PREDICTING SEVERAL SLEEP QUALITY METRICS BASED ON SAME DAY PHYSICAL ACTIVITY

Joao Palotti[*†], **Ignacio Perez-Pozuelo**[‡], **Abdulaziz Al Homaid**[†], **Raghvendra Mall**[†], **Marius Posa**[‡]
{jpalotti,abalhomaid,rmall}@hbku.edu.qa; {ip325,imp30}@cam.ac.uk;

Heather Berlin[*], **Di Jin**[*], **Faisal Farooq**[†] **& Peter Szolovits**[*]
{hberlin,jindi15,psz}@mit.edu; fafarooq@hbku.edu.qa;

ABSTRACT

Poor sleep quality has been linked to several chronic, psychiatric, and neurological conditions that can burden the public health system. This paper aims to develop methods to assess sleep quality from physical activity data collected from wearable devices. Our methods can be deployed on a large scale and used by recommender systems to guide the public towards higher quality sleep, hence a better life quality.

1 INTRODUCTION

Despite its ample public health impact, sleep health is rarely considered by policy makers whose policies and structures can adversely affect sleep (Jackson et al., 2015). The regularity, quantity and quality of sleep have been linked to a plethora of health and cognitive outcomes, such as memory consolidation or the removal of free radicals and neurotoxic waste (Stickgold, 2005; Inoué et al., 1995; Xie et al., 2013). Similarly, insufficient or disturbed sleep has been associated with a broad range of chronic, psychiatric and neurological disorders such as depression, anxiety, obesity, diabetes, anorexia, alcoholism, Alzheimer's or Parkinson's disease (Papadimitriou & Linkowski, 2005; Wulff et al., 2010; Simon et al., 2020).

Historically, the gold-standard method to monitor sleep and diagnose a subset of sleep disorders is polysomnography (PSG). While the value of polysomnography is unquestionable, it suffers from scalability issues due to its cost, requirement of trained staff for set-up and analysis and burdensome nature. PSG usually only entails one or two nights of recording in a foreign and often uncomfortable environment, limiting its use for the longitudinal characterization of an individual's sleep. Epidemiologists have traditionally employed sleep diaries and questionnaires for the longitudinal assessment of sleep to understand its role in a variety of diseases. While valuable in the lack of alternatives, these reports tend to suffer from recall bias and often provide partial information (Rosenman et al., 2011). As a result, longitudinal population studies have started to incorporate objective measures of human behaviors through wrist-worn actigraphy or accelerometers (Sudlow et al., 2015; Doherty et al., 2017; Perez-Pozuelo et al., 2019). Consumer-grade wearable devices continue to grow in popularity and their multimodal sensors, often including accelerometry, heart rate measurements through photoplethysmography and gyroscopes, make them attractive candidates to measure physical activity and sleep at scale and unobtrusively (Perez-Pozuelo et al., 2020). Recent work has systematically evaluated the performance of traditional heuristic, machine and deep learning approaches for sleep-wake and sleep-stage classification, showing the strengths and limitations of wrist-worn wearable sensors (Palotti et al., 2019; Yuda et al., 2017; Zhai et al., 2020). These studies show the potential of wearable devices to monitor physical behaviors and sleep at scale, which have important implications not only for the understanding of the role of sleep in disorders but also for designing and implementing interventions based on objectively measured behaviors.

An important area in which such devices shed light is the relationship between physical activity (PA), sedentarism and sleep. Most seminal studies on this area rely on self-reported measures of sleep

[*]Massachusetts Institute of Technology, USA; † Qatar Computing Research Institute, Qatar; ‡: Cambridge University, UK

and PA. For instance, in clinical trials, PA has been associated with better subjective and objective sleep (Kredlow et al., 2015). This finding has been replicated in epidemiological studies which showed that physically active individuals had higher sleep duration, efficiency, and quality (McClain et al., 2014; Lang et al., 2013; Kline et al., 2013; Master et al., 2019). PA has also been linked to lower risks of insomnia (Spörndly-Nees et al., 2017) and daytime sleepiness (Theorell-Haglöw et al., 2015). Recently, physically active individuals were shown to have better sleep efficiency than their sedentary counterparts in a study that objectively measured PA and sleep (Gubelmann et al., 2018).

We set to develop a method that leverages wearable devices' capabilities to objectively monitor PA and make inferences with regards to sleep quality metrics. Identifying and understanding how different behaviors may impact sleep quality provides insight into how behavioral changes could help individuals improve a future night's sleep, which, in turn, would impact their overall health. Our work is guided by the motivation to develop a method that can be deployed in large-scale, longitudinal cohorts and used to assess and improve (through future recommender systems) an individual's sleep quality continuously.

This paper introduces findings for our ongoing research. Here, we: (1) propose the use of an open and large dataset for sleep quality inference based on PA; (2) study metrics used for sleep quality inference and introduce a new multi-dimensional metric; (3) study how different representations of PA influence predicting the same night sleep quality. Our goal is to lay the foundations for understanding and quantifying the bi-directional relationship between sleep and physical activity.

2 METHODS

Dataset The increasing volume of research in mHealth often leads to contradictory findings (Munafò et al., 2017). Many studies suffer from small or constrained sample sizes which are only representative of a small part of the population (Button et al., 2013). Also, most datasets and their transformations tend to be not as open as in other research disciplines, severely impeding reproducible progress. The success of machine and deep learning methods largely depends on the quality and size of the data, making the use of large, diverse datasets particularly important (Banko & Brill, 2001). To ensure transparency and reproducibility, our experiments were carried in one large open dataset: the Hispanic Community Health Study (HCHS). HCHS is a community-based cohort study of 16,415 Hispanic/Latino adults 18–74 years old recruited from randomly selected households at 4 U.S. field centers between 2008 and 2011. This study has been described in detail in (Sorlie et al., 2010; LaVange et al., 2010). As part of the HCHS, the Sueño sleep ancillary study from 2010 to 2013 recruited a total of 2,252 participants. The study had two parts, with an average of 6 years in between each. In the first part, in-home PSG data was collected using an Apnea Risk Evaluation System. The participants were given questionnaires on their sleep habits, demographics and medical history. In the second part, wrist-worn actigraphy data for 1,887 participants was collected over a 7-day period using an Actigraph Spectrum device by Philips Respironics.

Data processing Phillips Actiwatch Spectrum actigraphs were used to collect 30-second resolution activity and light data in the HCHS dataset. This data was evaluated by expert scorers using a standardized approach to set rest intervals that relied on activity data, event markers, sleep diaries, and light intensity (Patel et al., 2015). For each day, rest intervals were identified by the scorer based on those diaries, lights and habitual patterns in the data. For our work, we need to identify one "day", defined as an *active* period followed by a *sleep* period. The sleep period is determined by taking the longest rest segment from 3 PM to 3 PM of the next day, and recursively appending preceding and succeeding rest segments as long as they occur less than 120 minutes apart (including segments in between). The active period is in between sleep periods. Non-wear periods are defined as zero count intervals of at least 90 minutes, allowing for 2-minute intervals of nonzero counts (accounting for artifact movement) surrounded by 30-minute intervals of zero counts, as defined by the Choi algorithm (Choi et al., 2011). To retain only high quality and reliable data, we opt for a strict data preprocessing and exclude days which: have more than 180 minutes of non-wear during the *active* periods; *sleep* periods are extremely short ($\leq$ 3H) or long ($\geq$ 12H); or account for less than 20 hours. Also, participants with less than 5 consecutive days of recording were excluded. A

Table 1: Sleep metrics, their description, age-divided clinical recommendations (■: recommended, ■: may be appropriate, ■: not recommended) and an example of their importance for health and well-being. These recommendations are used as classes in the classification task proposed here.

Metric	Definition and Medical Implications	Age	Implementation						
TST	Total sleeping time during the nighttime (hours). Clinical implications in hypertension (Meng et al., 2013)	18-25 26-64 ≥ 65	≤5:59	6-6:59	7-7:59	8-8:59	9-9:59	10-10:59	≥11
SEF	The percentage of time slept by an individual $SEF = 100. \times \frac{TST}{\text{TotalTimeInBed}}$ ($SEF \in \{0..100\}$) Disease activity in spine inflammation (İn et al., 2016)	18-25 26-64 ≥ 65	<65	65 - 69	70 - 74	75 - 79	80 - 84	85 - 89	≥90
AWA	Total number of awakenings in the nighttime (i.e. arousal lasting ≥5 minutes after sleep offset (Ohayon et al., 2017). Effects on glucose metabolism (Stamatakis & Punjabi, 2010)	18-25 26-64 ≥ 65	0	1	2	3	≥ 4		
CMB	We combine the scores of the metrics above by summing them up according to the following scores: 0 (■), 1 (■), or 2 (■).	All	0	1	2	3	4	5	6

total of 1,529 participants with 10,091 unique days from the HCHS study were included in our final analysis. The data processing was conducted with the Python package HypnosPy[1].

Sleep Quality Metrics The term "sleep-quality" often lacks a formal definition. Frequently, sleep quality is assessed through self-reporting, despite being highly subjective. One limitation of self-reports when assessing sleep quality is the loss of consciousness during sleep, which makes individuals poor self-observers of this particular behavior (Ohayon et al., 2017). Seeking more objective measures, the National Sleep Foundation (NSF) assembled an expert panel to survey the literature and recommend indicators of sleep quality (Hirshkowitz et al., 2015; Ohayon et al., 2017). They compiled sleep metrics divided into three categories ("recommended", "may be appropriate" or "not recommended") depending on the subject's age. In our work, we use three of NSF's metrics that can be reliably extracted from actigraphy devices: total sleep time, sleep efficiency, and the number of awakenings longer than 5 minutes after sleep onset. We also explore the multidimensional nature of sleep by proposing a composite measure for sleep quality that combines the three. These metrics and examples of how they are used for evaluating sleep and related disorders are provided in Table 1.

Machine Learning Methods Using the metrics in Table 1, we model the problem of sleep quality inference as a multi-class classification task with "recommended", "may be appropriate" and "not recommended" as labels. Our goal is to explore possible PA patterns, using them as features to predict each of the four sleep quality metrics. We evaluated several classifiers, from linear models (Linear Discriminant Analysis and Logistic Regression) to tree based and boosting techniques (Random Forests, Extremely Randomized Trees, XGBoost, CatBoost and LightGBM).

Our models were evaluated using common metrics in machine learning and medical sciences, namely: accuracy, Macro F_1 score, and the Matthews Correlation Coefficient (MCC). We chose MCC as our main metric because it is a more reliable, informative and truthful score than traditional accuracy and F_1 score (Chicco & Jurman, 2020). As a correlation metric, MCC varies from -1 (worst) to +1 (best) with 0 meaning no correlation.

We divided the HCHS dataset into 11 folds with roughly the same number of participants each, such that a participant's data is fully contained in a single fold. One of the folds was chosen as a held-out test set. The other ten folds were used for cross-validation, employing a hyperparameter search using the random search technique with 50 steps optimized for MCC (Bergstra & Bengio, 2012).

Feature sets and data transformations We consider a total of 723 PA features divided into four sets: statistical, autoencoder generated, intensity level, and circadian. The raw data obtained from actigraphy devices is given in activity counts per minute (CPM). The statistical features for the *active* period in a day are the overall and hourly maximum, minimum, mean, standard deviation, number of unique values, skewness, and kurtosis, of the raw CPM data. Autoencoder generated features are obtained from the bottleneck layer of linear autoencoders (AE) and variational autoencoders (VAE) trained on the daily mean CPM per hour. The models were trained with multiple linear layers with ReLU activation to minimize an L1 reconstruction loss. This was done to decrease input dimensionality, while retaining the meaningful information about the PA performed in a day (Esteban et al., 2017). Intensity levels are computed following a device-specific, validated conversion from CPM to four activity levels: sedentary, light, moderate and vigorous (Lee & Tse, 2019). The features

[1] https://github.com/HypnosPy/HypnosPy/

Table 2: Training (n=9,164) and test (n=927) results for multi-class classifiers trained to predict sleep quality given the same day physical activity.

HCHS Experiments			Training 10-CV (Mean Std)			Test (% Better/Worse than Baseline)		
Experiment	**Model**	**Target**	**Accuracy**	**Macro F1**	**MCC**	**Accuracy**	**Macro F1**	**MCC**
Baseline	Most Freq	SEF	.69 ± .02	.27 ± .00	.00 ± .00	.69	.27	.00
		AWA	.56 ± .02	.24 ± .01	.00 ± .00	.56	.23	.00
		TST	.40 ± .01	.19 ± .00	.00 ± .00	.39	.19	.00
		COM	.47 ± .02	.21 ± .01	.00 ± .00	.45	.21	.00
PA Features	CATB	SEF	.82±.01	.66±.03	.58±.03	.83 (20%)	.70 (158%)	.60 (+.60)
	LGBM	AWA	.67±.01	.58±.03	.40±.02	.66 (20%)	.60 (151%)	.40 (+.40)
	XGB	TST	.76±.02	.75±.01	.63±.03	.78 (97%)	.78 (314%)	.66 (+.66)
	XGB	COM	.72±.01	.67±.02	.53±.02	.70 (55%)	.64 (209%)	.49 (+.49)

used are the aggregated version of intensity levels (bouts of 1, 5, 10 and $\geq$ 20 minutes) following well-established PA guidelines (Pate et al., 1995). Circadian features are the following parameters extracted from the cosine function fitted to the CPM data: mesor (circadian rhythm-adjusted mean), cycle amplitude (half the distance between minimum and maximum values), cycle acrophase (the time of the day where the circadian cycle reaches its maximum) (Cornelissen, 2014; Moškon, 2020). Disruption of circadian rhythm has been associated with adverse health outcomes such as insomnia or circadian rhythm sleep disorders (Davis & Mirick, 2006; Sack et al., 2007a;b).

We ran experiments using each feature set individually, and their combination. However, we only report the results using the combination of all four feature sets since they performed the best.

3 Results and Discussion

The classification results in Table 2 show the average MCC, accuracy, and Macro F1 on the cross-validated training set and its comparison to a baseline that always selects the most frequent class. Observe that correctly estimating the number of awakenings (AWA) is the most challenging task, with the lowest MCC (0.40). Compared to the baseline, this was also the sleep quality metric with the least gains for all other evaluation metrics. The MCC results for sleep efficiency (SE), total sleep time (TST), and the combined score (COM) are in the same range.

Other works focused on predicting exclusively sleep efficiency (Fellger et al., 2020), which, in terms of accuracy, is the most straightforward metric to predict. We argue, however, that other sleep metrics are equally important to evaluate sleep quality. Because of that, the combined score is a better representative of the multidimensionality of sleep. As its prediction performance is competitive in relation to sleep efficiency, this is a promising measure to predict the overall sleep quality.

Despite experimenting with several complex machine learning algorithms, the Logistic Regression (LR) classifier outperformed all others in the four sleep quality metrics. We performed an ablation study (not shown) to understand each feature set's importance and found that the statistical features were the most influential. The mean amount of PA and the PA distribution skewness were among the top-ranked features for all sleep metrics.

There are many avenues to continue and expand this work. We plan to incorporate features extracted from demographics and questionnaires conducted during the study. We also plan to explore other machine learning models, such as stacking denoising autoencoder for automatic feature extraction from PA; and modeling a pure neural network end-to-end, potentially with multiple prediction heads for the four sleep quality metrics. Last, we would like to explore longitudinal predictions. Unfortunately, studies with actigraphy devices, such as HCHS, MESA (Dean et al., 2016) or the UK Biobank (Fatima et al., 2020), are limited to a single week. To circumvent this limitation, we should explore commercial wearable devices in our future work.

In conclusion, this work is based on clinically validated sleep quality metrics on highly trustworthy open data that provides a reproducible benchmark for future evaluation and exploration of the bi-directional relationship of sleep and physical activity. To ensure transparency and facilitate reproducibility of our results, our code is available on `https://github.com/HypnosPy/sleep_pa_interaction`.

Acknowledgements: We would like to thank Luis Fernandez-Luque, Chief Scientific Officer at Adhera Health Inc (Palo Alto, CA), for his valuable insights around this study and Giselle Reis for reviewing the text and type-setting the manuscript.

References

Michele Banko and Eric Brill. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th annual meeting on association for computational linguistics*, pp. 26–33, Toulouse, France, 2001. Association for Computational Linguistics.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(Feb):281–305, 2012.

Katherine S Button, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376, 2013.

Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13, 2020.

Leena Choi, Zhouwen Liu, Charles E Matthews, and Maciej S Buchowski. Validation of accelerometer wear and nonwear time classification algorithm. *Medicine and science in sports and exercise*, 43(2):357, 2011.

Germaine Cornelissen. Cosinor-based rhythmometry. *Theoretical Biology and Medical Modelling*, 11(1):16, 2014.

Scott Davis and Dana K Mirick. Circadian disruption, shift work and the risk of cancer: a summary of the evidence and studies in seattle. *Cancer causes & control*, 17(4):539–545, 2006.

Dennis A. Dean, Ary L. Goldberger, Remo Mueller, Matthew Kim, Michael Rueschman, Daniel Mobley, Satya S. Sahoo, Catherine P. Jayapandian, Licong Cui, Michael G. Morrical, Susan Surovec, Guo-Qiang Zhang, and Susan Redline. Scaling Up Scientific Discovery in Sleep Medicine: The National Sleep Research Resource. *Sleep*, 39 (5):1151–1164, 5 2016. ISSN 1550-9109. doi: 10.5665/sleep.5774. URL `http://www.ncbi.nlm.nih.gov/pubmed/27070134http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4835314https://academic.oup.com/sleep/article-lookup/doi/10.5665/sleep.5774`.

Aiden Doherty, Dan Jackson, Nils Hammerla, Thomas Plötz, Patrick Olivier, Malcolm H Granat, Tom White, Vincent T Van Hees, Michael I Trenell, Christoper G Owen, et al. Large scale population assessment of physical activity using wrist worn accelerometers: the uk biobank study. *PloS one*, 12(2), 2017.

Cristóbal Esteban, Stephanie L Hyland, and Gunnar Rätsch. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633*, 2017.

Yaqoot Fatima, Romola S Bucks, Abdullah A Mamun, Isabelle Skinner, Ivana Rosenzweig, Guy Leschziner, and Timothy C Skinner. Sleep trajectories and mediators of poor sleep: findings from the longitudinal analysis of 41,094 participants of the uk biobank cohort. *Sleep Medicine*, 76: 120–127, 2020.

Allison Fellger, Gina Sprint, Douglas Weeks, Elena Crooks, and Diane J Cook. Wearable device-independent next day activity and next night sleep prediction for rehabilitation populations. *IEEE Journal of Translational Engineering in Health and Medicine*, 8:1–9, 2020.

Cédric Gubelmann, Raphael Heinzer, José Haba-Rubio, Peter Vollenweider, and Pedro Marques-Vidal. Physical activity is associated with higher sleep efficiency in the general population: the colaus study. *Sleep*, 41(7):zsy070, 2018.

Max Hirshkowitz, Kaitlyn Whiton, Steven M Albert, Cathy Alessi, Oliviero Bruni, Lydia DonCarlos, Nancy Hazen, John Herman, Eliot S Katz, Leila Kheirandish-Gozal, et al. National sleep foundation's sleep time duration recommendations: methodology and results summary. *Sleep health*, 1(1):40–43, 2015.

Shojiro Inoué, Kazuki Honda, and Yasuo Komoda. Sleep as neuronal detoxification and restitution. *Behavioural brain research*, 69(1-2):91–96, 1995.

Chandra L Jackson, Susan Redline, and Karen M Emmons. Sleep as a potential fundamental contributor to disparities in cardiovascular health. *Annual review of public health*, 36:417–440, 2015.

Christopher E Kline, Leah A Irish, Robert T Krafty, Barbara Sternfeld, Howard M Kravitz, Daniel J Buysse, Joyce T Bromberger, Sheila A Dugan, and Martica H Hall. Consistently high sports/exercise activity is associated with better sleep quality, continuity and depth in midlife women: the swan sleep study. *Sleep*, 36(9):1279–1288, 2013.

M Alexandra Kredlow, Michelle C Capozzoli, Bridget A Hearon, Amanda W Calkins, and Michael W Otto. The effects of physical activity on sleep: a meta-analytic review. *Journal of behavioral medicine*, 38(3):427–449, 2015.

Christin Lang, Serge Brand, Anne Karina Feldmeth, Edith Holsboer-Trachsler, Uwe Pühse, and Markus Gerber. Increased self-reported and objectively assessed physical activity predict sleep quality among adolescents. *Physiology & behavior*, 120:46–53, 2013.

Lisa M LaVange, William D Kalsbeek, Paul D Sorlie, Larissa M Avilés-Santa, Robert C Kaplan, Janice Barnhart, Kiang Liu, Aida Giachello, David J Lee, John Ryan, et al. Sample design and cohort selection in the hispanic community health study/study of latinos. *Annals of epidemiology*, 20(8):642–649, 2010.

Paul Lee and CY Tse. Calibration of wrist-worn actiwatch 2 and actigraph wgt3x for assessment of physical activity in young adults. *Gait & posture*, 68:141–149, 2019.

Lindsay Master, Russell T Nye, Soomi Lee, Nicole G Nahmod, Sara Mariani, Lauren Hale, and Orfeu M Buxton. Bidirectional, daily temporal associations between sleep and physical activity in adolescents. *Scientific reports*, 9(1):1–14, 2019.

James J McClain, Daniel S Lewin, Aaron D Laposky, Lisa Kahle, and David Berrigan. Associations between physical activity, sedentary time, sleep duration and daytime sleepiness in us adults. *Preventive medicine*, 66:68–73, 2014.

Lin Meng, Yang Zheng, and Rutai Hui. The relationship of sleep duration and insomnia to risk of hypertension incidence: a meta-analysis of prospective cohort studies. *Hypertension Research*, 36(11):985–995, November 2013. ISSN 1348-4214. doi: 10.1038/hr.2013.70. URL `https://www.nature.com/articles/hr201370`. Number: 11 Publisher: Nature Publishing Group.

Miha Moškon. Cosinorpy: a python package for cosinor-based rhythmometry. *BMC bioinformatics*, 21(1):1–12, 2020.

Marcus R Munafò, Brian A Nosek, Dorothy VM Bishop, Katherine S Button, Christopher D Chambers, Nathalie Percie Du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J Ware, and John PA Ioannidis. A manifesto for reproducible science. *Nature human behaviour*, 1(1):1–9, 2017.

Maurice Ohayon, Emerson M Wickwire, Max Hirshkowitz, Steven M Albert, Alon Avidan, Frank J Daly, Yves Dauvilliers, Raffaele Ferri, Constance Fung, David Gozal, et al. National sleep foundation's sleep quality recommendations: first report. *Sleep health*, 3(1):6–19, 2017.

Joao Palotti, Raghvendra Mall, Michael Aupetit, Michael Rueschman, Meghna Singh, Aarti Sathyanarayana, Shahrad Taheri, and Luis Fernandez-Luque. Benchmark on a large cohort for sleep-wake classification with machine learning techniques. *NPJ digital medicine*, 2(1):1–9, 2019.

George N Papadimitriou and Paul Linkowski. Sleep disturbance in anxiety disorders. *International review of psychiatry*, 17(4):229–236, 2005.

Russell R Pate, Michael Pratt, Steven N Blair, William L Haskell, Caroline A Macera, Claude Bouchard, David Buchner, Walter Ettinger, Gregory W Heath, Abby C King, et al. Physical activity and public health: a recommendation from the centers for disease control and prevention and the american college of sports medicine. *Jama*, 273(5):402–407, 1995.

Sanjay R Patel, Jia Weng, Michael Rueschman, Katherine A Dudley, Jose S Loredo, Yasmin Mossavar-Rahmani, Maricelle Ramirez, Alberto R Ramos, Kathryn Reid, Ashley N Seiger, et al. Reproducibility of a standardized actigraphy scoring algorithm for sleep in a us hispanic/latino population. *Sleep*, 38(9):1497–1503, 2015.

Ignacio Perez-Pozuelo, Thomas White, Kate Westgate, Katrien Wijndaele, Nicholas J Wareham, and Soren Brage. Diurnal profiles of physical activity and postures derived from wrist-worn accelerometry in uk adults. *Journal for the Measurement of Physical Behaviour*, 1(aop):1–11, 2019.

Ignacio Perez-Pozuelo, Bing Zhai, Joao Palotti, Raghvendra Mall, Michaël Aupetit, Juan M Garcia-Gomez, Shahrad Taheri, Yu Guan, and Luis Fernandez-Luque. The future of sleep health: a data-driven revolution in sleep science and medicine. *npj Digital Medicine*, 3(1):1–15, 2020.

Robert Rosenman, Vidhura Tennekoon, and Laura G Hill. Measuring bias in self-reported data. *International journal of behavioural & healthcare research*, 2(4):320, 2011.

Robert L Sack, Dennis Auckley, R Robert Auger, Mary A Carskadon, Kenneth P Wright Jr, Michael V Vitiello, and Irina V Zhdanova. Circadian rhythm sleep disorders: part i, basic principles, shift work and jet lag disorders. *Sleep*, 30(11):1460–1483, 2007a.

Robert L Sack, Dennis Auckley, R Robert Auger, Mary A Carskadon, Kenneth P Wright Jr, Michael V Vitiello, and Irina V Zhdanova. Circadian rhythm sleep disorders: part ii, advanced sleep phase disorder, delayed sleep phase disorder, free-running disorder, and irregular sleep-wake rhythm. *Sleep*, 30(11):1484–1501, 2007b.

Eti Ben Simon, Aubrey Rossi, Allison G Harvey, and Matthew P Walker. Overanxious and underslept. *Nature Human Behaviour*, 4(1):100–110, 2020.

Paul D Sorlie, Larissa M Avilés-Santa, Sylvia Wassertheil-Smoller, Robert C Kaplan, Martha L Daviglus, Aida L Giachello, Neil Schneiderman, Leopoldo Raij, Gregory Talavera, Matthew Allison, et al. Design and implementation of the hispanic community health study/study of latinos. *Annals of epidemiology*, 20(8):629–641, 2010.

Søren Spörndly-Nees, Pernilla Åsenlöf, and Eva Lindberg. High or increasing levels of physical activity protect women from future insomnia. *Sleep medicine*, 32:22–27, 2017.

Katherine A. Stamatakis and Naresh M. Punjabi. Effects of Sleep Fragmentation on Glucose Metabolism in Normal Subjects. *Chest*, 137(1):95–101, January 2010. ISSN 0012-3692. doi: 10.1378/chest.09-0791. URL `http://www.sciencedirect.com/science/article/pii/S0012369210600207`.

Robert Stickgold. Sleep-dependent memory consolidation. *Nature*, 437(7063):1272–1278, 2005.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3), 2015.

Jenny Theorell-Haglöw, Torbjörn Åkerstedt, Johanna Schwarz, and Eva Lindberg. Predictors for development of excessive daytime sleepiness in women: a population-based 10-year follow-up. *Sleep*, 38(12):1995–2003, 2015.

Katharina Wulff, Silvia Gatti, Joseph G Wettstein, and Russell G Foster. Sleep and circadian rhythm disruption in psychiatric and neurodegenerative disease. *Nature Reviews Neuroscience*, 11(8):589–599, 2010.

Lulu Xie, Hongyi Kang, Qiwu Xu, Michael J Chen, Yonghong Liao, Meenakshisundaram Thiyagarajan, John O'Donnell, Daniel J Christensen, Charles Nicholson, Jeffrey J Iliff, et al. Sleep drives metabolite clearance from the adult brain. *science*, 342(6156):373–377, 2013.

Emi Yuda, Yutaka Yoshida, Ryujiro Sasanabe, Haruhito Tanaka, Toshiaki Shiomi, and Junichiro Hayano. Sleep stage classification by a combination of actigraphic and heart rate signals. *Journal of Low Power Electronics and Applications*, 7(4):28, 2017.

Bing Zhai, Ignacio Perez-Pozuelo, Emma AD Clifton, Joao Palotti, and Yu Guan. Making sense of sleep: Multimodal sleep stage classification in a large, diverse population using movement and cardiac sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(2):1–33, 2020.

Erdal İn, Teyfik Turgut, Arif Gülkesen, Servet Yolbaş, Gürkan Akgöl, and Süleyman Koca. Sleep Quality Is Related to Disease Activity in Patients With Ankylosing Spondylitis: A Polysomnographic Study. *Jcr: Journal of Clinical Rheumatology*, 22(5):248–252, August 2016. ISSN 1076-1608. doi: 10.1097/RHU.0000000000000355. URL `insights.ovid.com`.