Automated detection of sleep-boundary times using wrist-worn accelerometry

Johanna O'Donnell[1,2*], Sven Hollowell[3], Gholamreza Salimi-Khorshidi[2], Carmelo Velardo[1], Claire Sexton[4], Kazem Rahimi[2], Heidi Johansen-Berg[4], Lionel Tarassenko[1‡], Aiden Doherty[1,3‡*]

1 Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, United Kingdom
2 George Institute for Global Health, Nuffield Department of Obstetrics and Gynaecology, University of Oxford, Oxford, United Kingdom
3 Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom
4 Nuffield Department of Neuroscience, University of Oxford, Oxford, United Kingdom

‡These authors are joint last authors.
¤Current Address: Nuffield Department of Population Health, Richard Doll Building, Old Road Campus, Headington, Oxford OX3 7LF, United Kingdom
* johanna.ernst@st-annes.ox.ac.uk, aiden.doherty@ndph.ox.ac.uk

Abstract

Objective
Current polysomnography-validated measures of sleep status from wrist-worn accelerometers cannot be used in fully automated analysis as they rely on self-reported sleep-onset and –end (sleep-boundary) information. We set out to develop an automated, data-driven approach to sleep-boundary detection from wrist-worn accelerometer data.

Methods
On three separate occasions, participants were asked to wear a GENEActiv® wrist-worn accelerometer for nine days and concurrently complete sleep diaries with lights-off, asleep and wake-up information. We developed and evaluated three data-driven methods for sleep-boundary detection: a change-point detection based method , a thresholding method and a random forest classifier based method. Mean absolute errors between automatically-derived and self-reported sleep-onset and wake-up times were recorded in addition to kappa statistics for the minute-by-minute performance of each of the methods.

Results
46 participants provided 972 days of accelerometer recordings with corresponding self-reported sleep information. The three sleep-boundary detection methods resulted in mean absolute errors in sleep-onset and wake-up times per individual of 36 min, 34 min and 33 min and kappa statistics of 0.87, 0.89 and 0.89, respectively.

Conclusion
Our methods provide a data-driven approach to detect sleep-onset and -end times without the need for self-reported sleep-boundary information. The methods are likely

to be of particular use for large-scale studies where the collection of self-reported sleep diaries is impractical.

Significance
Objective measures of sleep are needed to reliably detect associations with health outcomes. This work lays the foundation for studies of objectively measured sleep duration and its health consequences in large studies.

Introduction

Alterations in sleep duration and changes in sleep-wake timing are associated with a wide range of negative health outcomes, including an increased risk of type 2 diabetes and cardiovascular outcomes [1–3] as well as increased incidence of psychiatric disorders [4]. However, evidence often relies on self-reported sleep information, which may be unreliable and affected by measurement errors due to memory bias [5]. As a result, many studies now use wrist-worn accelerometers in an attempt to objectively measure sleep durations [6, 7]. Publicly available sleep detection methods developed for wrist-worn accelerometers include the idleness-detecting method by *Borazio et al* [8] and the angular-movement based method by *van Hees et al* [6]. Methods such as these have traditionally been developed in a laboratory environment and tested against the gold standard for objective sleep analysis, polysomnography.

However, current polysomnography-validated measures of sleep status from wrist-worn accelerometers cannot be used in fully automated analysis as they rely on self-reported sleep-onset and –end (sleep-boundary) information [6]. This information is typically collected through means of a sleep diary [6] or through a button on the accelerometer which participants can press as they go to bed and wake up [5]. As a result it is not feasible to collected self-reported sleep-boundary times for all participants in large scale studies, such as the UK Biobank [7], and therefore a need exists for an automated way to extract sleep-onset and wake up times.

In this paper we set out to develop a fully-automated method for sleep-boundary detection from wrist-worn accelerometer data. We also set out to assess its performance in free-living scenarios in healthy UK adults aged 60-80.

Materials and methods

Study Population

Accelerometer recordings and self-reported sleep-boundary data were collected as part of a Cognitive Health in Ageing (CHA) Exercise study [9]. The purpose of this study was to analyse the effect of an anaerobic exercise intervention on brain MRI measures. Adults between the ages of 60 and 80 years that self-reported less than 60 minutes of heart-raising physical activity per week and no contraindications to MRI scanning or fitness testing were eligible to take part in the research study. Ethical Approval for this study was obtained from the Local Research Ethics Committee (Oxford REC B Ref 10/H0605/48).

Study Procedure

Participants wore the GENEActiv® accelerometer on their non-dominant wrist for nine consecutive days in weeks 0, 12 and 24 of the study and filled in daily sleep diaries with lights-off, asleep and wake-up information during these periods. The GENEActiv® sensor contains a tri-axial accelerometer, with a sensor resolution of +-8g, as well as a temperature sensor and has previously been validated for sleep research [10] [6]. Acceleration data were collected at a sampling frequency of 87.5 Hz.

Data Preparation

Accelerometer recordings with at least one day of corresponding asleep and wake-up labels were selected for the analysis. Accelerometer data followed careful UK Biobank preprocessing and quality control checks [7], including device calibration to local gravity [11], removal of machine noise and gravity [12], and imputation of non-wear data segments using the average of similar time-of-day vector magnitude data points from different days of the measurement. Device non-wear time was automatically identified as consecutive stationary episodes lasting for at least 60 minutes [7].

Sleep-boundary Detection

In this study, we employ three approaches for automatic sleep-boundary detection: (1) A statistical technique for detecting change points in the accelerometer time series, (2) A data-driven thresholding method to classify short intervals into sleep and wake and (3) A machine learning technique to classify short intervals into sleep and awake. All methods are illustrated on a real time series in Fig 1.

Fig 1. Processing steps. Visual representations automated detection of sleep-boundary times as described in this paper. (1) refers to the change-point detection based method, (2) refers to the threshold based method and (3) to the random forest based method.

Sleep-labeling using change-point detection

Accelerometer signals during wakefulness contain more severe and frequent changes in acceleration than during sleep, resulting in differences in both the mean and variance. For the purpose of this analysis the standard deviation of acceleration along the x-, y- and z-axes were calculated over one-minute epochs and combined as the vector magnitude of standard deviations, *s_vmag*.

The change-point detection technique developed by *Killick et al* [13], implemented in R package *'changepoint'* [14] was employed to identify changes in the mean or variance of *s_vmag*. This technique is based on maximum log-likelihood estimations (MLE), which provide a measure of the integrity/agreement within a signal. If the sum of the MLEs of the segments of a signal prior to and after a time n is larger than the MLE of the combined signal, n is marked as a change point within the signal. In order to identify multiple change points within the signal, a binary-segmentation approach was employed, whereby the dominant change point was identified first and either side of the change point was consecutively scanned for additional change points. This search stops when no more change points can be identified.

Whilst difference in acceleration between sleep and wake stages are expected to create dominant change points, there may also be changes in accelerometer patterns

throughout the day or night that result in false positive change points, e.g. a subject may show predominantly active behaviours in the morning and then become more sedentary in the afternoon (Fig 2 A). In order to improve the performance of the change-point-detection based sleep classification method, a threshold (τ_{CPD}) was included post change-point detection. Whenever the mean *s_vmag* of a section identified by the change-point-detection method fell above τ_{CPD}, the section was classified as wakefulness, whenever the mean *s_vmag* fell below τ_{CPD}, the section was classified as sleep (Fig 2 B).

Fig 2. Sleep-labeling using change-point detection. A: Visualisation of false positive change point showing accelerometer signal (grey line) and mean *s_vmag* values for segments identified using change-point detection (orange dashed line). B: Accelerometer signal (grey line) and sleep (0) and wake (1) classes (orange dashed line) after thresholding.

In order to find the optimal value of τ_{CPD}, participants were randomly assigned to a training or testing set following a 70:30 split. Thresholds between 20 mg and 150 mg were scanned in increments of 10 mg (limits were chosen based on minimum and maximum crossing points between the probability density functions of standard deviations during self-reported sleep and wakefulness). The threshold that resulted in the smallest mean absolute error (MAE) between automatically-derived and self-reported sleep-onset and wake-up times on the training set was chosen for further evaluation on the testing set.

Sleep-labeling using thresholding

As an alternative method, a data-driven threshold was used to classify each minute of the accelerometer recording as 'sleep' or 'wakefulness'. The threshold was found using probability-adjusted crossing-points of probability density functions (PDFs) during wake and sleep times (see example in Figure 3). In order to find the optimal threshold ($\tau_{i,j}$) for each recording in the training set, the PDFs of sleep and wake times were adjusted by their relative probabilities-of-state. Assuming an average amount of eight hours sleep a night for 24 hours worth of data, a wake-to-sleep ratio of 2:1 was used. The intersections between the adjusted wake and sleep PDFs were found and the corresponding s_{vmag} values provided $\tau_{i,j}$. The PDF-based thresholds were averaged across each individual (j) and consecutively across all participants (i) in the training set in order to find the final threshold τ_{PDF}. For the PDF-based sleep labels, all values of s_{vmag} above τ_{PDF} were classified as wake (1) and all others as sleep (0) (Equation 2).

$$\tau_{PDF} = \frac{1}{N}\sum_{i=0}^{I}\frac{1}{J}\sum_{j=0}^{J_i}\tau_{i,j}, \tag{1}$$

where I is the total number of participants in the training set and J_i is the number of recordings for participant i.

$$state_n = \begin{cases} 1, & \text{if } s\_vmag_n \geq \tau_{PDF} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Fig 3. Extraction of PDF-based thresholds for sleep detection.

Sleep-labeling using a random forest classifier

Finally, the random forest classifier developed by *Breiman* [15] and implemented in the Python package *scikit-learn* [16] was employed to classify each minute of the accelerometer recording as 'sleep' or 'wakefulness'. Random forests are bagging classifiers constructed of multiple de-correlated decision trees. Each tree is itself a classifier that is fed a random subset of the dataset. During training of the random forest, the combined trees act as a form of cross-validator, reducing the classification variance due to noise within individual classifiers.

The input to the random-forest classifier was an array of 42 time- and frequency-domain features, containing single-axis features computed using the vector magnitude of acceleration (*s_vmag*) as well as cross-axial features (Table 1) [17]; extracted across one-minute epochs. In order to study the impact of age and sex on classification performance, the random-forest classifier was trained twice, once including solely accelerometer-derived features and once including accelerometer-derived features as well as age and sex information.

Table 1. Overview of features used for random forest classification.

Single-axis Features	Multi-axes Features
Mean	Pearson correlation
Standard deviation	Covariance
Coeffiecient of variation	
Skewness	Mean roll
Kurtosis	Mean pitch
Median	Mean yaw
Minimum	Standard deviation of roll
Maximum	Standard deviation of pitch
Power contained in 1-15 HZ	Standard deviation of yaw
Dominant frequency	
Power in dominant frequency	
Entropy	

In order to find the optimal parameters for the random-forest classifier, the training set was split into a training (*rf_train*) and a testing (*rf_test*) set (split ratio 70:30). The classifier was trained on *rf_train* and tested on *rf_test*. Area under the curve (AUC) values for numbers of trees between 50 and 300 and minimum number of samples per leaf of 10 to 1000 were recorded for the *rf_test* set. The classification settings with the largest associated AUC were selected for further analysis.

Post-processing using a smoothing filter

Short periods of wakefulness during sleep were sometimes classified erroneously as being the start or end of sleep. Smoothing was used to reduce this using a rolling mean-filter of length N (shown in Fig 1). N is the number of minutes to smooth over, and was learned from the training set by minimising the Mean absolute errors (MAE) between self-reported and automatically-derived sleep-onset and wake-up times post filtering. The MAE was calculated as shown in Eq 3, where $tan_{i,k}$ is the self-reported and $talg_{i,k}$

is the algorithm-derived sleep-onset or -end time, K is the combined number of sleep-onset and wake-up times for a specific individual and I the number of individuals in the training set.

$$\text{MAE} = \frac{1}{I} * \sum_{i=0}^{I} \frac{1}{K} * \sum_{k=0}^{K} |(tan_{i,k} - talg_{i,k})| \tag{3}$$

Finally, the start and end times of the M largest sleep periods were selected as the sleep-boundary times identified by the change-point detection method, where M was the number of days (defined as number of 4:00 am instances) contained within a recording.

The code developed as part of this work is available at https://github.com/activityMonitoring. (From time of publication)

Statistical Analysis

Outcome statistics were recorded for the previously unseen test set. MAE between automatically-derived and self-reported sleep-onset and wake-up times were used as the primary outcome measure. In addition to this, kappa statistics [18] (Eq 4) between the automatically-derived and self-reported minute-by-minute labels were calculated. Kappa statistics reflect the inter-rater agreement between two sources taking into account both the observed agreement (P_o, also known as accuracy) and the likelihood of them agreeing by chance (P_e). Finally, Bland-Altman [19] plots were used to visualise the agreement between the automatically-derived and self-reported sleep-boundary times.

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{4}$$

The performance of all sleep-detection algorithms was evaluated on a per person (*individual*) and a per night (*episode*) basis. For the *individual* analysis, mean values across all recording incidences and all nights were extracted.

Results

Study Population

Out of the 51 participants taking part in the CHA Exercise study, 46 had at least one day worth of accelerometer recordings with corresponding self-reported sleep information. A total of 972 days of accelerometer recordings were included in the analysis. The mean age of all included participants was 66.7 (standard deviation: 5.4) years. 29 (63%) participants were female, 17 (37%) male. Table 2 provides an overview of participant characteristics and self-reported data grouped by tertiles of total sleep duration (tertile boundaries: 0h-7.3h, 7.4h-8.0h, 8.1h-9.2h).

Sleep-boundary detection

The threshold to differentiate between sleep and wake episodes identified by the change-point-detection method, τ_{CPD}, learned using the training set was 0.192g. The

Table 2. Baseline characteristics and self-reported sleep information of participants by tertiles of self-reported total sleep time.

	Tertile 1 (0-7.3h)	Tertile 2 (7.4-8h)	Tertile 3 (8.1-9.2h)
n	15	15	16
Age (years)	66.5 (5.0)	67.9 (6.2)	65.7 (5.2)
% female	60	67	56
Self-reported days	21.9 (5.8)	20.1 (8.0)	20.0 (5.4)
Total sleep time (h)	6.5 (1.0)	7.7 (0.2)	8.4 (0.4)

ideal filter length, N, learned using the training set was 60 minutes. The performance of the change-point detection based sleep-boundary detection on the testing set is summarised in Table 3. The method achieved a MAE of 36 min and a kappa statistic of 0.87 in the *individual* analysis. Figures 4.a-b show the Bland-Altman plots of sleep-onset and -end times.

The optimal data-driven threshold (τ_{PDF}) to classify each minute of the accelerometer recording as 'sleep' or 'wakefulness' was found to be 0.074g. The corresponding filter length N for the thresholding method was found to be 70 minutes. Using these settings the thresholding method achieved a MAE of 34 min and a kappa statistic of 0.89. Figures 4.c-d show the Bland-Altman plots of sleep-onset and -end times.

The feature importance learned by the random forest classifier excluding age and sex information is shown in Figure S1 Fig in the Appendix. Median, 75th percentile and 25th percentile were identified as the most important features. The choice of maximum number of trees and minimum numbers of leafs only minimally affected the results; the combination of 10 trees and a minimum of 10 leafs resulted in an area under the curve (AUC) value for minute-by-minute sleep classification of 0.95. The ideal filter length (N) was found to be 50 minutes, achieving a MAE of 33 minutes and a kappa statistic of 0.89 in the *individual* analysis (see Table 3).

Including age and sex information as features in the classification did not improve the method's performance, resulting in a MAE of 32.1 (40.1) and a kappa statistic of 0.89. The feature importance learned by the classifier including age and sex information is shown in Figure S2 Fig in the Appendix.

The Bland-Altman plots for both the change-point detection and random forest based classification are shown in Figure 4. They show on average good agreement between self-reported and automatically-derived sleep-boundary times. There are however outliers with more than 100 minutes difference between self-reported and automatically-derived sleep boundary times. On closer inspection, these outliers include potential errors in self-reported sleep-boundary times (mislabeling) as well as miss-classifications of sleep and wake labels. Miss-classifications occur primarily where low-activity wake times and sleep times are close in proximity and when sleep-interruptions occur in close proximity to sleep-boundary times.

Discussion

To our knowledge, this is the first study attempting a data-driven approach to automated sleep-boundary detection. Previous studies have used self-reported sleep

Table 3. Performance summary: Mean absolute error (MAE) and kappa statistic between automatically-derived and self-reported sleep-onset and wake-up times.

	Change-point detection (minutes)	Thresholding (minutes)	Random forest (minutes)
Individual (n=14)			
MAE combined (SD)	35.9 (14.8)	34.4 (9.9)	33.3 (11.3)
MAE wake (SD)	20.3 (20.3)	14.9 (9.9)	14.6 (10.5)
MAE sleep (SD)	20.3 (19.5)	17.6 (20.2)	14.7 (15.9)
MAE TST (SD)	31.7 (32.0)	25.7 (18.0)	19.3 (14.7)
Kappa statistic	0.87	0.89	0.89
Episode (n=264)			
MAE combined (SD)	31.8 (54.8)	32.8 (38.0)	31.1 (41.5)
MAE wake (SD)	26.3 (43.4)	30.2 (33.3)	27.8 (34.3)
MAE sleep (SD)	37.4 (47.4)	35.4 (42.0)	34.5 (47.3)
MAE TST (SD)	36.4 (45.9)	39.2 (37.8)	36.7 (40.6)
Kappa statistic	0.89	0.89	0.90

Fig 4. Bland-Altman plots of agreement between automatically-derived (t_alg) and self-reported (t_an) sleep-onset and wake-up times on an individual level. A: Change-point detection sleep-onset times. B: Change-point detection wake-up times. C: Thresholding sleep-onset times. D: Thresholding wake-up times. E: Random forest sleep-onset times. F: Random forest wake-up times.

Fig 5. Bland-Altman plots of agreement between automatically-derived (t_alg) and self-reported (t_an) sleep-onset and wake-up times on an episode level. A: Change-point detection sleep-onset times. B: Change-point detection wake-up times. C: Thresholding sleep-onset times. D: Thresholding wake-up times. E: Random forest sleep-onset times. F: Random forest wake-up times.

times [6] or restricted their analysis to night-time assessments [20]. As a result, current accelerometer studies are reliant on the collection of self-reported sleep-boundary information. Where this information is missing, analysis of sleep and wake behaviour of participants is severely limited. The proposed sleep-boundary detection methods allow for the identification of sleep onset and end times in the absence of self-reported data and open the door to large-scale fully-automated accelerometer analysis.

The methods were trained using self-reported sleep and wake information as their ground truth. Self-reported sleep times have been shown to be less accurate than objectively collected sleep-onset and end-times [5] and may be seen to provide a weaker ground truth than polysomnography assessments. Collecting free-living rather than laboratory data does however provide a more realistic setting for sleep analysis, where recordings may be affected by periods of rest prior to sleep and after wake-up that may otherwise be missed. In addition to this, allowing participants to stay in their natural sleeping environment reduces the impact of data collection on their sleeping behaviour [21].

Previous research by *Lauderdale et al* has shown that self-reported sleep information can be affected by mis-reporting [5]. Whilst some differences may be explained by errors in self-reporting [5], others are likely to be caused by miss-classifications. Such

miss-classifications have in the past been reported in cases where sleep episodes were followed by restful wake periods [22].

The algorithms described as part of this work have been developed using raw accelerometer data as opposed to proprietary count-values. Nevertheless, device- (e.g. on-board filtering) and protocol-specific (e.g. wearing location) characteristics need to be considered when applying the developed algorithms to new datasets.

Conclusion

We developed three automated sleep-boundary detection methods for the analysis of accelerometer data that will not need self-reported sleep-diary information to label new accelerometer data. The change-point detection, thresholding and random forest methods achieved MAE of 36 min, 34 min and 33 minutes, and kappa statistics of 0.87, 0.89 and 0.89, respectively. A kappa statistic of 0.7 and above is traditionally thought of as a substantial level of agreement [18]. Whilst the random forest method achieved the lowest MAE, the thresholding method performed only marginally worse and requires only a single input variable. We therefore recommend this approach for automated sleep-boundary detection. This work can facilitate the inclusion of polysomnography-validated measures of sleep status in population scale studies.

Supporting information

S1 Fig

Feature ranking as learned by random forest classifier (excluding age and sex information).

S2 Fig

Feature ranking as learned by random forest classifier (including age and sex information).

Acknowledgments

We would like to thank all participants for agreeing to volunteer in this research, and Jill Betts for organising this data collection effort. This analysis was supported by the British Heart Foundation Centre of Research Excellence at Oxford [grant number RE/13/1/30181 to AD], the National Institute for Health Research (NIHR), Oxford Biomedical Research Centre (BRC) based at Oxford University Hospitals NHS Trust and University of Oxford, the NIHR Oxford Health BRC and the Wellcome Trust. We would also like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work (http://dx.doi.org/10.5281/zenodo.22558). JOD acknowledges the support of the RCUK Digital Economy Programme with grant number EP/G036861 (Oxford Center for Doctoral Training in Healthcare Innovation). No funding bodies had any role in the analysis, decision to publish, or preparation of the manuscript.

References

1. Shan Z, Ma H, Xie M, Yan P, Guo Y, Bao W, et al. Sleep duration and risk of type 2 diabetes: a meta-analysis of prospective studies. Diabetes Care. 2015;38(3):529–537. doi:10.2337/dc14-2073.

2. Jackowska M, Steptoe A. Sleep and future cardiovascular risk: prospective analysis from the English Longitudinal Study of Ageing. Sleep Medicine. 2015;16(6):768–774. doi:10.1016/j.sleep.2015.02.530.

3. Cappuccio FP, Cooper D, D'Elia L, Strazzullo P, Miller MA. Sleep duration predicts cardiovascular outcomes: a systematic review and meta-analysis of prospective studies. European Heart Journal. 2011;32(12):1484–1492. doi:10.1093/eurheartj/ehr007.

4. Wulff K, Gatti S, Wettstein JG, Foster RG. Sleep and circadian rhythm disruption in psychiatric and neurodegenerative disease. Nature Reviews Neuroscience. 2010;11(8):589–599. doi:10.1038/nrn2868.

5. Lauderdale DS, Knutson KL, Yan LL, Liu K, Rathouz PJ. Sleep duration: how well do self-reports reflect objective measures? The CARDIA Sleep Study. Epidemiology (Cambridge, Mass). 2008;19(6):838–845. doi:10.1097/EDE.0b013e318187a7b0.

6. Hees VTv, Sabia S, Anderson KN, Denton SJ, Oliver J, Catt M, et al. A Novel, Open Access Method to Assess Sleep Duration Using a Wrist-Worn Accelerometer. PLOS ONE. 2015;10(11):e0142533. doi:10.1371/journal.pone.0142533.

7. Doherty A, Jackson D, Hammerla N, Plötz T, Olivier P, Granat MH, et al. Large Scale Population Assessment of Physical Activity Using Wrist Worn Accelerometers: The UK Biobank Study. PLoS ONE. 2017;12(2). doi:10.1371/journal.pone.0169649.

8. Borazio M, Berlin E, Kücükyildiz N, Scholl P, Laerhoven KV. Towards Benchmarked Sleep Detection with Wrist-Worn Sensing Units. In: 2014 IEEE International Conference on Healthcare Informatics (ICHI); 2014. p. 125–134.

9. Thomas AG, Dennis A, Rawlings NB, Stagg CJ, Matthews L, Morris M, et al. Multi-modal characterization of rapid anterior hippocampal volume increase associated with aerobic exercise. NeuroImage. 2016;131:162–170. doi:10.1016/j.neuroimage.2015.10.090.

10. te Lindert BHW, Van Someren EJW. Sleep Estimates Using Microelectromechanical Systems (MEMS). Sleep. 2013;36(5):781–789. doi:10.5665/sleep.2648.

11. van Hees VT, Fang Z, Langford J, Assah F, Mohammad A, da Silva ICM, et al. Autocalibration of accelerometer data for free-living physical activity assessment using local gravity and temperature: an evaluation on four continents. Journal of Applied Physiology. 2014;117(7):738–744. doi:10.1152/japplphysiol.00421.2014.

12. Sabia S, van Hees VT, Shipley MJ, Trenell MI, Hagger-Johnson G, Elbaz A, et al. Association Between Questionnaire- and Accelerometer-Assessed Physical Activity: The Role of Sociodemographic Factors. American Journal of Epidemiology. 2014;179(6):781–790. doi:10.1093/aje/kwt330.

13. Killick R, Haynes K, Eckley I, Fearnhead P, Lee J. changepoint: Methods for Changepoint Detection; 2016. Available from: `https://cran.r-project.org/web/packages/changepoint/index.html`.

14. Rebecca Killick, Kaylea Haynes, Idris Eckley, Paul Fearnhead, Jamie Lee. Package 'changepoint'; 2016. Available from: `https://github.com/rkillick/changepoint/`.

15. Breiman L. Random Forests. Machine Learning. 2001;45(1):5–32. doi:10.1023/A:1010933404324.

16. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. arXiv:12010490 [cs]. 2012;.

17. Ellis K, Godbole S, Marshall S, Lanckriet G, Staudenmayer J, Kerr J. Identifying Active Travel Behaviors in Challenging Environments Using GPS, Accelerometers, and Machine Learning Algorithms. 2014;2. doi:10.3389/fpubh.2014.00036.

18. Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement. 1960;20(1):37–46. doi:10.1177/001316446002000104.

19. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1. doi:10.1016/S0140-6736(86)90837-8.

20. Cole RJ, Kripke DF, Gruen W, Mullaney DJ, Gillin JC. Automatic sleep/wake identification from wrist activity. 1992;15(5):461–469.

21. Arora T, Omar OM, Taheri S. Assessment for the possibility of a first night effect for wrist actigraphy in adolescents. BMJ Open. 2016;6(10):e012172. doi:10.1136/bmjopen-2016-012172.

22. Ancoli-Israel S, Cole R, Alessi C, Chambers M, Moorcroft W, Pollak CP. The role of actigraphy in the study of sleep and circadian rhythms. Sleep. 2003;26(3):342–392.

Fig 1. Processing steps. Visual representations automated detection of sleep-boundary times as described in this paper. (1) refers to the change-point detection based method, (2) refers to the threshold based method and (3) to the random forest based method.

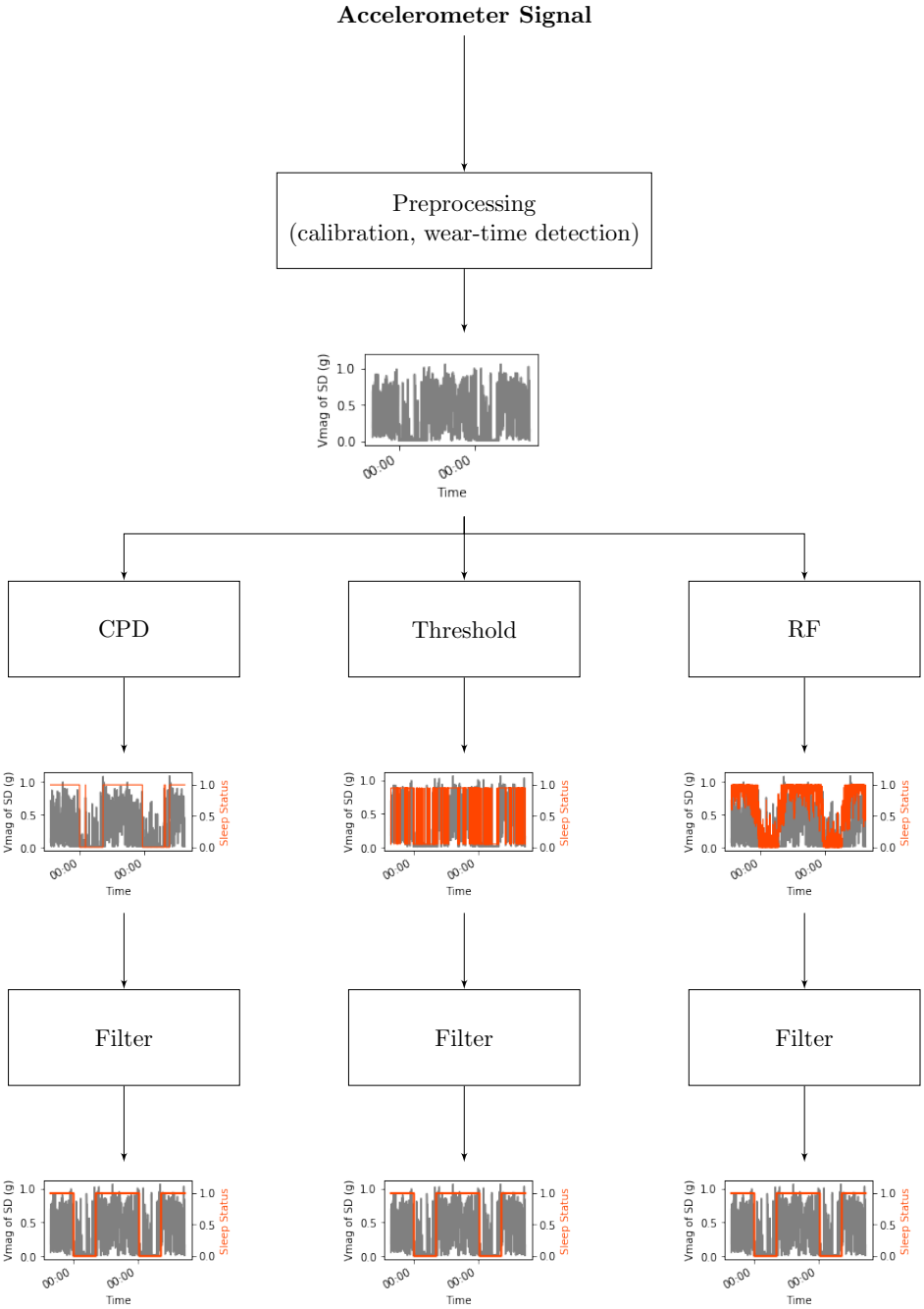


Fig 2. Sleep-labeling using change-point detection. A: Visualisation of false positive change point showing accelerometer signal (grey line) and mean s vmag values for segments identified using change-point detection (orange dashed line). B: Accelerometer signal (grey line) and sleep (0) and wake (1) classes (orange dashed line) after thresholding.

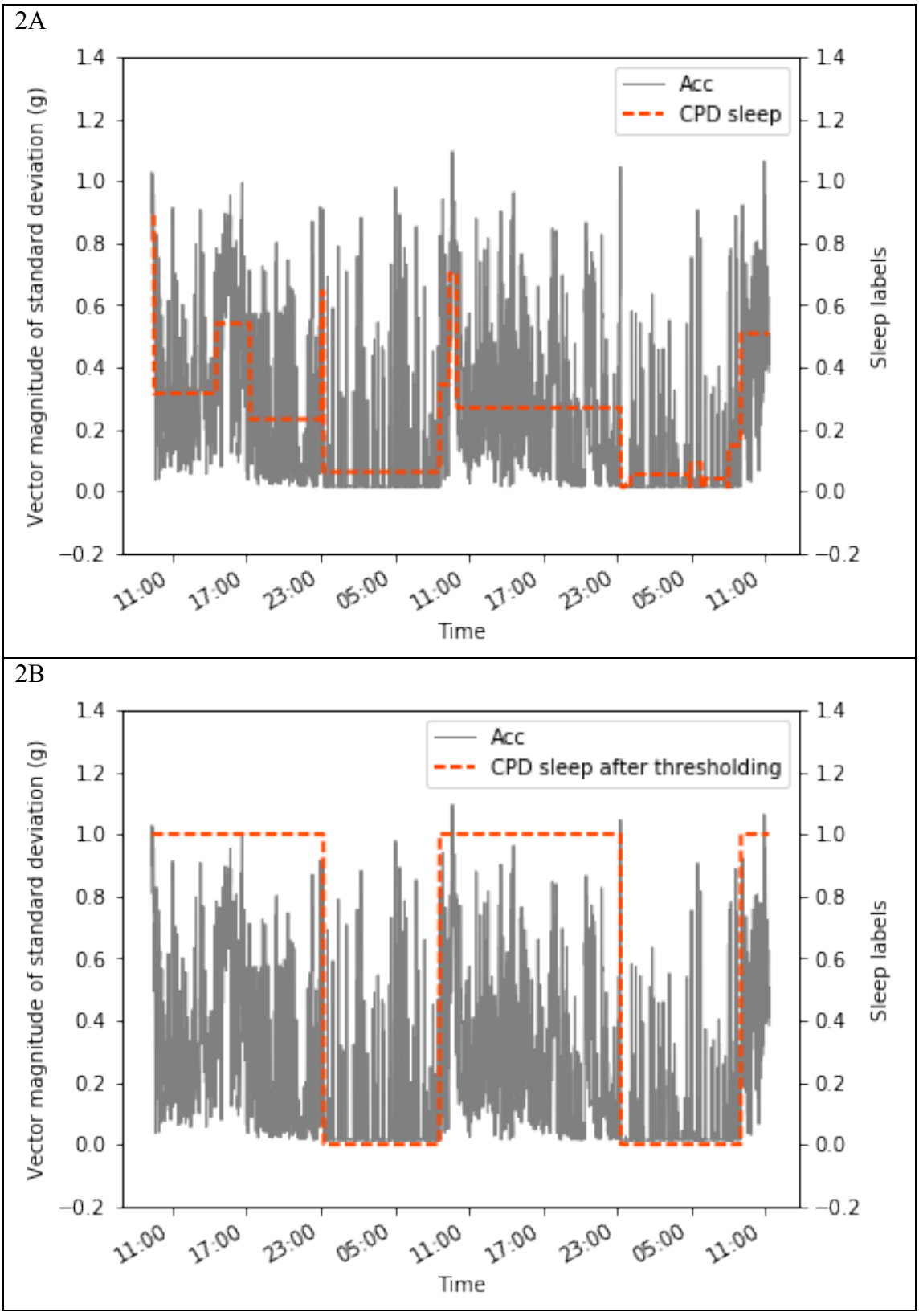


Fig 3. Extraction of PDF-based thresholds for sleep detection.

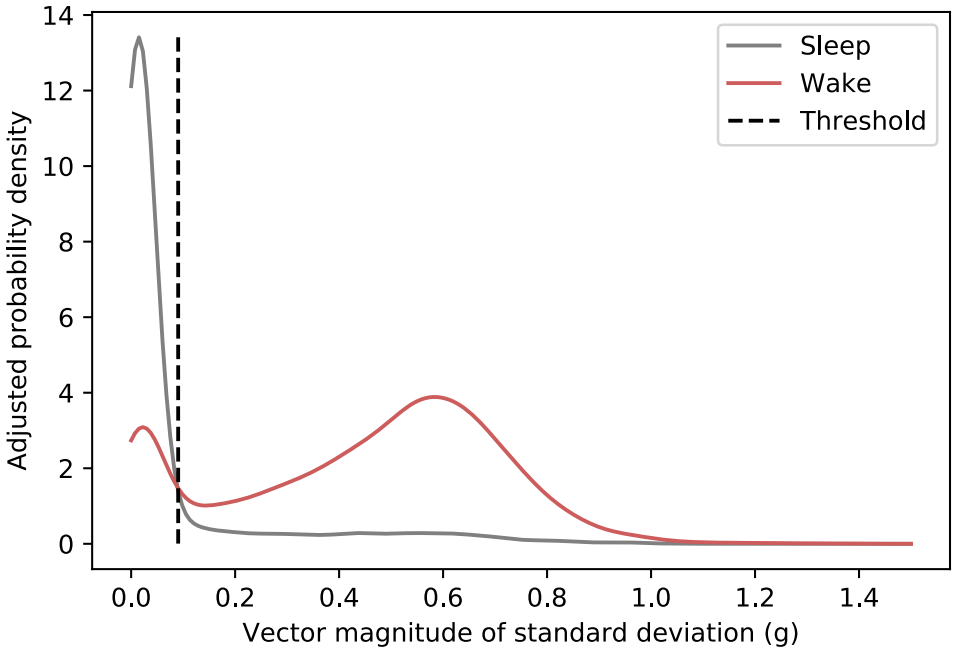


Fig 4. Bland-Altman plots of agreement between automatically-derived (t alg) and self-reported (t an) sleep-onset and wake-up times on an individual level. A: Change-point detection sleep-onset times. B: Change-point detection wake-up times. C: Thresholding sleep-onset times. D: Thresholding wake-up times. E: Random forest sleep-onset times. F: Random forest wake-up times.

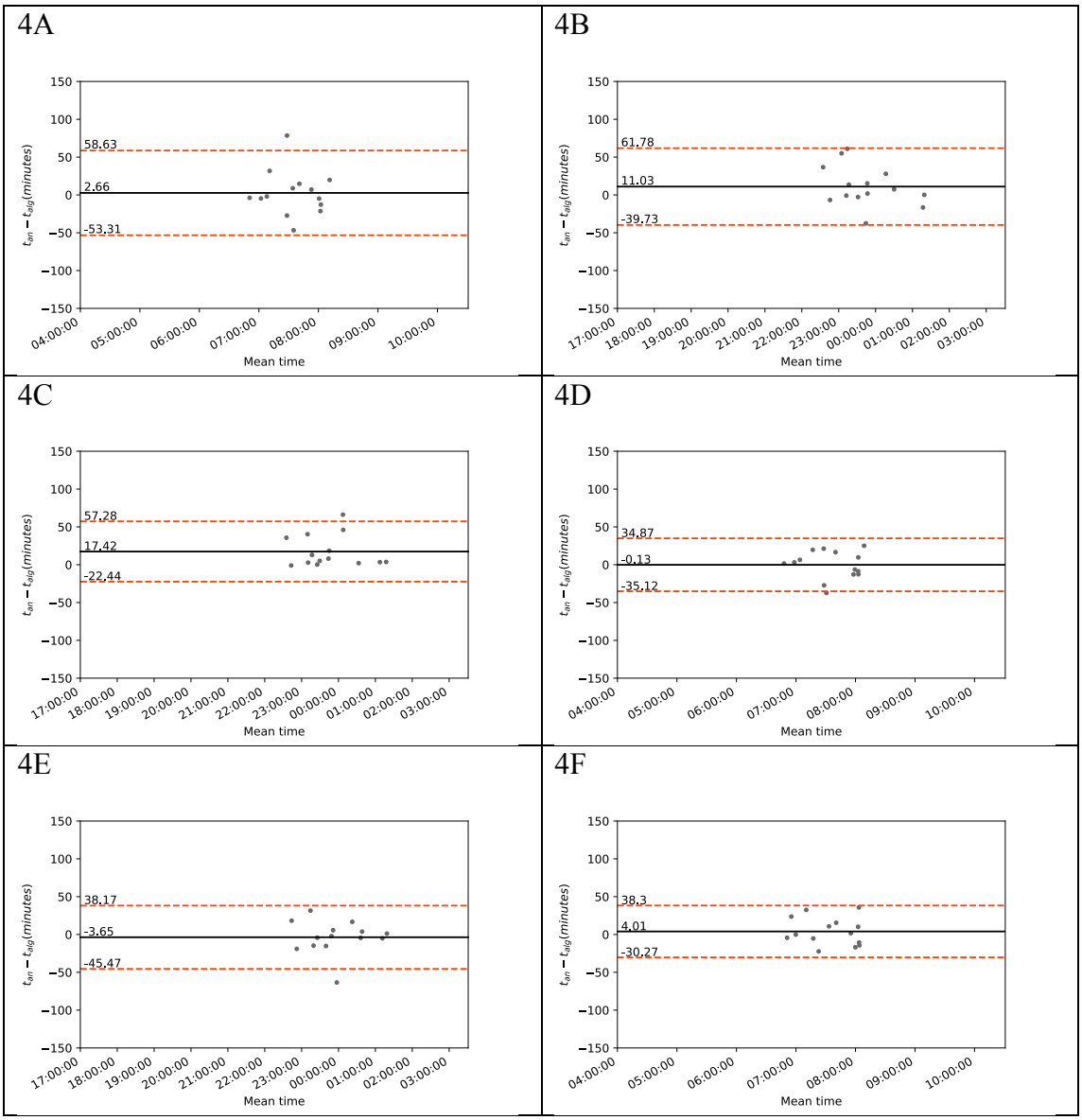
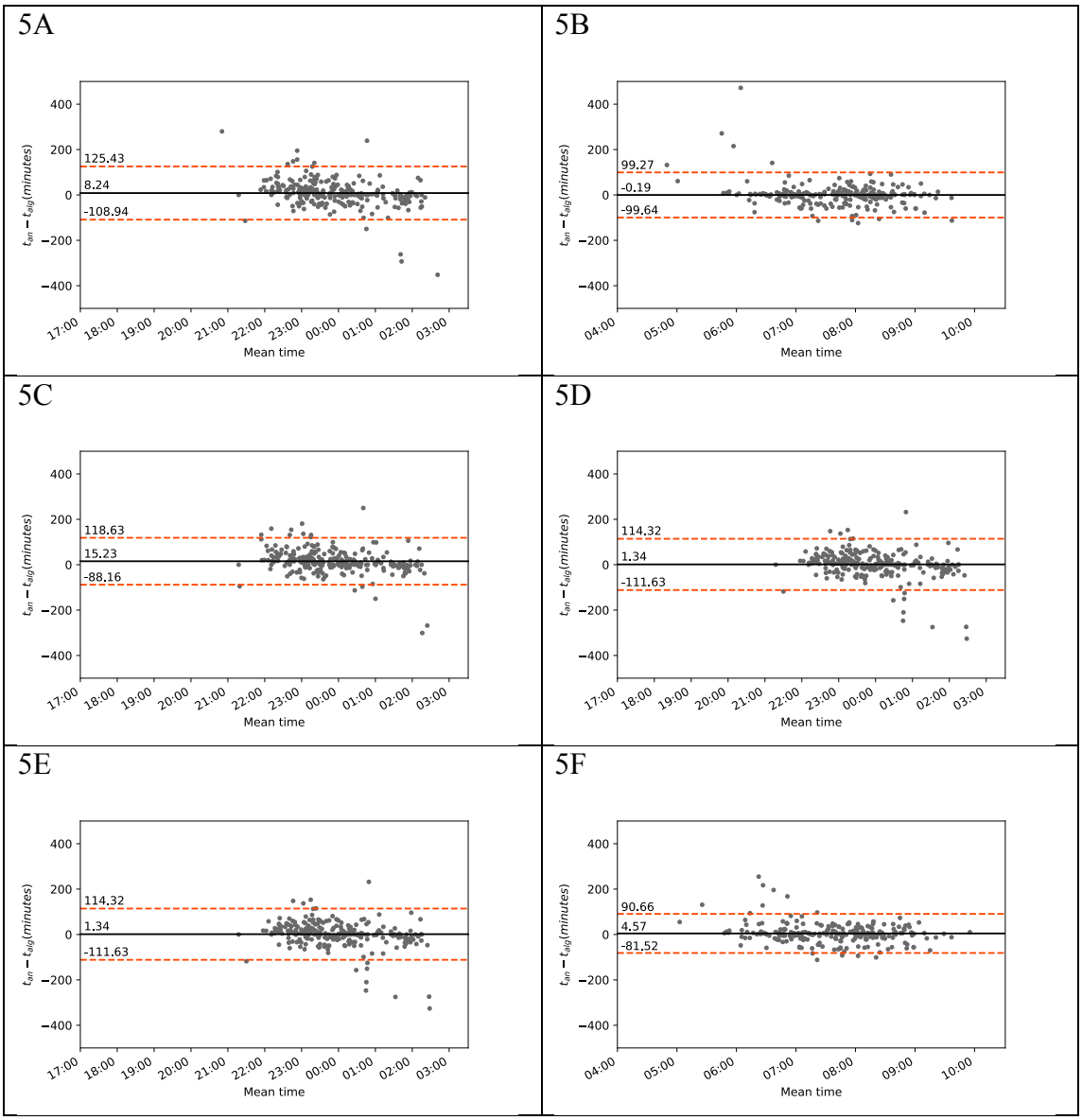


Fig 5. Bland-Altman plots of agreement between automatically-derived (t alg) and self-reported (t an) sleep-onset and wake-up times on an episode level. A: Change-point detection sleep-onset times. B: Change-point detection wake-up times. C: Thresholding sleep-onset times. D: Thresholding wake-up times. E: Random forest sleep-onset times. F: Random forest wake-up times.



S1 + S2 Fig. Feature ranking as learned by random forest classifier.