http://dx.doi.org/10.5665/sleep.1712

Assessing REM Sleep in Mice Using Video Data

Blakeley B. McShane, PhD[1]; Raymond J. Galante[2]; Michael Biber, MD[3]; Shane T. Jensen, PhD[4]; Abraham J. Wyner, PhD[4]; Allan I. Pack, MBChB, PhD[2]

[1]Kellogg School of Management, Northwestern University, Evanston, IL; [2]Center for Sleep and Circadian Neurobiology, University of Pennsylvania, Philadelphia, PA; [3]Beth Israel Deaconess Medical Center, Harvard Medical School, Boston MA; [4]The Wharton School, University of Pennsylvania, Philadelphia, PA

Study Objectives: Assessment of sleep and its substages in mice currently requires implantation of chronic electrodes for measurement of electroencephalogram (EEG) and electromyogram (EMG). This is not ideal for high-throughput screening. To address this deficiency, we present a novel method based on digital video analysis. This methodology extends previous approaches that estimate sleep and wakefulness without EEG/EMG in order to now discriminate rapid eye movement (REM) from non-REM (NREM) sleep.

Design: Studies were conducted in 8 male C57BL/6J mice. EEG/EMG were recorded for 24 hours and manually scored in 10-second epochs. Mouse behavior was continuously recorded by digital video at 10 frames/second. Six variables were extracted from the video for each 10-second epoch (i.e., intraepoch mean of velocity, aspect ratio, and area of the mouse and intraepoch standard deviation of the same variables) and used as inputs for our model.

Measurements and Results: We focus on estimating features of REM (i.e., time spent in REM, number of bouts, and median bout length) as well as time spent in NREM and WAKE. We also consider the model's epoch-by-epoch scoring performance relative to several alternative approaches. Our model provides good estimates of these features across the day both when averaged across mice and in individual mice, but the epoch-by-epoch agreement is not as good.

Conclusions: There are subtle changes in the area and shape (i.e., aspect ratio) of the mouse as it transitions from NREM to REM, likely due to the atonia of REM, thus allowing our methodology to discriminate these two states. Although REM is relatively rare, our methodology can detect it and assess the amount of REM sleep.

Keywords: Mouse sleep, inbred mouse strains, REM, high-throughput phenotyping

Citation: McShane BB; Galante RJ; Biber M; Jensen ST; Wyner AJ; Pack AI. Assessing REM sleep in mice using video data. *SLEEP* 2012;35(3):433-442.

INTRODUCTION

A major focus of current research in mice is to elucidate gene products that (1) regulate sleep and wake, (2) are regulated by sleep and wake, or (3) are affected by sleep deprivation. Multiple strategies are used to identify relevant genes (for review, see[1]). These include the analysis of changes in the transcriptome with sleep, wake, and sleep deprivation (for review, see[2]) as well as the use of specific transgenic mice. Transgenic strategies typically create altered gene function on a C57BL/6J background by using selective breeding strategies to minimize the effects of genetic background. There are currently large numbers of mice available with knockout of specific genes.[3] Use of these new mouse resources requires evaluation of the effect of knockout of a specific gene on sleep and its substages as well as wakefulness.

Currently, this evaluation is performed by assessing changes in the electroencephalogram (EEG) and electromyogram (EMG). This technique requires surgery on mice with the implantation of electrodes; mice cannot be studied until they recover from the surgery. Moreover, scoring of EEG/EMG records is labor intensive: if states are assessed in 10-second epochs across a 24-hour period, there are 8,640 epochs to be scored, whereas with 4-second epochs, there are 21,600 epochs. Thus, the necessity for surgery, time to recover from surgery, and scoring of large numbers of epochs adds expense and makes studies of sleep in mice very labor intensive.

We seek an alternative high-throughput strategy that will obviate the need for EEG/EMG recording. This strategy can be used (1) in studies that evaluate changes in mRNA, protein, etc. in response to sleep, wake, and sleep deprivation and (2) to screen the large panel of knockout mice that have already been created.[3]

Two approaches to high-throughput phenotyping have already been proposed.[4,5] One approach is based on determining inactivity either by electronic beam splits or by video analysis; any duration of inactivity that lasts 40 seconds or more is considered sleep. This approach, termed the 40-second Rule, has been validated by comparison with manual scores based on EEG/EMG recordings in both young[4] and old[6] C57BL/6J mice. The other strategy is based on piezoelectric detection of mouse movements by pressure sensors in the floor of the mouse cage; analysis of the data recorded by such sensors reveals patterns that are characteristic of sleep and wakefulness.[5]

Although these methods are quite accurate to determine wake and sleep, they cannot distinguish non-rapid eye movement sleep (NREM) from rapid eye movement sleep (REM). Nonetheless, our video recordings show a subtle signal for REM sleep. In particular, the area and aspect ratio of the mouse, respectively, increased and decreased when the mouse went from NREM to REM sleep (see Figure 1). This, we believe, is related to the mouse becoming more atonic in REM sleep.

The goal of this study was, therefore, to determine whether we could develop an algorithm to identify REM vs NREM

Submitted for publication June, 2011
Submitted in final revised form August, 2011
Accepted for publication September, 2011

Address correspondence to: Allan I. Pack, MBChB, PhD, Division of Sleep Medicine, Department of Medicine, Center for Sleep and Circadian Neurobiology, University of Pennsylvania Perelman School of Medicine, 125 South 31st Street, Suite 2100, Philadelphia, PA 19104-3403; Tel: (215) 746-4806; Fax: (215) 746-4814; E-mail: pack@mail.med.upenn.edu

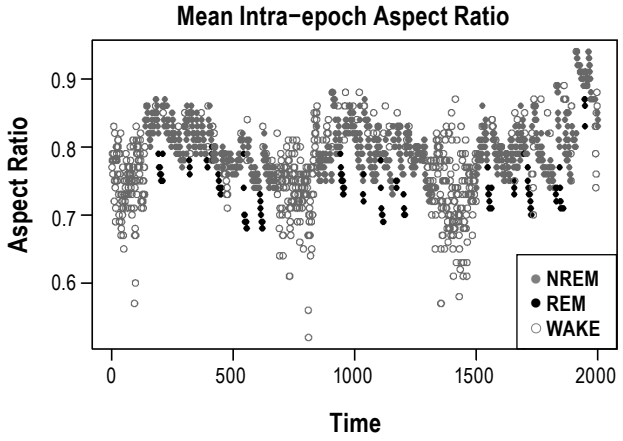

Figure 1—Mean intraepoch aspect ratio. A time-series plot of mean aspect ratio for 1 mouse with colors corresponding to the manually scored state (WAKE, non-rapid eye movement (NREM), and rapid eye movement (REM) sleep). Subtle differences among the 3 states can be detected visually.

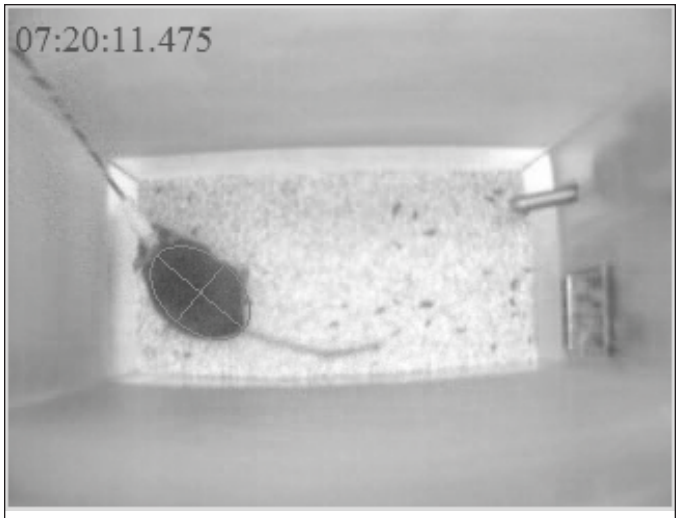

Figure 2—Mouse with tracking ellipse. One frame of video data with an ellipse imposed by our tracking software. Using the ellipse, we can calculate the size, aspect ratio, and velocity of the mouse.

sleep in mice based on digital video recordings. This is challenging because REM sleep is a relatively rare state compared with NREM sleep and wake and because episodes of REM sleep are short. We show that the identification of REM vs NREM is possible with reasonable accuracy, and we validate this by comparison with EEG/EMG assessments of REM sleep in C57BL/6J male mice. This new phenotyping strategy will be valuable for studies of molecular change in response to sleep, wake, or sleep deprivation and for screening of the recently created large number of knockout mice[3] to determine if they have altered sleep and wake.

ANIMAL STUDIES

One inbred strain of male mice was used in this study: C57BL/6J (n = 8, age: 10 to 12 weeks, weight: 18 to 23 g), purchased from Jackson Laboratory, Inc. (Bar Harbor, ME). Mice were individually housed in Plexiglas cages (4" wide × 8" long × 12" high) and maintained on a 12-hour light/dark cycle (lights on 0700; 80 lux at the floor of the cage) in a sound-attenuated recording room, temperature 22°C-24°C. Food and water were available *ad libitum*. Animals were acclimated to these conditions for 10-14 days before beginning any studies. All animal experiments were performed in accordance with the guidelines published in the NIH Guide for the Care and Use of Laboratory Animals and were approved by the University of Pennsylvania Animal Care and Use Committee.

Mice were implanted with EEG/EMG electrodes under deep anesthesia (intraperitoneal injection of ketamine [100 mg/kg] / xylazine [10 mg/kg]). For EEG recordings, 3 stainless-steel miniature screws (0-80 × 1/16, Plastics One, Inc., Roanoke, VA) were placed epidurally in the following locations: (1) right frontal cortex (1.7 mm lateral to midline and 1.5 mm anterior to bregma), (2) right parietal cortex (1.7 mm lateral to midline and 1 mm anterior to lambda), and (3) a reference electrode over the cerebellum (1 mm posterior to lambda on the midline). Two EMG electrodes were sutured onto the dorsal surface of the nuchal muscles immediately posterior to the skull. All leads from the electrodes were connected to an 8-pin plastic connector/pedestal (Plastics One, Inc.) and then bonded to the skull with dental acrylic. After the bonding agent cured, the animals were connected to our signal-amplifier system using a connecting cable and swivel contact (Plastics One, Inc.) mounted above each cage. All mice were given 10-14 days for postoperative recovery and habituation before beginning any recording.

EEG and EMG signal were amplified using the Neurodata amplifier system (Model M15, Astro-Med, Inc., West Warwick, RI). Signals were amplified (2000×) and conditioned using the following settings for EEG signals: low cut-off frequency (-6dB), 0.3 Hz and high cut-off frequency (-6dB), 30 Hz; for EMG signals: low cut-off frequency (-6dB), 10 Hz and high cut-off frequency (-6dB), 100 Hz. Signals were digitized at 100 Hz. All data were acquired and analyzed using Gamma software (Astro-Med, Inc.) and converted to European data format (EDF) for manual scoring and analysis in the Somnologica science software (Embla, Inc., Denver, CO).

WAKE, NREM, and REM sleep were manually scored using EEG/EMG in 10-second epochs during 24-hour baseline recordings. Sleep stages were determined as follows: epochs were scored as wake when the EMG amplitude ranged from activity slightly higher than baseline during quiet wakefulness to higher-amplitude activity during ambulation. EEG amplitude was low, with frequencies mostly above 10 Hz. NREM was characterized by high-amplitude delta (1-4 Hz). EMG was constant with low-amplitude activity. REM was scored when low-amplitude rhythmic theta waves (6-9 Hz) predominated, with the EMG remaining at baseline levels. Although our goal is to replace this manual scoring with an automated video-based system, these EEG/EMG-based manual scores will be our "gold standard" for comparison because they are currently the most widely accepted method for accurately scoring sleep.

Twenty-four hours of data divided into 10-second epochs implies 8,640 epochs for each of the 8 mice, giving us a total of 69,120 epochs that have been manually scored as REM, NREM, or WAKE. For each of these epochs, we also have video recordings captured at 10 frames per second, giving us 100 frames of

video data per epoch upon which to build our automated system (see Figure 2 for one such frame).

Tracking software was used to calculate, for each epoch with time index t, six continuous numerical covariates: the within-epoch mean of the velocity, aspect ratio, and size of the mouse and the within-epoch standard deviation of the velocity, aspect ratio, and size of the mouse (where the mouse is approximated by a tracking ellipse as shown in Figure 2). For velocity and size, we used the natural logarithms of the means and standard deviations as covariates. We also had one binary covariate which indicates whether or not the light in the cage was turned on (lights were on from 0700-1900). Henceforth, we denote the vector of our seven covariates for epoch t as X_t.

MATHEMATICAL APPROACHES

Model

The sequential classification problem we face (i.e., the automated sleep scoring of mice) can be conceptualized by considering the data as consisting of two components, an "in-sample" component and an "out-of-sample" component. The in-sample component consists of all of the data from a single mouse, namely (1) the sleep states (Y_1, Y_2,…,$Y_{8,640}$) where each Y_t is one of NREM, REM, or WAKE and (2) the video-based covariates (X_1, X_2,…,$X_{8,640}$). Using this in-sample data, we estimate a model that predicts the collection of Y_t from the collection of X_t. The out-of-sample data component, in contrast, comes from a different mouse and consists of only the video-based covariates denoted ($\widetilde{X}_1$, $\widetilde{X}_2$,…,$\widetilde{X}_{8,640}$). The goal is to predict the corresponding ($\widetilde{Y}_1$, $\widetilde{Y}_2$,…,$\widetilde{Y}_{8,640}$) using the estimated model and the collection of $\widetilde{X}_t$.

Our modeling strategy builds on a statistical technique known as random forests.[7] A random forest is a collection of classification (or decision) trees,[8] each of which is constructed using random subsamples of the data and the covariates. The random forest combines the predictions made by each tree by allowing them each to "vote" on a sleep state; the probability of each sleep state is determined by the fraction of votes it receives, and the predicted state is the one with the most votes.

Although the random forest algorithm is known to perform well in a wide variety of settings, it ignores a key feature of sleep data: namely, that the Y_t and X_t form sequences in time. This sequential nature leads to dependencies in the data. For example, if a mouse was awake in the last epoch (i.e., Y_{t-1} = WAKE), there is a high probability it will be awake this epoch (i.e., Y_t = WAKE). It should be possible to modify the basic random forest to account for these dependencies and to thus enhance performance.

To do so, we build on previous work, which combines conventional methods with Markov models.[9,10] The general structure of a Markov model is illustrated in Figure 3. The mouse starts at time $t = 1$ in sleep state Y_1 (i.e., one of NREM, REM, or WAKE), and we observe video-based covariates X_1 that depend on Y_1. Next, the mouse transitions to state Y_2, and the process repeats itself until time $t = T$ (in our case, $T = 8,640$).

In our modeling of sleep states in mice, we consider two particular Markovian enhancements of the basic random forest. First, we combine the random forest with a first-order Markov model. This enhances the random forest so that it takes account of local time dependencies (i.e., those that are nearby in time). Although accounting for such dependencies will likely substantially improve model performance, the first-order Markov assumption imposes several important restrictions. In particular, it implies that sleep-bout durations are (1) geometrically distributed and (2) do not depend on the previous state (e.g., the model assumes that WAKE bout lengths are distributed the same regardless of whether the previous bout was NREM or REM). Prior literature has found both of these assumptions untenable,[11] and, indeed, the unconditional fits of a geometric distribution to our data were quite poor.

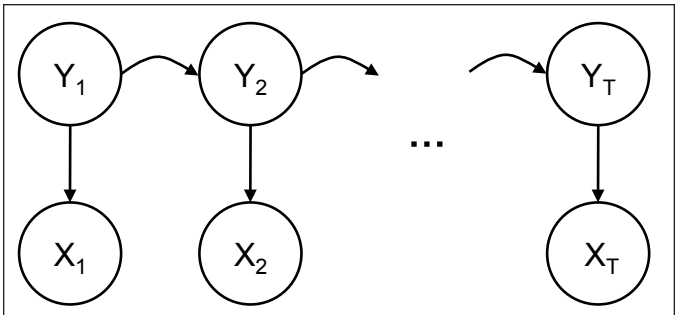

Figure 3—Markov chain structure.

Motivated by these observations, we therefore also combined the random forest with a transition-dependent generalized Markov model. This allows the random forest to take account of very general time-dependence structures, including (1) non-local dependence, (2) bout duration distributions that are not geometrically distributed, and (3) bout duration distributions that depend on the previous state.

When fit to data, our model provides an estimate of the probability that a mouse is in a given sleep state at a given epoch. Formally, our model estimates the probabilities $\lambda_i^t \equiv \mathbb{P}(Y_t = i \mid X, \widehat{\Theta})$ where i is one of NREM, REM, or WAKE, t indexes the epochs, X is the full set of video covariates (X_1,…,$X_{8,640}$), and $\widehat{\Theta}$ is an estimate of the model parameters. Our actual prediction $\widehat{Y}_t$ for epoch t is taken to be whichever state (NREM, REM, or WAKE) has the largest λ_i^t at epoch t. We note that our estimates λ_i^t should be superior to the probability estimates produced by the basic random forests algorithm, which ignores the time-series structure and predicts Y_t based only on X_t.

While the technical details pertaining to the estimation and computation of the models outlined above are beyond the scope of this manuscript, they can be found elsewhere.[12] Nonetheless, we note that the algorithm is fast, requiring only several seconds to estimate using the entire sequence of datapoints (i.e., all 24 hours worth of data) from a given mouse and only several minutes to predict on the entire sequence of datapoints (i.e., all 24 hours worth of data) from a different mouse. We also note that exploiting time dependencies greatly enhances our ability to detect signal in the data, particularly given the inherently high noise level. We will show that our proposed method is highly advantageous in terms of predicting REM sleep.

Evaluation

We focus our model evaluation on determining how well our model can track (1) the total amount of time spent in REM sleep, (2) the number of REM bouts, and (3) the median REM bout length using the values derived from EEG/EMG manual

scoring as the benchmark. In particular, we break our 24 hours worth of data into 12 two-hour blocks, and we examine these three metrics averaged across all mice for each of the blocks. We also examine how well the model performs at predicting total amount of time spent in each of the three states (REM, NREM, and WAKE) individually for each mouse.

A novel aspect of our methodology is that our model includes a threshold-tuning parameter that takes λ^t_{REM}, the probability of REM sleep in epoch t as given by our model, and "converts" it into a REM score for epoch t. This parameter can be set by the user to adjust the specificity and sensitivity of the model's predictions so that the predictions can take account of the relative costs of false positives and negatives (which typically vary from application to application). We discuss this parameter and how to optimally tune it more fully in the Aggregate Measures of REM subsection of the Results section.

Although we focus on the summary statistics discussed above, we also examine how well our model is able to match the gold standard manual scores on an epoch-by-epoch basis. Given that manual scoring is currently the most widely accepted method for accurately scoring sleep, matching manual scores to a reasonable degree is important. Nonetheless, there are several issues related to epoch-by-epoch matching worthy of mention. First, we anticipate that most applications of our methodology will focus on estimating the summary statistics rather than the epoch-by-epoch scores. While matching manual scores on an epoch-by-epoch basis is a *sufficient* condition for estimating the summary statistics, it is by no means a *necessary* one, and accurate estimates of the summary statistics can be obtained from models that are less precise on an epoch-by-epoch basis. Second, epoch-by-epoch manual scores are inconsistent: each of our epochs was scored independently by two different scorers who disagreed on approximately 5% of the epochs,[3] with disagreement rates highest among those epochs in which the sleep stage was transitional (in such cases, an independent third scorer was used to break the tie and to determine the "truth"). Consequently, the maximum possible epoch-by-epoch agreement rate between any model and manual scores will be below 100%.

RESULTS

Aggregate Measures of REM

We focus our model evaluation on aggregate measures of REM sleep. Specifically, we break the 24 hours worth of data into 12 two-hour blocks and look at how well the model predicts the number of minutes spent in REM, the number of REM bouts, and the median REM bout length—averaged over all 8 mice. We also examine the performance at predicting the number of minutes spent in REM for individual mice. We use the values of these quantities derived from the EEG/EMG manual scores as our target benchmark.

We first consider the number of minutes spent in REM during block j, whose "true value" derived from manual scores we denote by M^j_{REM}. To estimate M^j_{REM}, we sum the raw probability of REM over each of the 720 epochs that make up a 2-hour block (i.e., 2 hours is 7,200 seconds or 720 epochs). That is, we set $\widehat{M}^j_{REM} = \sum_{t \in \text{Block}_j} \lambda^t_{REM}$ where λ^t_{REM} is the model estimate of the probability of REM at epoch t (i.e., $\lambda^t_{REM} = \mathbb{P}(Y_t = REM \mid X, \widehat{\Theta})$). This yields an expected amount of time spent in REM for each 2-hour block.

However, since the raw probabilities λ^t_{REM} are not fully calibrated, we can improve on $\widehat{M}^j_{REM}$ by introducing a threshold tuning parameter θ. In particular, we let $\widehat{M}(\theta)^j_{REM} = \sum_{t \in \text{Block}_j} \theta \cdot \lambda^t_{REM} = \theta \cdot \widehat{M}^j_{REM}$. For low values of θ, the model will tend to underpredict M^j_{REM}, whereas, for high values, it will tend to overpredict it.

This notion is formalized in panel (a) of Figure 4, which gives the root mean square error (RMSE) between $\widehat{M}(\theta)^j_{REM}$ and M^j_{REM} for various values of θ; we also look at the RMSE for the first 12 hours (dark) vs the second 12 (light). As can be seen, the optimal value occurs around $\theta \approx 0.31$, regardless of whether one looks at light, dark, or all blocks. In panel (b) of Figure 4, we plot $\widehat{M}(\theta)^j_{REM}$ averaged across all 8 mice for the optimal value of $\theta = 0.31$. As can be seen, our video-based model's prediction of the amount of time spent in REM sleep quite accurately tracks that based on manual scoring.

The remaining panels of Figure 4 provide additional results for total time spent in REM sleep. In panel (c), we give the difference between the two methods ± 1 standard deviation; as can be seen, all differences lie less than one standard deviation from zero (for full details, see Table S1 of the supplement). In panels (d) and (e) of Figure 4, rather than averaging across all mice, we look at the algorithm's performance on two individual mice. Not surprisingly, the performance on individual mice is not quite as good as when averaged across all mice. Nonetheless, the curve for the video-based method tracks the contours of the curve for the manual scores. Furthermore, the differences between the two curves for individual mice appear calibrated with respect to the standard deviations in panel (c): 67% of the 96 individual 2-hour blocks (i.e., 8 mice × 12 two-hour blocks) are contained within 1 standard deviation and 94% are contained within 2. Nonetheless, this additional variability should be taken into consideration when our method is applied to individual mice.

A final point worth noting is that the amount of REM sleep is small. Fewer than 10 minutes are spent in REM per 2-hour block on average across all mice and in aggregate only about 5% of the time is spent in REM sleep. Furthermore, no single mouse spends more than about 15 minutes in REM in any 2-hour block.

In panels (a) and (b) of Figure 5, we examine how well the model performs at predicting the total number of REM bouts within a given 2-hour block (for full details, see Table S2 of the supplement). We estimate this quantity using an analogue of the threshold procedure used for the number of minutes spent in REM: (1) when $\theta \cdot \lambda^t_{REM}$ is larger than λ^t_{NREM} and λ^t_{WAKE}, we label epoch t a REM epoch; (2) using these labels for the 720 epochs within each 2-hour block, we can calculate the number of distinct bouts of REM. Although the model underpredicts the number of REM bouts, there appear to be no substantial differences between our video-based estimates and those based on manual scoring for any particular block. This is even more encouraging when one considers the fact that we again used $\theta = 0.31$, the value of θ that was optimal for the number of minutes spent in REM. There is no guarantee this value is also optimal for the number of bouts of REM, and, indeed, predictions would likely improve if we were to estimate a different value

of θ specifically for the number of bouts of REM. Nonetheless, doing so would also add an extra parameter to the model.

Finally, panels (c) and (d) of Figure 5 show the performance of the model at forecasting the median REM-bout length (for full details, see Table S3 of the supplement). As for number of REM bouts, we (1) label an epoch as REM when $\theta \cdot \lambda^t_{REM}$ is larger than λ^t_{NREM} and λ^t_{WAKE} and (2) calculate the median bout length using these labels for the 720 epochs in a given block.

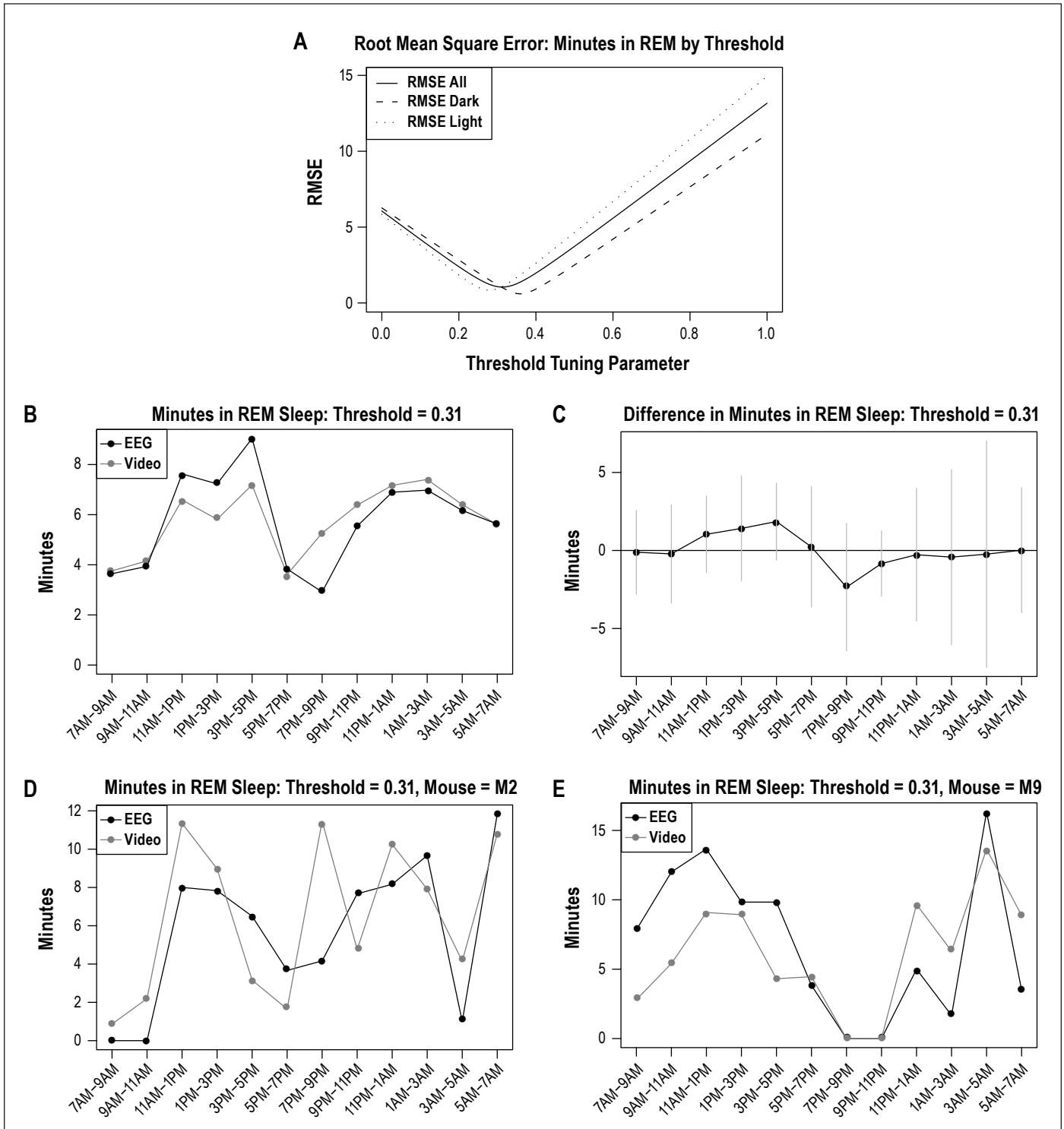

Figure 4—Minutes spent in rapid eye movement (REM) sleep. In panel **(A)**, we plot the root mean square error of our model's prediction of the total time spent in REM sleep averaged across all mice against the various values of the threshold timing parameter. The black line averages across all blocks, the dashed line across "lights out" blocks, and the dotted line across "lights on" blocks. There is a clear minimum at 0.31 and little difference between the minimum for the lights off and lights on period. In panel **(B)**, we give the total time spent in REM sleep averaged across all mice for each two-hour block based on electroencephalography (EEG; black) and video (gray) for the optimal parameter value of 0.31. In panel **(C)**, we give the difference between EEG and video ± 1 standard deviation. In panels **(D)** and **(E)**, we give total time spent in REM sleep for two individual mice for each two-hour block based on EEG (black) and video (gray).

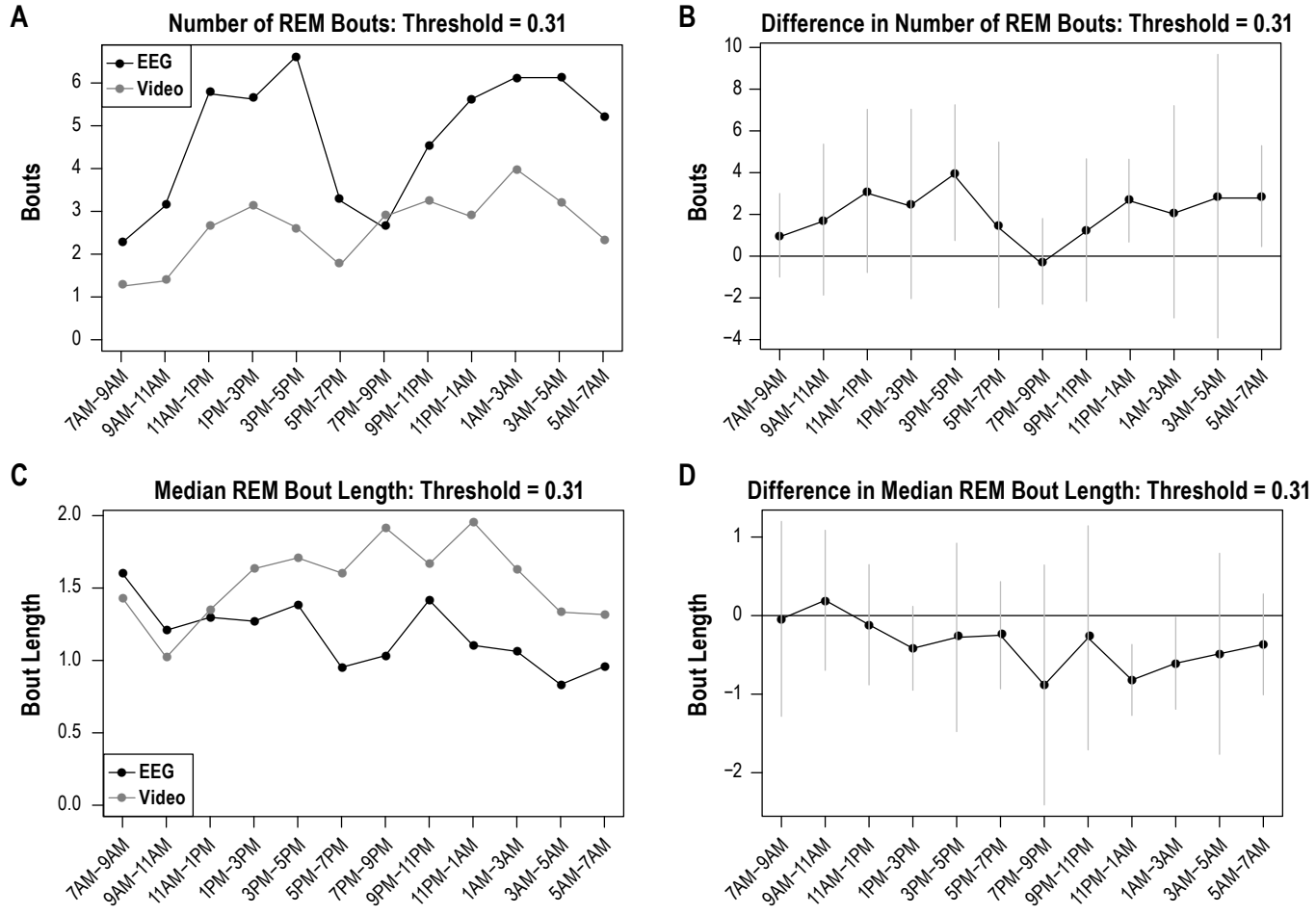

Figure 5—Number of rapid eye movement (REM) bouts and median REM bout length for the optimum threshold of 0.31. In panel **(A)**, we give the number of REM bouts averaged across all mice for each 2-hour block based on electroencephalography (EEG; black) and video (gray). In panel **(B)**, we give the difference between EEG and video ± 1 standard deviation. In panel **(C)**, we give the median REM bout length in minutes averaged across all mice for each 2-hour block based on EEG (black) and video (gray). In panel **(D)**, we give the difference between EEG and video ± 1 standard deviation.

The model consistently overpredicts the median bout length by about 20-30 seconds (2-3 epochs). This is the mirror image of the model's modest underprediction of number of bouts (since number of bouts times median bout length is roughly equivalent to total time spent REM). Again, we used $\theta = 0.31$ here, and, although predictions would likely improve if a value of θ were specifically estimated for the median bout length, doing so would add yet another parameter to the model.

Aggregate Measures of NREM Sleep and Wake

Although our primary focus is how well our model estimates REM sleep, we also provide data on the estimation of both NREM and WAKE amounts. We again do so using the value $\theta = 0.31$ for our threshold tuning parameter. That is, we set $\widehat{M}(\theta)^j_{NREM} = \sum_{t \in \text{Block}_j} \phi \cdot \lambda^t_{NREM}$ and $\widehat{M}(\theta)^j_{WAKE} = \sum_{t \in \text{Block}_j} \phi \cdot \lambda^t_{WAKE}$ where $\phi = \frac{1-\theta\lambda^t_{REM}}{1-\lambda^t_{REM}}$ and $\theta = 0.31$ (normalization by ϕ ensures that the total time spent in all states sums to the proper value of two hours per block).

In Figure 6, we provide the analogue of Figure 4 but for NREM sleep (for full details, see Table S4 of the supplement). Panel (a) gives the total time spent in NREM sleep averaged across all mice based on EEG (black) and video (gray). In panel (b), we give the difference between the two methods ± 1 standard deviation. There are no significant differences between our model and the "truth" as given by EEG/EMG data. In panels (c) and (d) of Figure 6, rather than averaging across all mice, we look at the algorithm's performance for the two individual mice considered in the lower panels of Figure 4. As can be seen, the video-based method tracks the manual scores closely with no major divergences when evaluated both in aggregate across all mice and for individual mice.

In Figure 7, we provide the same plots but for WAKE (for full details, see Table S5 of the supplement). The model's estimates of time spent awake track the manual scores extremely well again, with no major divergences from the manual scores. This again holds both at the aggregate and individual level.

Epoch-by-Epoch Scoring Evaluation

Though our principal focus is on estimating measures of REM sleep—such as time spent in REM, number of REM bouts, and median REM bout duration—we also examined whether our algorithm could replicate the manual scores on an epoch-by-epoch basis. In particular, our epoch-by-epoch scoring evaluation directly compares the performance of 5 different methods: (1)

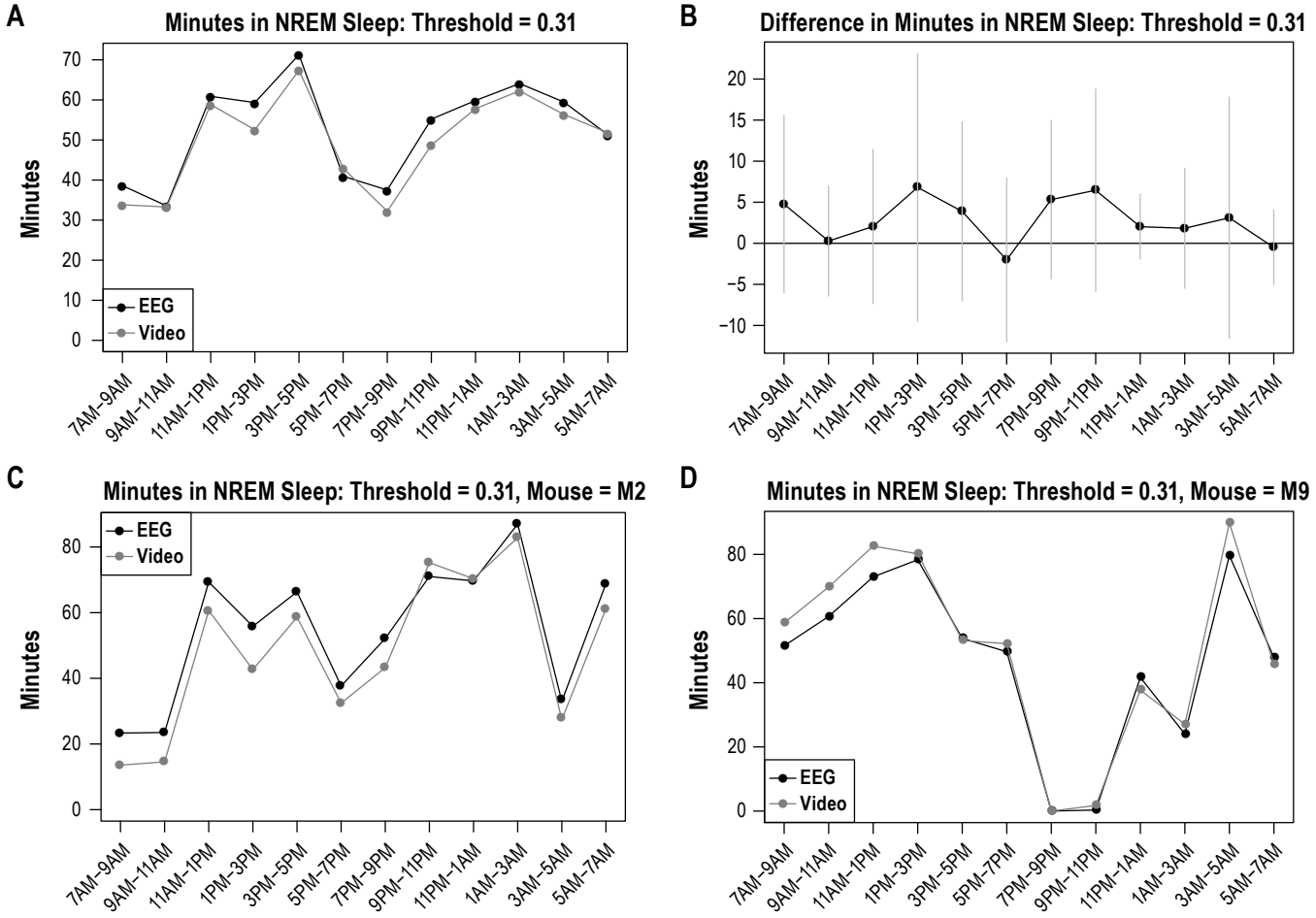

Figure 6—Minutes spent in non-rapid eye movement (NREM) sleep for the optimum threshold of 0.31. In panel **(A)**, we give the total time spent in NREM sleep averaged across all mice for each 2-hour block based on electroencephalography (EEG; black) and video (gray) for the optimum threshold value. In panel **(B)**, we give the difference between EEG and video ± 1 standard deviation. In panels **(C)** and **(D)**, we give total time spent in NREM sleep for two individual mice for each 2-hour block based on EEG (black) and video (gray).

multinomial logistic regression, (2) random forests, (3) random forests combined with a first-order Markov model (RF+1MM), (4) random forests combined with a transition-dependent generalized Markov model (RF+TDGMM), and (5) the so-called "40-second Rule."[4] The fourth method, RF+TDGMM, is our method, which we have been examining thus far. The second and third represent various simplifications of it. Finally, the first and fifth are more common in the literature.

We also examine the "error rate" for the gold standard of manually scored EEGs. In particular, we declare the gold standard to be in error if the two original scorers scored the same epoch differently.

Before proceeding, we note the error rates for four of the five of the methods are completely "out of sample" in the sense that the models are tuned and fit for each mouse and then applied and evaluated on different mice. The only exception is the 40-second Rule. This algorithm considers a mouse "inactive" in a given 10-second epoch if the mean intraepoch velocity is less than 3 pixels per second; it then rules a mouse asleep when there are four or more consecutive inactive epochs. A single parameter (i.e., 40 seconds/4 epochs as opposed to some other multiple of 10 seconds/1 epoch) has been optimized in the sample so as to minimize the error rate with respect to the gold standard. An additional point worth noting about the 40-second Rule is that it can only distinguish sleep from wakefulness, whereas all other methods considered can distinguish among REM, NREM, and wakefulness.

Before trying to discriminate REM from NREM, we first consider the simpler 2-state problem of forecasting SLEEP vs WAKE. The "true" score for an epoch is SLEEP if the manual scorers scored it as REM or NREM, and it is WAKE otherwise (as mentioned earlier, when the two manual scorers disagreed, an independent third scorer was used to break the tie and determine the "truth"). The various classification methods are then trained using this 2-state SLEEP/WAKE score as the response.

We declare an epoch to be in error if a given method classifies the epoch as something other than the "true" score and present the error rates in the second column of Table 1. As can be seen, one can achieve error rates lower than 10%. Although the 40-second Rule performs well, this method can be defeated by models that account for the additional information beyond velocity which is present in the video data. Indeed, the best overall error rate of 8.8% is achieved by our RF+TDGMM method;

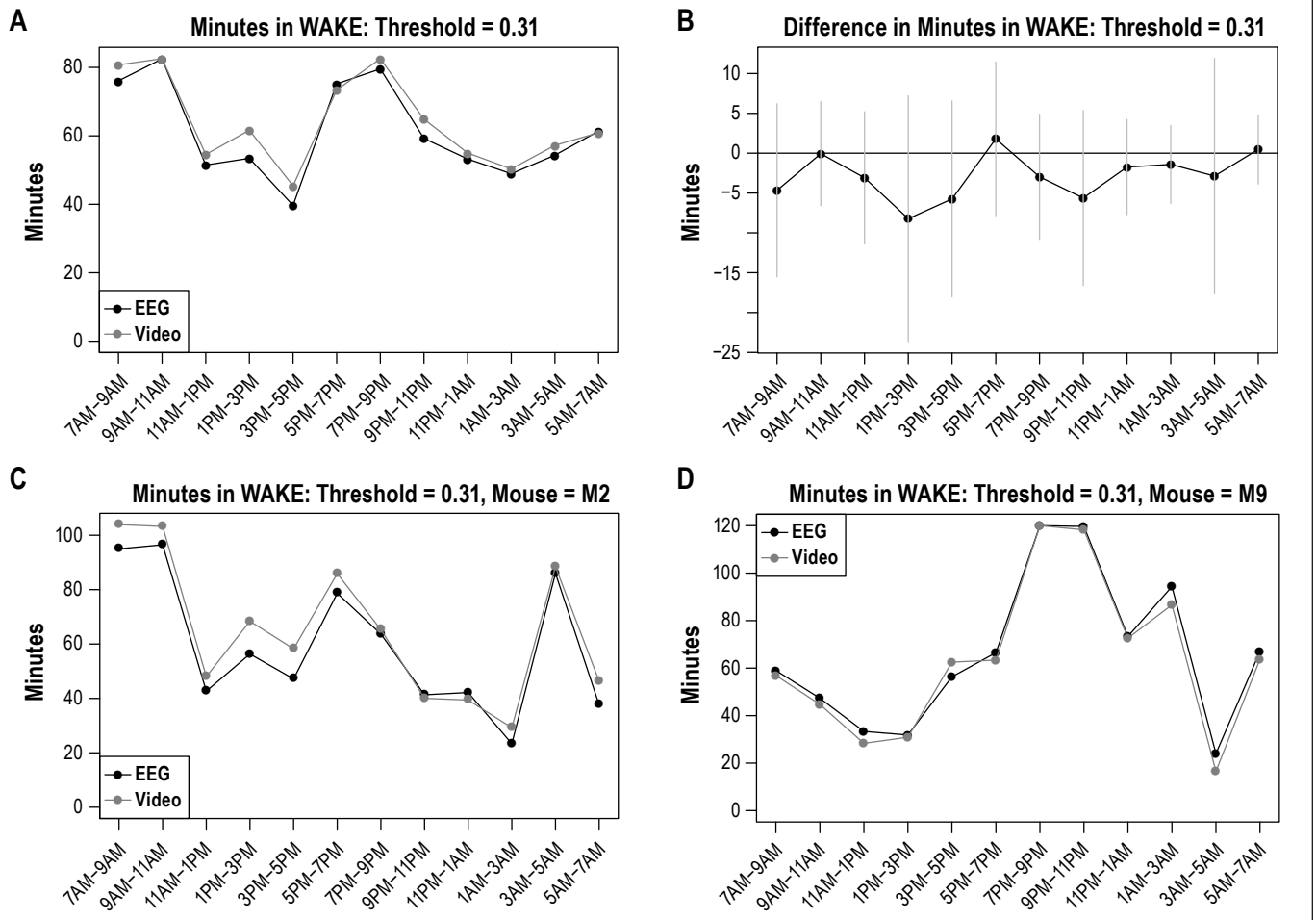

Figure 7—Minutes spent in WAKE for the optimum threshold of 0.31. In panel **(A)**, we give the total time spent in WAKE averaged across all mice for each 2-hour block based on electroencephalography (EEG; black) and video (gray) for the optimum threshold value. In panel **(B)**, we give the difference between EEG and video ± 1 standard deviation. In panels **(C)** and **(D)**, we give total time spent in WAKE for two individual mice for each 2-hour block based on EEG (black) and video (gray).

Table 1—Error rates, in percentage, for various methods

Method	Two-State Error Rate	Three-State		
		Error Rate	REM FP	REM FN
Logistic Regression	9.7	14.9	1.2	95.3
Random Forests	10.4	16.2	1.9	90.9
RF+1MM	8.9	24.7	15.9	53.0
RF+TDGMM	8.8	23.3	14.0	54.2
40-second Rule	10.1	NA	NA	NA
Manual Scores	4.8	5.8	NA	NA

The first column gives the methodology, the second column gives the overall error rate on the two-state SLEEP/WAKE problem, and the third through fifth columns give, respectively, the overall error rate, the rapid eye movement (REM) false positive rate, and the REM false negative rate on the three-state non-REM (NREM)/REM/WAKE problem. RF+1MM denotes the random forest combined with a first-order Markov model whereas RF+TDGMM denotes the random forest combined with the transition-dependent generalized Markov model.

this compares favorably to the 4.8% disagreement rate among manual scorers.

Our second evaluation considers the 3-state problem (i.e., REM vs NREM vs WAKE), and we present our results in the third through fifth columns of Table 1 (since the 40-second Rule can only discriminate sleep from wakefulness but not REM from NREM, it is listed as NA in these columns). This problem is much more difficult for classification methodologies since they now must choose among three alternatives rather than two. Further complicating this difficulty is the fact that the REM occurs only about 5% of the time and looks somewhat similar to NREM in terms of video covariates. Consequently, the overall error rate for each method is higher in the third column vs the second column of the table.

In addition to this overall error rate, which is determined as outlined above, we also consider the false positive and false negative rate for REM, which is of special interest. Again, using the manual scores as "truth" (with ties broken by an independent third scorer when necessary), an epoch is classified as a REM false positive if the classification method declares it to be REM but the manual scorer does not; the REM false positive rate is thus the number of such epochs divided by the total number of epochs declared to be other than REM by manual scoring. Similarly, an epoch is declared to be a REM false negative if the manual scorers score it as REM but the classification

method does not; the REM false negative rate is the number of such epochs divided by the total number of epochs scored as REM by the manual scorers.

The table reveals what is already known: REM is difficult to classify correctly, with REM false positive and false negative rates that are much higher than the overall error rates. The challenge here is to discover a method with any power to detect REM sleep. That is, there is an inherent trade-off between (1) obtaining a low REM false negative rate accompanied by a higher overall and REM false positive rate or (2) obtaining lower overall and REM false positive rates while having a high REM false negative rate. Since high REM false negative rates mean our models have little or no power to detect REM, we prefer to err on the side of (1) rather than (2).

Indeed, logistic regression and random forests have the best overall error rates and low REM false positive rates, but this is because they largely ignore the REM state (i.e., they very rarely classify an epoch as REM) leading to extremely high REM false negative rates. On the other hand, our RF+TDGMM methodology is able to achieve a good balance relative to other methods: it has a REM false negative rate that is much lower than the competitor models while remaining competitive on both the overall error rate and the REM false positive rate. By accounting for the time dependence of the data, our procedure is able to capture a greater proportion of the REM signal. Furthermore, by retaining a reasonable false positive rate relative to the other methods, our model does not sacrifice specificity in order to gain substantial improvements in sensitivity.

In sum, our RF+TDGMM methodology can detect REM sleep in video data. In achieving a lower REM false negative rate (i.e., actually detecting REM), it does have a commensurately higher overall and REM false positive rate as compared with methods such as logistic regression and random forests which tend to ignore the REM state. Finally, as demonstrated in the previous subsections, the RF+TDGMM can be combined with a threshold-tuning parameter to provide accurate assessments of aggregate measures of sleep and wakefulness such as the amount of time spent in REM sleep over 2-hour blocks.

DISCUSSION

In this study, we demonstrate that there is signal in video recordings of mice that is capable of distinguishing NREM from REM sleep. There are subtle changes in the area and shape of the mouse as it transitions from NREM to REM sleep, likely as a result of the atonia of REM sleep. Although REM sleep is a relatively rare state, as compared with NREM and WAKE, our methodology can provide reasonable estimates of it. This new methodology extends previous approaches[4,5] that do not require EEG/EMG recording to now differentiate REM from NREM as opposed to merely SLEEP from WAKE.

This new method has several applications to which it can be applied immediately (i.e., with no further estimation of the model parameters including the threshold parameter). First, in studies in which mRNA changes or protein changes with sleep and wake are being assessed, this approach is much more cost-effective for estimating sleep states. EEG/EMG recording requires surgical implantation of electrodes, time to recover from surgery, and labor-intensive manual scoring of EEG/EMG recordings. There is, moreover, a concern that results at the molecular level could be affected by the recent surgery and the insertion of foreign objects into the mouse skull. A second application of our procedure is for high-throughput phenotyping, something that is increasingly important for studying the large number of knockout mice that are now available.[3]

Currently, the only other automated approach being applied is assessment of mouse behavior by piezoelectric data.[5] This method measures pressure changes in the floor of the mouse cage produced by movement. There are highly variable signals during wakefulness as the mouse moves around; signals during sleep reflect breathing. It is conceivable that the piezo technology could also identify REM sleep because breathing in REM sleep is more irregular than in NREM sleep.[13] At present, however, this possibility has not been assessed.

The sensitivity and specificity of video-based methods to estimate sleep and its substages might be improved if the mouse behavior was observed by video not only from above but also from the side. A 3-dimensional assessment of the mouse using high-resolution video would likely improve assessment of its behavior, including the problem addressed here (i.e., identifying NREM and REM sleep). Such a system would likely be able to determine breathing, as is possible with piezo, as well as the small twitches that occur during REM sleep. Video analysis also provides the opportunity to identify other behaviors, and it is likely that analytic strategies could be developed to study a whole range of mouse behaviors.

In our studies, we used 10-second epochs to score wake and the stages of sleep. We did so because (1) this is the most commonly used epoch length for scoring of behavioral state[4] and (2) the original papers assessing behavioral state by non-EMG/EMG based approaches used this epoch length.[4,5] Behavioral states of wake and sleep can, however, occur in quite short episodes, and, hence, examining in detail the architecture of sleep (bout length) requires scoring in 4-second epochs.[11,14] It is conceivable that if we had used 4-second epochs in this study, we might have found better agreement with bout lengths, etc. Future studies need to assess the impact of different epoch lengths on agreement between video and EEG analysis.

Epoch lengths aside, there are several differences between the results for time spent in REM sleep on one hand and number of bouts, bout length, and epoch-by-epoch scores on the other hand. First, there was a methodological choice: since our focus was on the time spent in REM, we tuned our parameter θ to that quantity, whereas for the latter we either fixed it at the value that was optimal for time spent in REM (number of bouts and bout length) or at the default of one (epoch-by-epoch scoring). Future studies focusing on these metrics should consider tuning our method's predictions specifically for them. Second, beyond modeling choices, there are fundamental differences between time spent in REM and the other metrics. The expected time spent in REM does not require the conversion of a model's probabilities for each state at each epoch into a predicted sleep state for that epoch; rather, these probabilities can be summed across all epochs, yielding the expected time in the state. On the other hand, computing epoch-by-epoch scores, number of bouts, and bout length requires the conversion of these probabilities into sleep scores on an epoch-by-epoch basis. This fundamental difference underlies the varying results observed. Finally, there is a third difference that applies

to number of bouts and bout length. When long stretches of REM are briefly interrupted (e.g., an epoch or two of NREM or WAKE surrounded on either side by many epochs of REM), the model's estimates of number of bouts and bout length—assuming it cannot detect these brief interruptions—will be strongly negatively impacted whereas there will be little impact on time spent in REM. Despite this major difference, our method is still quite competitive at estimating these more difficult quantities.

The application of the video methodology will be in uninstrumented mice (i.e., mice without EEG/EMG headstages). It is conceivable that the changes in shape (i.e., aspect ratio) and area as the mouse transitions from NREM to REM sleep are sufficiently different in uninstrumented mice that the model presented here (which was fit to data from instrumented mice) will be inaccurate on uninstrumented mice. We believe that this is unlikely for two reasons. First, the cable connected to the mouse's head is carefully counterbalanced so that the mouse moves freely and there is no excessive force on the head; thus, it seems unlikely that the cable will result in different changes in shape and area as the mouse becomes more atonic in REM sleep. Second, it is the *changes* in aspect ratio and area that are most important for differentiating NREM from REM sleep; the absolute magnitudes of these variables are of secondary importance, and, indeed, they vary from mouse to mouse. Although it seems that the need for instrumentation will therefore not affect the accuracy of our approach, the question is ultimately unanswerable, since validation requires EEG/EMG recordings; such recordings, in turn, require some form of instrumentation, whether by the methodology used here or by telemetry (which also could potentially alter mouse shape and area).

In conclusion, this study shows that video analysis can distinguish REM from NREM sleep in mice. Future elaborations of this technological approach could lead to further improvements in these estimates. Thus, high-throughput phenotyping of sleep and wake in mice is feasible and will facilitate studies of the role of specific genes using the large number of mice with knockout of specific genes that are more available[3] and investigation of chemical libraries to determine compounds that affect sleep and wake, as has been done in zebra fish.[15]

ACKNOWLEDGMENTS

The authors thank Mr. Daniel Barrett for assistance in preparation of this manuscript. This research was supported by NIH grants T32 HL07713, P01 AG17628, and MH081491.

DISCLOSURE STATEMENT

This was not an industry supported study. Funds for the endowment of Dr. Pack's professorship (John Miclot Professorship) are provided by the Phillips/Respironics Foundation. Dr. Biber is the Co-President of Neurocare, Inc. and has received compensation in excess of $10,000 per annum; in this role, he has served as Principal Investigator of several clinical trials funded by pharmaceutical companies which have compensated Neurocare, Inc. though not Dr. Biber directly for this work. The other authors have indicated no financial conflicts of interest.

REFERENCES

1. Sehgal A, Mignot E. Genetics of sleep and sleep disorders. Cell 2011;146:194-207.
2. Mackiewicz M, Zimmerman JE, Shockley KR, et al. What are microarrays teaching us about sleep? Trends Mol Med 2009;15:79-87.
3. Guan C, Ye C, Yand X, et al. A review of current large-scale mouse knockout efforts. Genesis 2010;48:73-85.
4. Pack AI, Galante RJ, Maislin G, et al. Novel method for high-throughput phenotyping of sleep in mice. Physiol Genomics 2007;28:232-8.
5. Flores AE, Flores JE, Deshpande H, et al. Pattern recognition of sleep in rodents using piezoelectric signals generated by gross body movements. IEEE Trans Biomed Eng 2007;54:225-33.
6. Naidoo N, Ferber M, Master M, Zhu Y, Pack AI. Aging impairs the unfolded protein response to sleep deprivation and leads to proapoptotic signaling. J Neurosci 2008;28:6539-48.
7. Breiman L. Random forests. Machine Learning 2001;45:5-32.
8. Breiman L, Friedman J, Olshen R, Stone C. Classification and Regression Trees. New York, NY: Wadsworth; 1984.
9. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE. 1989;77:257-86.
10. Smyth P. Markov monitoring with unknown states. IEEE Journal of Selected Areas in Communications, Special Issue on Intelligent Signal Processing for Communications. 1994;12:1600-12.
11. McShane BB, Galante RJ, Jensen ST, Naidoo N, Pack AI, Wyner A. Characterization of the bout durations of sleep and wakefulness. J Neurosci Methods 2010;193:321-33.
12. McShane BB. Machine learning methods with time series dependence. The Wharton School of the University of Pennsylvania; 2010.
13. Friedman L, Haines A, Klann K, et al. Ventilatory behavior during sleep among A/J and C57BL/6J mouse strains. J Appl Physiol 2004;97:1787-95.
14. Franken P, Malafosse A, Tafti M. Genetic determinants of sleep regulation in inbred mice. Sleep 1999;22:155-69.
15. Rihel J, Prober DA, Arvanites A, et al. Zebrafish behavioral profiling links drugs to biological targets and rest/wake regulation. Science. 2009;327:348-51.

Supplement to "Assessing REM Sleep in Mice Using Video Data"

In this supplement, we present tables detailing the data plotted in Figures 4a and 4b (Table S1), Figures 5a and 5b (Table S2), Figures 5c and 5d (Table S3), Figures 6a and 6b (Table S4), and Figures 7a and 7b (Table S5) of the main text.

In Table S1, we present the mean and standard deviation of the number of minutes spent in REM sleep in each two hour block (column one) across all eight mice for both EEG/EMG manual scores (columns two and three) and for our video-based model (columns four and five). We also give the mean and standard deviation of the difference between manual scores and the model across all eight mice (columns six and seven).

Table S2 presents the same information as Table S1 but for the number of REM bouts rather than the number of minutes spent in REM sleep. Table S3 presents the same information but for the median REM bout length. Table S4 presents the same information but for the number of minutes spent in NREM sleep. Table S5 presents the same information but for the number of minutes spent in WAKE.

Table S1: Minutes Spent in REM Sleep for the Optimum Threshold $\theta = 0.31$.

Block	EEG		Video		Difference	
	Mean	SD	Mean	SD	Mean	SD
7AM - 9AM	3.65	4.12	3.76	4.03	-0.11	2.68
9AM - 11AM	3.94	4.82	4.15	4.84	-0.22	3.16
11AM - 1PM	7.63	4.94	6.59	4.91	1.04	2.45
1PM - 3PM	7.23	2.58	5.82	2.75	1.41	3.36
3PM - 5PM	9.04	2.61	7.19	4.09	1.85	2.47
5PM - 7PM	3.83	3.18	3.61	2.39	0.22	3.86
7PM - 9PM	2.92	2.95	5.27	5.43	-2.35	4.08
9PM - 11PM	5.54	3.11	6.38	4.07	-0.84	2.09
11PM - 1AM	6.90	2.99	7.17	5.17	-0.27	4.26
1AM - 3AM	6.98	3.33	7.41	3.81	-0.43	5.61
3AM - 5AM	6.17	5.25	6.40	5.95	-0.23	7.26
5AM - 7AM	5.63	4.74	5.61	4.16	0.02	4.01

We give the total time spent in REM sleep averaged across all mice for each two hour block for EEG and video assessments. We also give the standard deviation of each, their difference, and the standard deviation of the difference.

Table S2: Number of REM Bouts for Threshold $\theta = 0.31$.

Block	EEG		Video		Difference	
	Mean	SD	Mean	SD	Mean	SD
7AM - 9AM	2.25	2.43	1.25	1.04	1.00	2.00
9AM - 11AM	3.13	4.02	1.38	2.13	1.75	3.62
11AM - 1PM	5.75	3.28	2.63	4.27	3.13	3.91
1PM - 3PM	5.63	1.92	3.13	3.64	2.50	4.54
3PM - 5PM	6.63	3.42	2.63	3.66	4.00	3.25
5PM - 7PM	3.25	2.25	1.75	2.55	1.50	3.96
7PM - 9PM	2.63	2.62	2.88	3.23	-0.25	2.05
9PM - 11PM	4.50	3.21	3.25	3.37	1.25	3.41
11PM - 1AM	5.63	1.92	2.88	2.95	2.75	1.98
1AM - 3AM	6.13	2.80	4.00	3.34	2.13	5.08
3AM - 5AM	6.13	3.91	3.25	3.54	2.88	6.79
5AM - 7AM	5.25	4.33	2.38	3.02	2.88	2.42

We give the number of REM bouts averaged across all mice for each two hour block for EEG and video assessments. We also give the standard deviation of each, their difference, and the standard deviation of the difference.

Table S3: Median REM Bout Length for Threshold $\theta = 0.31$.

Block	EEG		Video		Difference	
	Mean	SD	Mean	SD	Mean	SD
7AM - 9AM	1.60	0.27	1.43	0.84	-0.04	1.24
9AM - 11AM	1.21	0.42	1.02	0.37	0.19	0.89
11AM - 1PM	1.30	0.43	1.35	0.61	-0.12	0.77
1PM - 3PM	1.27	0.22	1.63	0.45	-0.42	0.53
3PM - 5PM	1.39	0.64	1.71	0.65	-0.28	1.20
5PM - 7PM	0.95	0.30	1.60	0.72	-0.25	0.68
7PM - 9PM	1.03	0.41	1.92	1.30	-0.88	1.53
9PM - 11PM	1.42	0.40	1.67	1.49	-0.28	1.43
11PM - 1AM	1.10	0.16	1.96	0.35	-0.82	0.45
1AM - 3AM	1.06	0.29	1.63	0.62	-0.61	0.58
3AM - 5AM	0.83	0.29	1.33	1.06	-0.49	1.28
5AM - 7AM	0.96	0.62	1.32	0.32	-0.37	0.64

We give the median REM bout length in minutes averaged across all mice for each two hour block for EEG and video assessments. We also give the standard deviation of each, their difference, and the standard deviation of the difference.

Table S4: Minutes Spent in NREM Sleep for Threshold $\theta = 0.31$.

Block	EEG		Video		Difference	
	Mean	SD	Mean	SD	Mean	SD
7AM - 9AM	38.60	17.49	33.83	23.22	4.77	10.80
9AM - 11AM	33.56	29.49	33.28	32.71	0.28	6.70
11AM - 1PM	60.90	25.95	58.85	32.88	2.05	9.39
1PM - 3PM	59.33	15.41	52.52	24.64	6.81	16.30
3PM - 5PM	71.29	15.40	67.40	21.72	3.89	10.93
5PM - 7PM	40.96	18.22	42.98	17.63	-2.02	9.95
7PM - 9PM	37.52	27.86	32.20	24.25	5.32	9.66
9PM - 11PM	55.21	29.25	48.75	29.52	6.46	12.33
11PM - 1AM	59.85	24.03	57.83	25.78	2.02	3.95
1AM - 3AM	64.08	21.27	62.25	21.59	1.83	7.31
3AM - 5AM	59.48	18.10	56.37	27.37	3.11	14.70
5AM - 7AM	51.31	27.64	51.81	27.52	-0.50	4.57

We give the total time spent in NREM sleep averaged across all mice for each two hour block for EEG and video assessments. We also give the standard deviation of each, their difference, and the standard deviation of the difference.

Table S5: Minutes Spent in WAKE Sleep for Threshold $\theta = 0.31$.

Block	EEG		Video		Difference	
	Mean	SD	Mean	SD	Mean	SD
7AM - 9AM	76.08	21.27	80.74	26.62	-4.66	10.87
9AM - 11AM	82.50	33.80	82.56	37.07	-0.06	6.53
11AM - 1PM	51.48	30.45	54.56	37.23	-3.09	8.29
1PM - 3PM	53.44	17.31	61.66	24.77	-8.22	15.43
3PM - 5PM	39.67	16.97	45.41	24.51	-5.74	12.31
5PM - 7PM	75.21	20.36	73.41	18.49	1.80	9.66
7PM - 9PM	79.56	30.50	82.53	28.09	-2.97	7.86
9PM - 11PM	59.25	32.21	64.87	32.40	-5.62	10.99
11PM - 1AM	53.25	26.81	55.00	28.93	-1.75	5.98
1AM - 3AM	48.94	24.04	50.34	21.32	-1.40	4.89
3AM - 5AM	54.35	21.27	57.23	30.86	-2.87	14.76
5AM - 7AM	61.40	31.13	60.91	29.91	0.48	4.35

We give the total time spent in WAKE averaged across all mice for each two hour block for EEG and video assessments. We also give the standard deviation of each, their difference, and the standard deviation of the difference.