitting the wake bout in two. Thus, every misclassification of a wake epoch effectively increases NA by one, resulting in higher NA error, as shown by the peak in NA error for staging error between 0.05 and 0.35. In general, some parameters' error grows gradually as staging error increases, but others (particularly WASO, REM latency and NA) seem to be much more sensitive to minor staging errors.

The dotted lines in Fig. 5 show the RMSE of the baseline sleep parameter estimators. While some sleep parameters (such as WASO and NA) seem to be estimated below baseline error irrespective of staging accuracy, others (such as deep sleep and REM duration) are estimated with an error surpassing baseline error, even when staging error is low.

The simulation shows that a highly accurate staging model (one far surpassing the wearable staging accuracy of today's state of art) could conceivably perform worse than the baseline model on the estimation of derived sleep parameters. Under the assumptions made here, models capable of 90% sleep staging accuracy were, on average, estimated to perform below a baseline model in two out of eleven sleep parameters, essentially making the estimations worse than a random guess. For comparison, the current state-of-the-art in machine learning 5-class sleep staging on PSG data is around 85% [5, 7].

In recent years, a large body of literature has focused on developing machine learning models to automate the analysis of sleep measurements in order to ease the labour-intensive analysis and allow the use of wearables and other screening devices [7, 42, 43]. However, these models are seldom used in clinical practice [5, 44].

Clinically, sleep staging accuracy is an irrelevant metric by itself, yet it is the most popular task for training machine learning models in the literature. The results of the simulations performed here may indicate that machine learning models should not use staging accuracy as a target variable because of how the error propagates from sleep stages to sleep parameters. If the underlying purpose of estimating sleep stages is to calculate sleep parameters, it may prove more effective and robust to train regression models to predict sleep parameters directly and to report sleep parameters and their errors when training sleep staging models. This could lead to more accurate results, which is important for clinical practice.

Treating sleep parameter estimation as a regression task poses some new challenges. Because sleep parameters are typically properties of a whole night of sleep, in contrast to sleep stages, which are properties of individual epochs, the training data needed for sleep parameter regression may be much larger than for training sleep staging models. Furthermore, not all clinically important information is contained in sleep parameters. For example, following a sleep study, a clinician would typically review the hypnogram in order to get an overview of the sleep architecture. However, epoch-by-epoch sleep stage classification may not be the best way to generate the hypnogram and the staging error may not be indicative of the hypnogram error, as shown in Fig. 3.

Conclusions

We have presented a simulation framework for exploring how sleep staging error propagates to the calculations of sleep parameters. Our findings suggest that some sleep parameters are highly sensitive to staging error and that the error in sleep parameters may be worse than a baseline model, even when the staging accuracy is high. We have also presented a baseline sleep stage estimation model, which can be used as a reference for future work.

The simulation framework was evaluated using an open-source dataset of sleep studies and four-stage classification metrics from studies using wearable sleep trackers. However, the framework can be expanded to apply to other scenarios, such as five-class sleep staging and machine learning on PSG data. Future work could use this framework to investigate sleep parameter error propagation in other approaches, such as EEG or ECG, where sleep parameter error is more widely reported. Future work performing automated analysis of sleep data should not focus solely on attaining higher sleep staging accuracy. Error statistics of sleep parameters and other clinically relevant quantities should be reported.

In conclusion, this simulation study explores nuances around the topic of error in sleep parameters derived from an estimated hypnogram. Machine learning experts and sleep scientists should have a dialogue about alternative goals for the future.

Acknowledgements This work was part of the Sleep Revolution project [42], which is funded by the European Union's 2020 Research and Innovation Program under Grant 965417, and NordForsk (NordSleep project 90458) via the Icelandic Research Fund.

Declarations

Competing Interests Dr. Arnardottir discloses lecture fees from Nox Medical, Philips, ResMed, Jazz Pharmaceuticals, Linde Healthcare, Alcoa - Fjardaral, Visitor (Novo Nordisk) and Wink Sleep. She is also a member of the Philips Sleep Medicine & Innovation Medical Advisory Board. The other authors declare that they have no conflict of interest.

Human and animal rights This research involved data from human participants. The data is publically available [37, 38]

References

1. Arnardottir ES, Islind AS, Óskarsdóttir M. The Future of Sleep Measurements: A Review and Perspective. Sleep Med Clinics. 2021;16(3):447–64. https://doi.org/10.1016/j.jsmc.2021.05.004.
2. de Zambotti M, Cellini N, Goldstone A, Colrain IM, Baker FC. Wearable Sleep Technology in Clinical and Research Settings. Med Sci Sports Exercise. 2019;51(7):1538–57. https://doi.org/10.1249/MSS.0000000000001947.
3. Phan H, Mikkelsen K (2022) Automatic sleep staging of EEG signals: recent development, challenges, and future directions. Physiological Measurement, 43(4):04TR01. ublisher: IOP Publishing. https://doi.org/10.1088/1361-6579/ac6049.
4. Faust O, Razaghi H, Barika R, Ciaccio EJ, Acharya UR. A review of automated sleep stage scoring based on physiological signals for the new millennia. Comput Methods Prog Biomed. 2019;176:81–91. https://doi.org/10.1016/j.cmpb.2019.04.032.
5. Fiorillo L, Puiatti A, Papandrea M, Ratti PL, Favaro P, Roth C, et al. Automated sleep scoring: a review of the latest approaches. Sleep Med Rev. 2019;48: 101204. https://doi.org/10.1016/j.smrv.2019.07.007.
6. Berry RB, Quan SF, Abreu AR, Bibbs ML, DelRosso L, Harding SM, et al. The AASM Manual for the Scoring of Sleep and Associated Events: Rules. American Academy of Sleep Medicine: Terminology and Technical Specifications; 2020.
7. van Gorp H, Huijben IAM, Fonseca P, van Sloun RJG, Overeem S, van Gilst MM (2022) Certainty about uncertainty in sleep staging: a theoretical framework Sleep 45(8):zsac134. https://doi.org/10.1093/sleep/zsac134.
8. Kang DY, DeYoung PN, Tantiongloc J, Coleman TP, Owens RL (2021) Statistical uncertainty quantification to augment clinical decision support: a first implementation in sleep medicine. npj Digital Medicine 4(1):1–9. Number: 1 Publisher: Nature Publishing Group. https://doi.org/10.1038/s41746-021-00515-3.
9. Anderson GM. Error propagation by the Monte Carlo method in geochemical calculations. Geochimica et Cosmochimica Acta. 1976;40(12):1533–8. https://doi.org/10.1016/0016-7037(76)90092-2.
10. Kager J, Herwig C. Monte Carlo-Based Error Propagation for a More Reliable Regression Analysis across Specific Rates in Bioprocesses. Bioengineering. 2021;8(11):160. https://doi.org/10.3390/bioengineering8110160.
11. Li Y, Zhang S, Hobbs R, Caiado C, Sproson A, Selby D, et al. Monte Carlo sampling for error propagation in linear regression and applications in isochron geochronology. Sci Bulletin. 2019;64. https://doi.org/10.1016/j.scib.2018.12.019.
12. Óskarsdóttir M, Islind AS, August E, Arnardóttir ES, Patou F, Maier AM (2022) Importance of Getting Enough Sleep and Daily Activity Data to Assess Variability: Longitudinal Observational Study. JMIR Format Res 6(2):e31807. Company: JMIR Formative Research Distributor: JMIR Formative Research Institution: JMIR Formative Research Label: JMIR Formative Research Publisher: JMIR Publications Inc., Toronto, Canada. https://doi.org/10.2196/31807.
13. Khosla S, Deak MC, Gault D, Goldstein CA, Hwang D, Kwon Y, et al (2018) Consumer Sleep Technology: An American Academy of Sleep Medicine Position Statement. J Clinical Sleep Med 14(05):877–880. Publisher: American Academy of Sleep Medicine. https://doi.org/10.5664/jcsm.7128.
14. Depner CM, Cheng PC, Devine JK, Khosla S, de Zambotti M, Robillard R, et al (2020) Wearable technologies for developing sleep and circadian biomarkers: a summary of workshop discussions. Sleep 43(2):zsz254. https://doi.org/10.1093/sleep/zsz254.
15. Menghini L, Cellini N, Goldstone A, Baker FC, de Zambotti M (2021) A standardized framework for testing the performance of sleep-tracking technology: step-by-step guidelines and open-source code. Sleep 44(2):zsaa170. https://doi.org/10.1093/sleep/zsaa170.
16. Rentz LE, Ulman HK, Galster SM (2021) Deconstructing Commercial Wearable Technology: Contributions toward Accurate and Free-Living Monitoring of Sleep Sensors 21(15):5071. Number: 15 Publisher: Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/s21155071.
17. de Zambotti M, Cellini N, Menghini L, Sarlo M, Baker FC. Sensors capabilities, performance, and use of consumer sleep technology. Sleep Medicine Clinics. 2020;15(1):1–30. https://doi.org/10.1016/j.jsmc.2019.11.003.
18. • Radha M, Fonseca P, Moreau A, Ross M, Cerny A, Anderer P, et al (2021) A deep transfer learning approach for wearable sleep stage classification with photoplethysmography. npj Digital Medicine 4(1):135. https://doi.org/10.1038/s41746-021-00510-8. This study described the use of machine learning for sleep staging on data from wrist-worn wearables. Classification error statistics from this study were used in our simulations.
19. : Wearable device market share by category worldwide (2021) Available from: https://www.statista.com/statistics/690731/wearables-worldwide-shipments-by-product-category/.
20. : Global smartwatch shipments market share (2022) Available from: https://www.statista.com/statistics/910862/worldwide-smartwatch-shipment-market-share/.
21. Imtiaz SA (2021) A Systematic Review of Sensing Technologies for Wearable Sleep Staging. Sensors 21(5):1562. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/s21051562.
22. Lujan MR, Perez-Pozuelo I, Grandner MA (2021) Past, Present, and Future of Multisensory Wearable Technology to Monitor Sleep and Circadian Rhythms. Frontiers in Digital Health 3
23. Roy Y, Banville H, Albuquerque I, Gramfort A, Falk TH, Faubert J (2019) Deep learning-based electroencephalography analysis: a systematic review. J Neural Eng 16(5):051001.Publisher: IOP Publishing. https://doi.org/10.1088/1741-2552/ab260c.
24. Boostani R, Karimzadeh F, Nami M. A comparative review on sleep stage classification methods in patients and healthy individuals. Computer Methods and Programs in Biomedicine. 2017;140:77–91. https://doi.org/10.1016/j.cmpb.2016.12.004.
25. • Gashi S, Alecci L, Lascio ED, Debus ME, Gasparini F, Santini S (2022) The Role of Model Personalization for Sleep Stage and Sleep Quality Recognition Using Wearables. IEEE Pervasive Comput 21(2):69–77. Conference Name: IEEE Pervasive Comput. https://doi.org/10.1109/MPRV.2022.3164334. This study described the use of machine learning for sleep staging on data from

wrist-worn wearables. Classification error statistics from this study were used in our simulations.

26. • Wulterkens BM, Fonseca P, Hermans LW, Ross M, Cerny A, Anderer P, et al (2021) It is All in the Wrist: Wearable Sleep Staging in a Clinical Population versus Reference Polysomnography. Nature and Science of Sleep 13:885–897. Publisher: Dove Press. https://doi.org/10.2147/NSS.S306808. This study described the use of machine learning for sleep staging on data from wrist-worn wearables. Classification error statistics from this study were used in simulations.
27. • Fonseca P, Weysen T, Goelema MS, Møst EIS, Radha M, Lunsingh Scheurleer C, et al (2017) Validation of Photoplethysmography-Based Sleep Staging Compared With Polysomnography in Healthy Middle-Aged Adults. Sleep 40(7):zsx097. https://doi.org/10.1093/sleep/zsx097. This study described the use of machine learning for sleep staging on data from wrist-worn wearables. Classification error statistics from this study were used in our simulations.
28. • Anusha AS, Preejith SP, Akl TJ, Sivaprakasam M,. Electrodermal activity based autonomic sleep staging using wrist wearable. Biomedical Signal Processing and Control. 2022;75: 103562. https://doi.org/10.1016/j.bspc.2022.103562. This study described the use of machine learning for sleep staging on data from wrist-worn wearables. Classification error statistics from this study were used in our simulations
29. • Ko YF, Kuo PH, Wang CF, Chen YJ, Chuang PC, Li SZ, et al (2022) Quantification Analysis of Sleep Based on Smartwatch Sensors for Parkinson's Disease. Biosensors 12(2):74. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/bios12020074. This study described the use of machine learning for sleep staging on data from wrist-worn wearables. Classification error statistics from this study were used in our simulations.
30. • Li Q, Li Q, Cakmak AS, Poian GD, Bliwise DL, Vaccarino V, et al (2021) Transfer learning from ECG to PPG for improved sleep staging from wrist-worn wearables. Physiological Measurement 42(4):044004. Publisher: IOP Publishing. https://doi.org/10.1088/1361-6579/abf1b0. This study described the use of machine learning for sleep staging on data from wrist-worn wearables. Classification error statistics from this study were used in our simulations.
31. • Sundararajan K, Georgievska S, te Lindert BHW, Gehrman PR, Ramautar J, Mazzotti DR, et al (2021) Sleep classification from wrist-worn accelerometer data using random forests. Scientific Reports. 11(1):24. Number: 1 Publisher: Nature Publishing Group. https://doi.org/10.1038/s41598-020-79217-x. This study described the use of machine learning for sleep staging on data from wrist-worn wearables. Classification error statistics from this study were used in our simulations.
32. • Walch O, Huang Y, Forger D, Goldstein C (2019) Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device. Sleep 42(12):zsz180. https://doi.org/10.1093/sleep/zsz180. This study described the use of machine learning for sleep staging on data from wrist-worn wearables. Classification error statistics from this study were used in our simulations.
33. Brupbacher G, Straus D, Porschke H, Zander-Schellenberg T, Gerber M, von Känel R, et al. The acute effects of aerobic exercise on sleep in patients with depression: study protocol for a randomized controlled trial. Trials. 2019;20(1):352. https://doi.org/10.1186/s13063-019-3415-3.
34. Zhang J (2021) Modern Monte Carlo methods for efficient uncertainty quantification and propagation: a survey. WIREs Comput Stat 13(5):e1539 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wics.1539.
35. Couto PRG, Damasceno JC, Oliveira SPd, Couto PRG, Damasceno JC, Oliveira SPd (2013) Monte Carlo Simulations Applied to Uncertainty in Measurement. IntechOpen. Publication Title: Theory and Applications of Monte Carlo Simulations. Available from: https://www.intechopen.com/state.item.id.
36. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med. 2019;38(11):2074–102. https://doi.org/10.1002/sim.8086.
37. Kemp B, Zwinderman AH, Tuk B, Kamphuisen HAC, Oberye JJL (2000) Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. IEEE Trans Biomed Eng 47(9):1185–1194. Conference Name: IEEE Transactions on Biomedical Engineering. https://doi.org/10.1109/10.867928.
38. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation. 2000;101(23):E215-220. https://doi.org/10.1161/01.cir.101.23.e215.
39. Mourtazaev MS, Kemp B, Zwinderman AH, Kamphuisen HA. Age and gender affect different characteristics of slow waves in the sleep EEG. Sleep. 1995;18(7):557–64. https://doi.org/10.1093/sleep/18.7.557.
40. Rechtschaffen A, Kales A. A manual of standardized terminology, techniques and scoring system of sleep stages in human subjects. Los Angeles: University of California, Brain Information Service/Brain Research Institute; 1968.
41. Li Q, Li Q, Liu C, Shashikumar SP, Nemati S, Clifford GD (2018) Deep learning in the cross-time frequency domain for sleep staging from a single-lead electrocardiogram. Physiological Measurement 39(12):124005. Publisher: IOP Publishing. https://doi.org/10.1088/1361-6579/aaf339.
42. Arnardottir ES, Islind AS, óskarsdóttir M, Ólafsdóttir KA, August E, Jónasdóttir L, et al. The Sleep Revolution project: the concept and objectives. J Sleep Res. 2022;31(4): e13630. https://doi.org/10.1111/jsr.13630.
43. Mallett J, Arnardottir ES. Improving Machine Learning Technology in the Field of Sleep. Sleep Med Clinics. 2021;16(4):557–66. https://doi.org/10.1016/j.jsmc.2021.08.003.
44. Kim D, Lee J, Woo Y, Jeong J, Kim C, Kim DK. Deep Learning Application to Clinical Decision Support System in Sleep Stage Classification. J Personal Med. 2022;12(2):136. https://doi.org/10.3390/jpm12020136.