Using Machine Learning for Wearables to Understand the Association of Sleep with Future Morbidity



Hang Yuan

Keble College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2024

To my grandma, who nurtured me,
and to my mom, who taught me how to see the light in darkness.

Acknowledgements

First, I would like to thank both of my amazing supervisors, Aiden Doherty and Simon Kyle, who gave me invaluable support on both academic and personal levels. Research sometimes is not just about doing the work but also about taking care of oneself. I am particularly grateful that they supported me in taking a break from work with my family after enduring 1.5 months of isolation for COVID quarantine when I flew back home in 2021.

I would also like to thank all the amazing people that I have met during the course of my DPhil. I will give my gratitude to them in chronological order.

In the first year of the DPhil, I was fortunate to be part of the inaugural cohort of the EPSRC Centre for Doctoral Training (CDT) Program. Because of the exposure to a wide arrange of research topics in health, I could identify the kind of work that would interest me most intellectually in the following years. Obviously, I would like to thank all the program directors, Jim Davis, Tom Nicholas, Nina Hallowell, and Jens Rittscher. Another key person who made this program so special is Leema Chapman, who made sure that we were looked after. Nina and Leema gave us insights into the CDT program and were outspoken advocates for us students. It is unfortunate that Nina passed away recently but I am sure she would be proud of who I have become. I would also like to thank my other fellow CDT students, with whom I learnt and grew together, including Henrique Aguiar, Andres Tamm, Yifan Yu, Kangning Zhang and many more.

During the second year of my DPhil, I joined the OxWearables and Experimental and Clinical Sleep Medicine groups. Through my interactions with these lovely people, I became a more independent thinker and a researcher with many skills. I would like to thank Shing Chan, Rosemary Walmsley, Andrew Creagh, Catherine Tong, Aiden Acquah, Alain Shreves, Scott Small, Laura Portas, Scott Small, Stefan Van Duijvenboden, Abram Schonfeldt, Lizzie Hill, and Rachel Sharman. I would like to highlight the conversations that I have had with Shing Chan and Rosemary Walmsley. Both were my go-to people if I encountered any technical difficulties or was simply feeling hopeless at times. Those light-hearted conversations really made the whole DPhil experience far more enriching.

As my DPhil project developed, I also had much help from the collaborators elsewhere, some of whom contributed data they collected, and some others provided

critical feedback for my work in progress. I would give my warm regards to my collaborators in Australia, Leicester and Pennsylvania: Peter Eastwood, Amy Reynolds, Kathleen Maddison, Tatiana Plekhanova, Philip Gehrman and others whom I might not be able to have direct contact with. Even though Amy and I only had a few encounters during our collaboration, she gave me strong support in how to put writing into perspective and helped me understand how sleep clinicians think.

On a personal level, I met so many friends by playing tennis at the University. It is the sport that kept me sane during COVID. Thank you to my coaches, Matej Lukic and Jarlath Brine, who taught me to be a better player and master the mental game of tennis. Thank you to my fellow tennis mates, Renyuan Xu and Tiger Yuan and friends from OULTC, who I still managed to see on a regular basis during the lockdown in the UK for sports with social distancing rules.

In the summer of 2022, I was fortunate to have taken a short departure from my study program to do an internship at Meta, London. I have to say it was a very intense 12-week program, but I learnt a huge deal about how to build machine learning systems at scale and the best practices in the industry. Thank you to my manager Yanping Xie who made sure that I had impactful projects and listened to some of my concerns. Thank you to my coworkers at Meta AI/Applied AI teams who also mentored me along the way, Antoine Toisoul, Chengyang Fu, and Nikhila Ravi.

I would also like to thank my funder Novo Nordisk who generally provided the resources to make my DPhil project financially feasible. A special shout out to Eneida Parizotto, and Joanna McCammond McGill Howson, who made our collaboration as smooth as possible.

There is a special group of people that I have never met but are an essential part of my DPhil project. Thank you to the study participants in the UK Biobank, the Raine study and all the participants who were involved in the datasets that I used. It is you who really made everything possible.

Last but not least, I would like to give my greatest gratitude to my family. As the first person in my family to ever attempt to study at the doctoral level, my family gave me unquestioned support for a long time. Even before I wanted to go abroad to America for high school, they rarely said no to my desire to go after education opportunities. Especially, I would like to say thanks to a few female characters in my family who gave me the chance and support needed to explore what my heart desired, my mother, Guiyuan Shen, my grandmother, Wuxiu Zeng and my aunty, Min Shen. At the same time, I would also like to thank my partner Zhuocheng Xiao. We fight all the time, and we still do. But it was you who helped me see what I could do and gave me the reassurance I needed. I do hope our time together will be free from conflicts and filled with more harmony in the future.

Abstract

Sleep is essential to life and is structurally complex. Our understanding of how sleep is associated with health and morbidity primarily draws on studies that use self-report sleep diaries, which capture the subjective experience. However, sleep diaries are limited because they are measured at one point in time, single dimensional, often only capturing sleep duration, and have a low correlation with objective device-measured sleep parameters. The accepted standard for sleep measurement is laboratory-based polysomnography, which is not feasible for use at scale due to its high cost and technical complexity. Instead, wrist-worn accelerometers are more viable to deploy in large-scale epidemiological studies because of their portability and low user burden. Therefore, I aimed to develop a machine learning method for sleep stage classification and evaluate its utility to provide insights into the association between sleep and health outcomes.

I conducted a systematic review to assess the agreement between accelerometer-based sleep staging and polysomnography with a secondary aim of understanding the agreement level from different methods used. This review found that existing sleep staging methods were limited by a reliance on hand-crafted features and the use of labelled datasets of small sizes, highlighting the need to improve accelerometer-based sleep staging.

A self-supervised deep neural network was first developed to automatically extract features from 700,000 person-days of unlabelled raw accelerometry. To systematically evaluate the generalisability of the self-supervised features, the pre-trained network was tested in seven human activity recognition datasets, for which more open-access benchmark datasets were available. The self-supervised network showed generalisability across activity classes, devices, device placements and populations with an F1 relative improvement of 2.5%-100% (median: 18.4%) compared to the network without self-supervision.

The self-supervised feature extractor was then used to develop a sleep stage classifier (SleepNet) using a deep recurrent neural network. The SleepNet was able to obtain a state-of-the-art performance for the classification of sleep and the stages of sleep using 1,500 nights of multi-centre polysomnography as the ground truth. Overall, the derived sleep parameters had a fair agreement with polysomnography with a Kappa score of 0.37 (SD: 0.16) for three-class classification between wake,

rapid-eye-movement sleep (REM), and non-rapid-eye-movement sleep (NREM). The difference between polysomnography and the model classifications on the external validation was 34.7 minutes (95% limits of agreement (LoA): -37.8 to 107.2 minutes) for total sleep duration, 2.6 minutes for REM duration (95% LoA: -68.4 to 73.4 minutes) and 32.1 minutes (95% LoA: -54.4 to 118.5 minutes) for NREM duration.

Finally, SleepNet was used to infer overnight sleep duration and sleep efficiency (the proportion of time asleep when in bed) in the UK Biobank accelerometer dataset to understand the associations with mortality outcomes. Short sleepers (<6 hours) had a higher risk of mortality compared to participants with normal sleep duration 6 to 7.9 hours, regardless of whether they had low sleep efficiency (Hazard ratios (HRs): 1.58; 95% confidence intervals (CIs): 1.19 to 2.11) or high sleep efficiency (HRs: 1.45; 95% CIs: 1.16 to 1.81).

In conclusion, accelerometer-based sleep classification had a fair agreement with polysomnography. Using the derived sleep parameters on datasets with longitudinal follow-up could transform our understanding of how sleep contributes to human health and well-being.

Contents

List of Tables **xiii**

List of Figures **xvii**

List of Abbreviations **xxi**

1 Introduction **1**
- 1.1 Sleep . . . 1
- 1.2 Sleep and human health . . . 4
- 1.3 Beyond sleep measurement using questionnaires . . . 6
  - 1.3.1 Self-reported sleep . . . 6
  - 1.3.2 Device-measured sleep . . . 8
- 1.4 Challenges . . . 10
- 1.5 Thesis aims and objectives . . . 13
- 1.6 Thesis outline . . . 13
- 1.7 Summary . . . 14

2 Systematic review of the performance of wrist-worn accelerometers in measuring sleep stages **15**
- 2.1 Review goals . . . 16
- 2.2 Review protocol . . . 17
  - 2.2.1 Eligible studies . . . 17
  - 2.2.2 Search strategies . . . 20
  - 2.2.3 Study selection and data extraction . . . 20
  - 2.2.4 Data evaluation . . . 21
- 2.3 Review results . . . 24
  - 2.3.1 Study populations . . . 24
  - 2.3.2 Accelerometer devices . . . 24
  - 2.3.3 Sleep staging algorithms . . . 27
  - 2.3.4 Performance metrics and assessment . . . 27
  - 2.3.5 Risk of bias of included studies . . . 30
- 2.4 Discussions . . . 31
- 2.5 Summary . . . 33

3 Raine Study and UK Biobank 35
3.1 Raine Study . . . 36
3.1.1 Recruitment . . . 36
3.1.2 Accelerometer and polysomnography follow up . . . 37
3.1.3 Population characteristics . . . 37
3.2 UK Biobank . . . 40
3.2.1 Study protocol . . . 40
3.2.2 Accelerometry in the UK Biobank . . . 42
3.2.3 Population characteristics . . . 43
3.3 Summary . . . 47

4 Self-supervised representation learning for accelerometers 49
4.1 Introduction . . . 50
4.2 Methods . . . 53
4.2.1 Datasets . . . 54
4.2.2 Multi-task self-supervised-learning . . . 55
4.2.3 Network training . . . 57
4.2.4 Evaluation – human activity recognition . . . 59
4.3 Results . . . 61
4.3.1 Weighted single task training . . . 61
4.3.2 Multi-task self-supervised learning . . . 64
4.3.3 Downstream performance - human activity recognition . . . 65
4.3.4 Transfer learning using labelled pre-training . . . 68
4.3.5 Ablation studies . . . 68
4.3.6 Understanding the representation . . . 71
4.4 Discussions . . . 73
4.5 Summary . . . 74

5 Deep learning for sleep stage detection using accelerometers 77
5.1 Introduction . . . 78
5.2 Methods . . . 79
5.2.1 Study design and participants . . . 79
5.2.2 Accelerometer devices and data preprocessing . . . 82
5.2.3 Ascertainment of sleep stages via polysomnography . . . 83
5.2.4 Sleep stage classification using wrist-worn accelerometers . . . 84
5.3 Results . . . 90
5.3.1 Baseline SleepNet performance . . . 90
5.3.2 Subject-wise performance external evaluation . . . 92
5.3.3 Epoch by epoch performance evaluation . . . 96
5.3.4 The performance of sleep parameters estimation . . . 101

5.4 Discussion . . . 103
5.4.1 Comparison to existing sleep staging methods . . . 103
5.4.2 Sleep parameters agreement between SleepNet and polysomnography . . . 104
5.4.3 Strengths and limitations . . . 105
5.5 Summary . . . 107

6 Associating device-measured sleep with all-cause mortality 109
6.1 Introduction . . . 110
6.2 Methods . . . 112
6.2.1 Reliability assessment for device wear time exclusion criterion 112
6.2.2 Measurements of sleep in 100,000 UK Biobank participants . 114
6.2.3 Association analysis for mortality . . . 117
6.3 Results . . . 119
6.3.1 Population characteristics for the UK Biobank . . . 119
6.3.2 Descriptive analysis of device-measured sleep . . . 122
6.3.3 Association of device-measured sleep with all-cause mortality 128
6.4 Discussion . . . 132
6.4.1 Descriptive analysis in the UK Biobank . . . 132
6.4.2 Association analysis between device-measured sleep and mortality . . . 133
6.4.3 Strengths and limitations . . . 134
6.5 Summary . . . 135

7 Discussion 137
7.1 Summary of findings . . . 138
7.2 Thesis limitations . . . 141
7.3 Ethical implications . . . 143
7.4 Future work . . . 145
7.4.1 Multi-modal sleep classification . . . 145
7.4.2 Towards consumer-grade devices . . . 146
7.4.3 Beyond accelerometer-based sleep staging . . . 147
7.5 Conclusion . . . 147

Appendices

A Data for the systematic review of the performance of actigraphy in measuring sleep stages 151
A.1 Data extraction item . . . 151
A.2 Search terms . . . 153

B Data for self-supervised learning for human activity recognition 165
B.1 Datasets 166
B.2 Results 167

C Data for sleep classification using wrist-worn accelerometers 169
C.1 Additional data for Agreement between derived sleep parameters and polysomnography 170

List of Tables

1.1 **Pathophysiology of sleep stages.** R: rapid-eye-movement sleep; N1, N2, N3: non-rapid-eye-movement sleep 1, 2, 3. Summarisation of [15, 19, 20]. 4
1.2 Characteristics of sleep diary, accelerometry and polysomnography . 6

2.1 **Eligibility criteria**: ACC, accelerometer. 19
2.2 **Modified QUADAS-2 form**: Bold text questions are newly added. 23
2.3 **Study characteristics.** 26
2.4 **Sleep staging algorithms and performance.** Wake: W; Non-rapid-eye-movement sleep 1, 2, and 3: NREM, N1, N2, and N3; rapid-eye-movement sleep: R; polysomnography: polysomnography. 29
2.5 **Risk of bias using assessment Quality Assessment of Diagnostic Accuracy Studies tool (QUADAS-2).** 30

4.1 Wrist-worn accelerometer datasets used to evaluate the utility of self-supervised learning for human activity recognition tasks. 54
4.2 Dataset characteristics 54
4.3 Downstream human activity recognition performance (subject-wise F1 (±SD)) when using different self-supervised learning tasks after fine-tuning: Arrow of the Time (AoT), Permutation (P), Time warping (TW). The tasks were trained using the same 1,000 UK-Biobank participants. 63
4.4 Downstream human activity recognition performance (subject-wise F1 (±SD)) for different self-supervised task combinations using 1,000 randomly sampled UK Biobank participants. N is the number of samples. 64
4.5 Subject-wise F1 and Kappa (κ) for downstream human activity recognition tasks (mean ± SD) using 100,000 UK Biobank participants for pre-training. The relative improvement compares the performance between the model that is trained from scratch and fine-tuning using all the layers. 66
4.6 Transfer learning (subject-wise F1) performance comparison between supervised pre-training with self-supervised pre-training. 67

5.1 **Characteristics of the datasets used for internal validation, external validation and health association analyses** "Patient" indicates whether a cohort consists of sleep patients in a clinic. . . 81
5.2 **Hand-crafted features** 86
5.3 **Model performance metric definitions (TP: true positive; TN: true negative; FP: false positive; FN: false negative)** . 88
5.4 **Sleep parameter definitions: total sleep duration (TSD), rapid-eye-movement (REM), non-rapid-eye-movement (NREM), sleep onset latency (SOL), wake after sleep onset (WASO), and sleep efficiency (SE).** 89
5.5 **Subject-wise sleep stage classification for benchmark models using internal validation datasets with the Raine Study and the Newcastle cohort:** The random forest model was trained using hand-crafted features. SleepNet is the deep recurrent network without pre-training. SleepNet-SSL is the network pre-trained using self-supervision. Five-fold subject-wise performance metrics (mean ± SD) are reported using the internal validation data. REM: rapid-eye-movement sleep, NREM: non-rapid-eye-movement sleep, Kappa score: κ. 91
5.7 **Model characteristics on the internal validation datasets (wake versus sleep):** subject-wise performance metrics (mean ± SD) are reported using the internal validation data. Sen: sensitivity, Spe: specificity. Wake is the negative class and the sleep is the positive class when calculating model performance. 94
5.8 **Model characteristics on the internal validation datasets (wake versus REM versus NREM):** subject-wise performance metrics (mean ± SD) are reported using the internal validation data. REM: rapid-eye-movement, NREM: non-rapid-eye-movement, Kappa score: κ. 95
5.9 **Model characteristics on the internal validation datasets (wake versus REM versus NREM I, II, III):** subject-wise performance metrics (mean ± SD) are reported using the internal validation data. REM: rapid-eye-movement, NREM: non-rapid-eye-movement, Kappa score: κ. 95

6.1 **Code table for UK Biobank variables used in the study.** . . 116

A.1 Search terms for Ovid. 153
A.2 Search terms for Web of Science. 154

A.3 **Exclusion reasons for studies that might be eligible.** AASM: American Academy of Sleep Medicine. 156

B.1 The license, consent and device used for each dataset 166
B.2 Test accuracies on different tasks 1000 subjects. 167

List of Figures

2.1 **PRISMA flow diagram of studies screened and included to review the validity of wrist-worn accelerometers for the measurement of sleep stages.** 25

4.1 Overview of the proposed self-supervised learning pipeline: step 1 involves multi-task self-supervised learning on 700,000 person-days of data from the UK Biobank. Then, utility of the pre-trained network is evaluated in seven benchmark human activity recognition baselines via transfer learning. Reproduced and modified with permissions from [133]. Figure produced by Andrew P. Creagh. 52

4.2 Data collected from free-living environment tends to have more stationary periods than the data collected in lab environment. SD: standard deviation; oppo: opportunity dataset. 57

4.3 Accuracy test curves for training four self-supervised tasks individually using 1,000 randomly sampled participants from the UK Biobank with and without weighted sampling. The performance shown was on the training data. The patience for early-stopping was five. AoT: arrow of time; TW: time-warping. 62

4.4 **Left**: More #unlabelled subjects in the self-supervision leads to better downstream performance. **Right**: Unlabelled data per subject does not improve downstream performance when using 10,000 subjects for pre-training. Oppo: opportunity dataset. 69

4.5 How different models perform in the downstream tasks when we change the number of labelled subjects: Capture-24 (**left**) and Rowlands (**right**) 70

4.6 **Cluster analysis on raw inputs, untrained features and self-supervised pretrained features:** The colour gradients denote activity intensities. Results suggest that self-supervised features were better at clustering similar activities (e.g. walking, stair climbing vs. sitting, writing, typing) as well as their intensities (e.g. lying down, sitting, standing vs. jogging, sports). Figure produced by Shing Chan. 72

5.2 **Sleep stage distribution for all the datasets used.** 83

5.3 **Receiver operating characteristics curves for two-class (wake/sleep) and three-class (wake/REM/NREM) settings on the internal validation dataset using the best performing model self-supervised SleepNet.** REM: rapid-eye-movement sleep, NREM: non-rapid-eye-movement sleep. 92

5.4 **Three class classification (wake/REM/NREM) confusion matrix**: epoch-to-epoch Kappa and balanced accuracies are shown. The number of predictions and proportion ratios are shown for each pair of ground-truth and prediction class. REM: rapid-eye-movement sleep; NREM: non-rapid-eye-movement sleep. 96

5.5 **Three-class sleep staging (wake/REM/NREM) for internal validation: epoch-to-epoch Kappa and balanced accuracies are shown.** The number of predictions and proportion ratios are shown for each pair of ground-truth and prediction class. REM: rapid-eye-movement sleep; NREM: non-rapid-eye-movement sleep. . 97

5.6 **Five-class sleep staging (wake/REM/N1/N2/N3) for internal validation: epoch-to-epoch kappa and balanced accuracies are shown.** The number of predictions and proportion ratios are shown for each pair of ground-truth and prediction class. REM: rapid-eye-movement sleep, N1, N2, N3: non-rapid-eye-movement sleep 1, 2, 3. 98

5.7 **Three-class sleep staging (wake/REM/NREM) for external validation: epoch-to-epoch kappa and balanced accuracies are shown.** The number of predictions and proportion ratios are shown for each pair of ground-truth and prediction class. REM: rapid-eye-movement sleep; NREM: non-rapid-eye-movement sleep. 99

5.8 **Five-class sleep staging (wake/REM/N1/N2/N3) for external validation: epoch-to-epoch kappa and balanced accuracies are shown.** The number of predictions and proportion ratios are shown for each pair of ground-truth and prediction class. REM: rapid-eye-movement sleep, N1, N2, N3: non-rapid-eye-movement sleep 1, 2, 3. 99

5.9 **A sample actigram, hypnogram ground truth and prediction for a participant whose sleep stages are well captured:** the **top** hypnogram is the ground-truth and the **bottom** hypnogram is the prediction generated by SleepNet based on the actigram. REM: rapid-eye-movement sleep, N1, N2, N3: non-rapid-eye-movement sleep 1, 2, 3. 100

6.2 **How the intraclass correlation coefficient (ICC) changes with respect to the non-wear hours (h) (left) and the number of wear days (right) in a reliability simulation using data from 27,870 participants that had zero non-wear time across a seven-day period.** Mean and 95% confidence intervals are plotted. 112
6.3 **The distribution of non-wear time for all the participants from the UK Biobank.** 113
6.4 **Participant flow diagram for the analysis of sleep and all-cause mortality in the UK Biobank.** TDI: Townsend deprivation index, BMI: body mass index, SR_health: self-reported overall health, SR_insomnia: self-reported insomnia symptoms, CVD: Cardiovascular disease. 120
6.5 **Correlation matrix for device-measured and self-reported sleep parameters on the UK Biobank.** The self-reported total sleep duration was obtained via questionnaire at baseline assessment in the UK Biobank. REM: rapid-eye-movement sleep, NREM: non-rapid-eye-movement sleep. 123
6.6 **Box plots showing the distributions of device-measured overnight sleep duration against self-reported total sleep duration.** The box whiskers reflect the lowest and highest data points that are 1.5 times of the inter-quartile-range from the median. 124
6.9 **Adjusted marginal mean (95% confidence interval) device-measured mean overnight sleep duration and mean sleep efficiency by self-reported overall health status and insomnia history in the UK Biobank.** Mean overnight sleep duration and sleep efficiency were adjusted for age and sex. 126
6.10 **Device-measured sleep probability trajectories throughout the day for the UK Biobank participants (weekday vs weekend).** Top: variations of the average overnight sleep probability for the participants with self-reported "morning" and "evening" chronotype (a) and the overnight sleep distributions across thirds of device-measured physical activity level (b). Bottom: variations of the average REM (c) and NREM (d) probability in participants with a history of self-reported insomnia symptoms versus those without. Rapid-eye-movement sleep (REM), and non-rapid-eye-movement sleep (NREM). 127

C.1 **Agreement assessment via Bland-Altman plots for internal validation: total sleep duration (TSD), non-rapid-eye-movement sleep (NREM), and rapid-eye-movement sleep (REM).** 171

C.2 **Agreement assessment via Bland-Altman plots for external validation: total sleep duration, wake after sleep onset (WASO), non-rapid-eye-movement sleep (NREM), and rapid-eye-movement sleep (REM).** 172
C.3 **Agreement assessment via Bland-Altman plots for internal validation: non-rapid-eye-movement sleep (NREM) ratio, and rapid-eye-movement sleep (REM) ratio.** 173
C.4 **Agreement assessment via Bland-Altman plots for external validation: non-rapid-eye-movement sleep (NREM) ratio, and rapid-eye-movement sleep (REM) ratio.** 174
C.5 **Agreement assessment via Bland-Altman plots for internal validation: wake after sleep onset (WASO), and sleep efficiency (SE).** 175
C.6 **Agreement assessment via Bland-Altman plots for internal validation: wake after sleep onset (WASO), and sleep efficiency (SE).** 176

List of Abbreviations

CVD . . . . . Cardiovascular diseases.

PSG . . . . . Polysomnography.

ECG . . . . . Electrocardiogram.

Raine . . . . . Australian Pregnancy Cohort.

SSL . . . . . Self-supervised learning.

AoT . . . . . Arrow of time.

CAM . . . . . Class Activation Mapping.

W . . . . . Wake.

REM . . . . . Rapid-eye-movement.

NREM . . . . Non-rapid-eye-movement.

CNN . . . . . Convolutional neural network.

HMM . . . . . Hidden Markov Model.

SOL . . . . . Sleep onset latency.

SE . . . . . Sleep efficiency.

TST . . . . . Total sleep time.

WASO . . . . . Wake after sleep onset.

Contributions

The work presented in the thesis is my own work under the supervision of Professor Aiden Doherty and Professor Simon K. Kyle. I conceptualised the project, conducted the data cleaning and analysis, and the final writing.

Several other people also contributed to part of the work presented:

- Shing Chan helped to conceptualise the self-supervised learning project with me and conducted part of the experiments and the writing of the manuscript. Shing Chan and Andrew P. Creagh also contributed to the explainable AI experiments for the self-supervised learning project.
- Rosemary Walmsley laid the base for the epidemiological analysis of sleep in the UK Biobank and provided critical feedback in many aspects of this work.
- My co-authors in [1–3] also guided me in the data analysis, interpretation and writing of the work presented in this thesis.

Figures and tables are my own work unless explicitly stated otherwise.

Outputs

At the time of submission, this work expands on the following first-author manuscripts:

- **Hang Yuan**, Tatiana Plekhanova, Rosemary Walmsley, Amy C. Reynolds, Kathleen J. Maddison, Maja Bucan, Philip Gehrman, Alex Rowlands, David W. Ray, Derrick Bennett, Joanne McVeigh, Leon Straker, Peter Eastwood, Simon D. Kyle, and Aiden Doherty. "Self-supervised learning of accelerometer data provides new insights for sleep and its association with mortality". In: *npj Digital Medicine* (2024)

- **Hang Yuan**, Elizabeth A Hill, Simon D Kyle, and Aiden Doherty. "A systematic review of the performance of actigraphy in measuring sleep stages". In: *Journal of Sleep Research* (2023)

- **Hang Yuan***, Shing Chan*, Andrew P Creagh, Catherine Tong, Aidan Acquah, David A Clifton, and Aiden Doherty. "Self-supervised Learning for Human Activity Recognition Using 700,000 Person-days of Wearable Data". In: *npj Digital Medicine* (2024)

  - **Best poster award** @ CDT for Health Data Science Symposium, 2022

I have also contributed to the following manuscripts:

- Organizers of Queer in AI, Anaelia Ovalle, Arjun Subramonian, Ashwin Singh, Claas Voelcker, Danica J. Sutherland, Davide Locatelli, Eva Breznik, Filip Klubicka, **Hang Yuan**, J Hetvi, Huan Zhang, Jaidev Shriram, Kruno Lehman, Luca Soldaini, Maarten Sap, Marc Peter Deisenroth, Maria Leonor Pacheco, Maria Ryskina, Martin Mundt, Milind Agarwal, Nyx McLean, Pan Xu, A Pranav, Raj Korpan, Ruchira Ray, Sarah Mathew, Sarthak Arora, St John, Tanvi Anand, Vishakha Agrawal, William Agnew, Yanan Long, Zijie J. Wang, Zeerak Talat, Avijit Ghosh, Nathaniel Dennler, Michael Noseworthy, Sharvani Jha, Emily Baylor, Aditya Joshi, Natalia Y. Bilenko, Andrew McNamara, Raphael Gontijo-Lopes, Alex Markham, Evyn Dǒng, Jackie Kay, Manu Saraswat, Nikhil Vytla, and Luke Stark. "Queer In AI: A

Case Study in Community-Led Participatory AI". in: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (2023)
Best paper award @ FAcct, 2023

- Andrew P Creagh, Valentin Hamy, Hang Yuan, Gert Mertes, Ryan Tomlinson, Wen-Hung Chen, Rachel Williams, Christopher Llop, Christopher Yee, Mei Sheng Duh, et al. "Digital health technologies and machine learning augment patient reported outcomes to remotely characterise rheumatoid arthritis". In: *npj Digital Medicine* 7.1 (2024), p. 33

- **Hang Yuan***, Claudia Vanea*, Federica Lucivero, and Nina Hallowell. "Training Ethically Responsible AI Researchers: a Case Study". In: *Navigating the Broader Impacts of AI Research Workshop at NeurIPS* (2020)

- Federica Lucivero, Gabrielle Samuel, Gordon Blair, Sarah J Darby, Tina Fawcett, Mike Hazas, Carolyn Ten Holter, Marina Jirotka, Michael Parker, Helena Webb, and **Hang Yuan**. "Data-driven unsustainability? An interdisciplinary perspective on governing the environmental impacts of a data-driven society". In: *An Interdisciplinary Perspective on Governing the Environmental Impacts of a Data-Driven Society (June 19, 2020)* (2020)

I have authored and contributed to three software packages:

- Author: OxWearables/asleep | Stars: 2 | Forks: 2
- Author: OxWearables/ssl-wearables | Stars: 41 | Forks: 13
- Contributor: OxWearables/biobankAccelerometerAnalysis | Stars: 140 | Forks: 55

I have also presented my work externally and internally

External:

2023 Third Richard Doll Prospective Cohort Studies Symposium, University of Oxford: *Device measured sleep and all-cause mortality.*

2023 World Sleep Congress, Rio de Janeiro, Brazil: *Self-supervised learning yields insights into sleep and all-cause mortality.*

2023 5th MobiUK Research Symposium, Lancaster University: *Self-supervised learning for human activity recognition using 700,000 person-days of wearable data.*

2022 Colby College, Virtual: *Self-supervised learning for wearables.*

2021 ML4Health, NeurIPS, Virtual: junior chair for round table in machine learning for population health.

2021 European Insomnia Network, Virtual: *Predicting polysomnography from accelerometry using machine learning.*

Internal:

2023 Oxford Sleep and Light Initiative, St. Hilda College, University of Oxford: *Towards population inference for light and sleep in large health datasets: a wearable perspective.*

2023 Phenome Seminar, Big Data Institute, University of Oxford: *Time-series machine learning of wearable sleep datasets.*

2023 Oxford Centre for Human Brain Activity Analysis Group, Wellcome Centre for Integrative Neuroimaging, University of Oxford: *Machine learning for time series using mobile sensing data.*

2021 CDT Symposium, Trinity College, University of Oxford: *Understanding sleep's association with cardiovascular diseases using machine learning for wearables.*

2021 OXSTATS: *Machine learning for population health - a wearable perspective.*

1 Introduction

Contents

1.1 Sleep 1
1.2 Sleep and human health 4
1.3 Beyond sleep measurement using questionnaires 6
1.3.1 Self-reported sleep 6
1.3.2 Device-measured sleep 8
1.4 Challenges 10
1.5 Thesis aims and objectives 13
1.6 Thesis outline 13
1.7 Summary 14

1.1 Sleep

Sleep is a reversible state of perceptual disengagement with reduced response to external stimuli [8]. Historically, sleep had been considered as a mere absence of wakefulness, without much differentiation from other physiological states with immobility, such as coma, intoxication and hypnosis. The discovery of the dynamic nature of sleep only began with the observations of electrophysiological activities of the brain, which then led to the differentiation between rapid-eye-movement sleep

(REM) and non-rapid-eye-movement sleep (NREM) [9]. Only then, the pathology of sleep was identified, forming the basis for the field of sleep medicine.

The American Academy of Sleep Medicine (AASM) divides sleep into five distinct stages, consisting of wake, rapid-eye-movement sleep (REM) and non-rapid-eye-movement sleep (NREM) I, II and & III [10]. Sleep cycles through different stages in a single night, and a cycle typically lasts around 90 minutes [11]. For human adults, it is estimated that NREM sleep makes up about 75% to 80% of the total sleep duration and the rest of the time is spent in REM sleep. An individual starts to fall asleep by transitioning from drowsy wakefulness to NREM sleep, which features slightly reduced mobility, reduced eye movement, slower heartbeat and slow-wave electroencephalogram (EEG) patterns. After spending the first hour in NREM, an individual will transition to REM sleep in with features of paralysis, rapid eye movement, fluctuating heartbeat and the presence of fast rhythms and theta waves EEG [12]. Figure 1.1 describes the variations in mobility, heart rate, and brain waves across different sleep stages.

Sleep is necessary for humans. However, the exact function of sleep remains to be determined. Theories have been put forward to explain the function of sleep, including neural maturation, facilitation of learning and memory, cognition, synaptic homeostasis, clearance of metabolic waste and conservation of energy [13–15]. Different sleep stages have been found to have distinct roles in various functions. For example, in learning and memory, NREM II sleep was found to be associated with improvement in motor-skill memories [16]. An example could be in physical restoration, for which sleep contributes to the release of hormones. Most growth

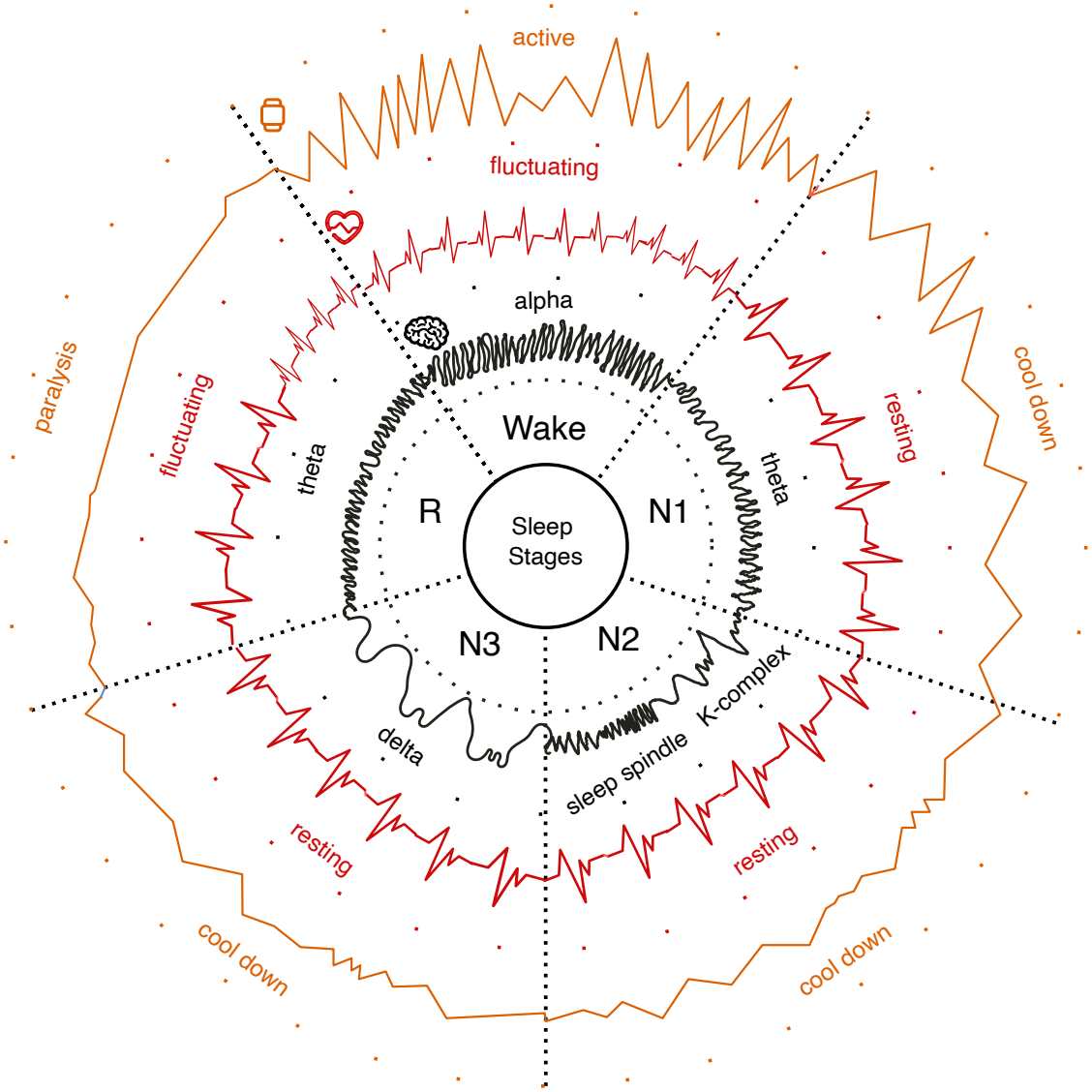


Figure 1.1: Variations in brain wave (black), heart rate (red) and mobility (orange) during different sleep stages. R: rapid-eye-movement sleep. N1, N2, N3: non-rapid-eye-movement sleep 1, 2, 3.

hormone releases occur during slow-wave sleep (NREM III), the majority of which is released shortly after the first phase of slow-wave sleep [17].

In healthy individuals, the length and architecture of sleep change throughout one's lifetime to best support the biological functions of different stages of life [18]. However, in sleep disorder patients, sleep architecture can deviate from its normal form. Table 1.1 describes the common sleep disruptions observed in sleep apnea,

REM sleep disorder, narcolepsy (inability to control sleep-wake cycle), and somnambulism (sleepwalking) [15, 19, 20]. Some of the sleep disorders involve changes in the duration of certain stages. In contrast, others may involve physical movement during certain sleep stages, for instance, acting out dreams with vocalisations in REM sleep, for patients with REM sleep disorder.

Table 1.1: Pathophysiology of sleep stages. R: rapid-eye-movement sleep; N1, N2, N3: non-rapid-eye-movement sleep 1, 2, 3. Summarisation of [15, 19, 20].

Sleep stage	N1	N2	N3	R
Insomnia	-	-	-	↓
Sleep apnea	-	-	↓	↓
REM sleep disorder	-	-	-	Act out dreams with vocalisations
Narcolepsy	-	-	↓	Shorter R onset latency
Somnambulism	-	-	Sleepwalk	-

1.2 Sleep and human health

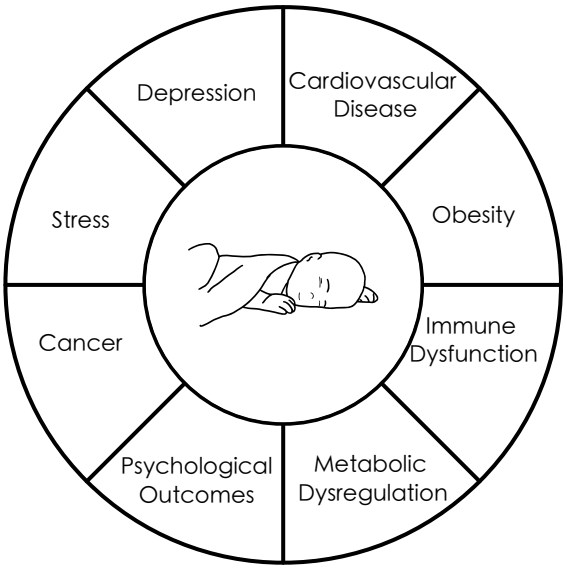


Figure 1.2: Physiological and social outcomes of sleep [21, 22]

Sleep is essential to life. Poor and insufficient sleep has been linked with a wide range of diseases such as cardiovascular diseases, cancer, and depression [22, 23].

However, sleep has been traditionally under-appreciated in population health [24, 25], until the recent decade when sleep health became recognised as a public health research priority in national health agendas [26, 27].

Prior epidemiological studies have found strong associations between short and long sleep duration with all-cause mortality [28, 29]. Typically, a U-shaped association is identified: sleep duration between 7-8 hours per night is considered optimal, and any deviation from this range could lead to an increased risk of mortality. Several pathways have been proposed to explain how short and long sleep could lead to mortality, including cardiovascular diseases (CVD), cancer and stress (Figure 1.2).

Sleep is a multi-dimensional behaviour. Other than sleep duration, sleep quality and timing are also key determinants of sleep health. For instance, the MORGEN Study followed 20,432 men and women between the age of 20-65 with no prior CVD and found that self-reported short sleep had a higher risk of incident CVD compared to normal sleepers, and the risk was even higher in short sleepers who had a poor self-reported sleep quality [30]. Another example could be an investigation of the association between the timing of sleep onset and CVD risks. Even after factoring sleep duration, regularity, and other CVD risk factors, the sleep onset time of 10:00 p.m. - 10:59 p.m. was associated with the lowest CVD risks [31].

However, much of the existing epidemiological understanding of sleep relies on the use of single-item sleep questionnaires, which are limited by: (1) sleep questionnaires are collected at one point in time; (2) only single-item questions about sleep duration to the exact hour is available; (3) the questionnaires only capture sleep duration but not other dimensions of sleep such as sleep continuity,

efficiency, timing or the architecture of sleep. Even though sleep questionnaires measure subjective sleep duration with minimal user burden, the lack of precise measurement from sleep questionnaires could introduce random errors that lead to an underestimation of the effect of sleep on human health [32].

1.3 Beyond sleep measurement using questionnaires

Table 1.2: Characteristics of sleep diary, accelerometry and polysomnography

Type	Device	User burden	Precision	Cost	Modality
Sleep questionnaire	✗	Low	Low	Low	Memory
Sleep diary	✗	Medium	Low	Low	Experience
Accelerometry	✓	Low	Medium	Medium	Movement
Polysomnography	✓	High	High	High	Physiology

Accurate measurement of sleep is at the foundation to reliably assess the association of sleep with health outcomes of interest. When selecting the most suitable sleep measurement approach, one must consider user burden, precision, and cost (Table 1.2).

1.3.1 Self-reported sleep

Sleep questionnaires and diaries are two common ways to report self-monitored sleep in clinical and epidemiological studies [33, 34]. The biggest difference between sleep questionnaires and diaries is the timings of the reports. Retrospective questionnaires are used as the first-line assessment of sleep in primary care and cohort studies, for which the participants are asked about their general sleep duration and quality. Sleep questionnaires are subject to recall bias because the questionnaire responses

are only collected at one point in time [35, 36]. On the other hand, sleep diaries are filled in every day. The diary items need to be filled in at different times throughout the day, namely, before bedtime, about bedtime and after waking up to capture different aspects of sleep [37–39]. Nonetheless, sleep diaries are not available in most cohort studies. Instead, the participants typically are asked to fill in a single-item sleep questionnaire about their sleep duration to the exact hour [40, 41].

In normal sleepers, sleep diaries tend to overestimate sleep duration and efficiency (proportion of total time asleep when in bed) compared with polysomnography, the gold standard for sleep measurement [42]. However, in insomnia patients, sleep diaries underestimate sleep duration instead [43]. Few studies have examined the long-term stability of single-item sleep questionnaires in cohort studies. Even when sleep duration is reported daily using sleep diaries, at least seven consecutive days are needed to obtain adequate reliability [44]. Using single-item sleep questionnaires could introduce random errors, leading to an underestimation of the association between sleep and mortality.

Despite a low correlation between self-reported sleep and polysomnography, self-reported sleep remains an important tool for measuring subjective sleep quality that device-measured sleep cannot capture. Both self-reported and device-measured sleep provide overlapping and some orthogonal dimensions of sleep. Nonetheless, most epidemiological studies consider only self-reported sleep [34], and thus, this thesis focuses on using device-measured sleep in cohort studies that might provide new insights into the association between sleep and human health.

1.3.2 Device-measured sleep

Polysomnography Polysomnography involves concurrently measuring brain activity, heart rate, eye movement, and mobility in a lab environment. Then, a trained clinician will score each 30-second epoch of the recording into one of five sleep stages using the AASM scoring guideline [10] or Rechtschaffen and Kales rules [45]. Because of the high fidelity of the data, polysomnography is the gold standard for sleep measurement in clinics and sleep disorder diagnoses.

The main limitation of laboratory-based polysomnography is that sleeping in a lab environment with many sensors attached may worsen the sleep quality, which might not reflect the sleep quality in habitual sleep [46]. A solution to the sleep variations is to have multiple nights of polysomnography recordings or home-based polysomnography, which is often not feasible.

Wrist-worn accelerometers

Activity count Wrist-worn accelerometers are wristbands, with an inertial measurement unit that tracks the wrist movement of an individual in a passive, continuous fashion. Wrist-worn accelerometers, also referred to as actigraphy in sleep medicine, have long been used for sleep detection [47–50]. Because of limited battery and memory space, many devices used before recorded wrist movement in activity count, a metric that divides movement over a specified time window into different activity levels. Historically, sleep detection was conducted using rule-based methods on activity count data [51–53], which were easy to interpret and implement. Increasingly, sleep studies are moving away from devices that

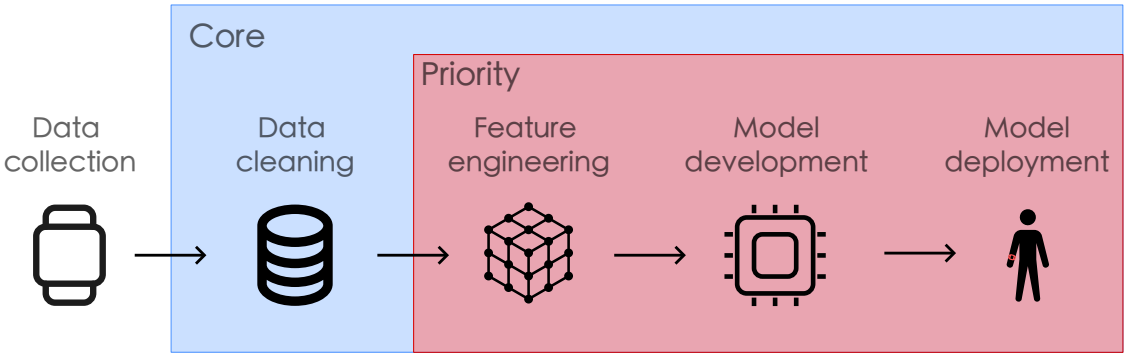


Figure 1.3: The lifecycle of sleep classification model development: this thesis performs secondary data analysis on collected data and focuses on the feature engineering, model development and the deployment of the model in epidemiological applications.

only store activity count to devices that store the raw accelerometry because raw accelerometry has better comparability across devices [54] and the ease of data harmonisation across studies [55].

Raw accelerometry Modern devices can store raw tri-axial accelerometry for up to several weeks without charging, providing opportunities for longitudinal sleep monitoring and analysis with greater data fidelity than activity count. Another advantage of working directly with raw accelerometry is that raw accelerometry can be converted back to activity count, but activity count cannot be converted to raw accelerometry [56].

Thanks to the recent advances in ubiquitous computing, we can leverage wearable devices to measure sleep in a cost-effective manner for population-scale studies. Large-scale wrist-worn accelerometers have been collected in cohort studies, opening up new avenues to investigate the association between behavioural risk factors like sleep and physical activities with important health outcomes [57, 58]. Accelerometer datasets with longitudinal follow-up have provided insights into chronotype [59],

sleep duration [60], and the joint effects of physical activity and sleep on incident diseases [61] and mortality [62].

Photoplethysmography Photoplethysmography (PPG) has also been used heavily in wearable devices such as smartwatches and smart rings. PPG usually contains a pulse oximeter, which illuminates the skin and measures the cardiac cycle by monitoring the changes in the light absorption from the skin. PPG devices can be used with or without the accelerometers for sleep stage classification. The raw PPG signal is often directly converted to heart rate as the input for the sleep classification model. Because of the variations of the heart rates at different stages of sleep, PPG-based devices have been shown to have a higher performance than accelerometer-based sleep staging devices [63, 64]. This thesis will not consider PPG-based sleep detection because of the lack of a large-scale PPG dataset with longitudinal health follow-up.

1.4 Challenges

This thesis thus focuses on using wrist-worn accelerometers to measure habitual sleep as they are being made available in large-health datasets such as the UK Biobank and NHANES [57, 58], which allow for subsequent health association analyses. Challenges persist in translating the raw accelerometry data into health insights.

Limited labelled data One major challenge in classifying sleep from wrist-worn accelerometers is a lack of sufficiently large laboratory-based polysomnography to be used as ground truth. One night of lab-based polysomnography could cost up to

1,000 pounds, making it difficult to obtain large-scale lab-based polysomnography dataset especially with concurrent wrist-worn accelerometer. Others have achieved successes in using self-supervised learning on unlabelled datasets to improve the performance of accelerometers-based human activity recognition with limited labelled data by obtaining a pre-trained network as feature extractor [65–67]. The utility of self-supervised learning in sleep detection is not clear, which is a research aim of the thesis.

Sleep staging using wrist movement Prior works have derived common sleep parameters, including sleep window, the time window when dominant sleep happens, and sleep duration, the amount of time spent in sleep using methods ranging from simple heuristics to machine learning techniques [68–70]. While most studies emphasise the classification between sleep and wakefulness, few studies have tried to infer the stages of sleep from raw accelerometry as sleep stagings from movement signals are much more challenging [71, 72]. Sleep stages exhibit different physiologic changes and metabolic functions. Currently, there have been few comprehensive studies to investigate the structural and electrophysiological functions of sleep. Going beyond the sleep-wake classification to sleep stage classification will be a step change in understanding the significance of sleep health. As shown in Section 1.1, sleep stages feature more pronounced differences in the electrophysiologic signals instead of wrist movement, potentially limiting the signal-to-noise ratio for accelerometers in sleep staging. Traditional sleep detection algorithms only use accelerometry from each 30-second window to be consistent with the polysomnography definition.

Methods that factor in longer sequences of data could improve the classification performance. Therefore, the second aim of this work is to establish the upper limit for accelerometry-based sleep stage classification with the eventual goal of deploying the model on large-scale datasets for health association analysis. The inter-rate agreement for polysomnography is around a Kappa score of 0.8 for wake/NREM/REM and a Kappa score of 0.6 for wake/NREM I, II, III/REM [73]. Even though we cannot obtain the same level of agreement from accelerometer-based sleep staging, achieving a Kappa score greater than 0.4 is desirable for a moderate agreement [74].

Associating device-measured sleep with mortality Existing studies have a primary focus on using self-reported sleep in epidemiological analysis [34]. Numerous analytical challenges remain when determining the representative sleep parameters to capture true habitual sleep. Examples include design choices for inclusion criteria for usable accelerometer data, expected population variations to confirm face validity, and the appropriate groupings of continuous sleep parameters self-reported on a dichotomous scale. This thesis will evaluate the utility of device-measured sleep for epidemiological analysis by investigating its association with mortality risk.

1.5 Thesis aims and objectives

This thesis sets out to improve the measurement of sleep and sleep staging using wrist-worn accelerometers to estimate the true effect of sleep on human health. To accomplish this aim in light of the challenges outlined above, I will investigate the following questions:

- How can we use unlabelled data to improve the performance of accelerometer-based sleep detection?

- How reliably can we classify sleep and the stages of sleep using wrist-worn accelerometers alone?

- Is there an association between accelerometer-measured sleep and all-cause mortality?

1.6 Thesis outline

The outline of this thesis is as follows:

Chapter 2 systematically reviews the validity of wrist-worn accelerometers for sleep stage classification.

Chapter 3 provides an overview of all the datasets used in thesis and the data preprocessing pipeline.

Chapter 4 investigates how to develop a self-supervised pre-trained model using a large-scale unlabelled dataset to improve the performance of human activity recognition and sleep.

Chapter 5 develops a self-supervised recurrent neural network for sleep stage classification and quantifies the classification error in population subgroups for the common sleep parameters.

Chapter 6 investigates the association between device-measured sleep and mortality outcomes by deploying the developed sleep classification model on the UK Biobank.

Chapter 7 synthesises the findings in this thesis, and interprets the results in the broader context of wearable sensing, machine learning and public health.

1.7 Summary

Sleep is an important modifiable risk factor for diseases and yet under-valued in public health research. Large-scale accelerometer datasets with longitudinal follow-up make it possible to develop a better understanding of the association between sleep and important health outcomes. However, challenges remain in accelerometer-based sleep classification and the interpretation of their use in subsequent health association analysis. This thesis aims to improve the measurement of sleep and evaluate how device-measured sleep could provide new insights into health using mortality outcomes as an example. The next Chapter describes a systematic review of the current state-of-the-art methods in classifying sleep stages using wrist-worn accelerometers and identifies potential areas for improvement.

2

Systematic review of the performance of wrist-worn accelerometers in measuring sleep stages

Contents

2.1 Review goals 16
2.2 Review protocol 17
2.2.1 Eligible studies 17
2.2.2 Search strategies 20
2.2.3 Study selection and data extraction 20
2.2.4 Data evaluation 21
2.3 Review results 24
2.3.1 Study populations 24
2.3.2 Accelerometer devices 24
2.3.3 Sleep staging algorithms 27
2.3.4 Performance metrics and assessment 27
2.3.5 Risk of bias of included studies 30
2.4 Discussions 31
2.5 Summary 33

The materials for this chapter have been made available online under the following citation: Yuan H., Hill A. L., Kyle D. S., and Doherty A. "A systematic review of the performance of actigraphy in measuring sleep stages". medRxiv, 2023 [3]

2.1 Review goals

Wrist-worn accelerometers have long been used to measure sleep-wake patterns in clinical and daily living environments. The broad adoption of accelerometry is driven by its relatively low cost, minimal user burden and ease of continuous recordings over weeks [75]. The American Academy of Sleep Medicine (AASM) published a meta-analysis on accelerometers for sleep detection in comparison with polysomnography and concluded that accelerometers were suitable to assess total sleep time and sleep onset latency for clinical decision-making in adults with sleep disorders [76]. Extensive evaluations of accelerometers have been done on their population validity [77, 78], ecological validity [79] and device validity [80, 81] for sleep-wake classification. Fundamentally, accelerometers and polysomnography measure different dimensions of sleep. polysomnography measures various physiological and movement signals, including brain activity (electroencephalography, EEG), ocular movement (electrooculography, EOG), muscle functions (electromyography, EMG) and heart rhythm (electrocardiography, ECG), with the combination of EEG, EOG, and EMG forming the basis of objective determination of sleep stages in line with published international guidelines [82]. In contrast, accelerometers measure peripheral limb movement alone, using activity as a proxy for wake and inactivity as a proxy for sleep, and by extension, have limited capacity to infer the stages of sleep due to the lack of physiological signals.

Nonetheless, the validity of accelerometers for sleep staging is relatively unknown. Understanding the performance limits of accelerometer-based sleep staging may open the door to investigation of the association between sleep architecture and important clinical outcomes in large-scale biobank datasets, in which only wrist-worn accelerometers are available [41, 57]. The most relevant systematic review evaluated wearable sensing technologies, including wrist-worn accelerometers for sleep staging [83], which had a broad coverage of different data modalities but a limited discussion for each sensing technology and the technical parameters that could influence the algorithm performance for each device.

Therefore, this review sets out to 1) systematically evaluate the validity of sleep staging using wrist-worn accelerometers and 2) analyse the effect of technical parameters on sleep staging performance.

2.2 Review protocol

This review was conducted according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) guidelines [84]. The protocol was registered with the International Prospective Register of Systematic Reviews (PROSPERO) following the PRISMA-P guidelines (PROSPERO 2021 CRD42021237456) prior to commencement.

2.2.1 Eligible studies

Studies were included if they (1) compared the sleep staging agreement between wrist-worn accelerometers and polysomnography for adults, (2) reported agreement results on AASM sleep stages (wake (W)/non-rapid-eye-movement sleep (NREM)

1, 2, and 3 (N1, N2, and N3)/rapid-eye-movement sleep (R)) or performance on the composite of different sleep stages, and (3) scored the polysomnography manually or used an automated polysomnography scoring first and then corrected the scoring by a human rater. Studies were included regardless of the polysomnography recording montage used, provided the scoring followed the AASM guidelines.

The exclusion criteria were: (1) editorials, reviews, or commentaries; (2) single-channel EEG as the reference when scoring sleep stages due to the lack of reliability; (3) studies that only reported the agreement performance for sleep and wakefulness; (4) scoring using Rechtschaffen and Kales rules [45] which is subject to greater inter-rater variability as compared to the AASM standard [85]. Table 2.1 lists the criteria for selecting eligible studies in detail.

Table 2.1: Eligibility criteria: ACC, accelerometer.

	Inclusion	Exclusion
Literature language	English	Not English
Literature type	Empirical studies	Non-empirical works such as: • Editorials • Reviews • Commentaries
Study design	Studies with parallel ACC and polysomnography.	Studies that have only ACC or polysomnography.
Populations	Adults age $\geq$ 18 years	Pediatric or adolescent populations with age $<$ 18 years
Measurements	Wrist-worn ACC and (polysomnography or EEG). Level I/II (attended/unattended) polysomnography.	• Studies that do not have ACC • Studies that only has either ACC or (polysomnography or EEG) • Studies that have both ACC and only single-channel EEG.
polysomnography scoring	Only consider manual or semi-automated (initially scored automatically and then corrected by a human rater)	fully-automated scoring
Sleep parameters	Should at least report one of the following • N1 • N2 • N3 • REM (R) • NREM • Other metrics, being composed by any of the above-mentioned measures, that are more specific than sleep.	Only reports sleep vs wake results or does not report any of the following • N1 • N2 • N3 • REM (R) • NREM • Other metrics, being composed by any of the above-mentioned measures, that are more specific than sleep.

2.2.2 Search strategies

The searches were performed using MEDLINE (Ovid), Web of Science, Google Scholar and Embase (Ovid) on 12 April 2021. Preprints that appeared in Google Scholar were also included to mitigate publication bias. The exact search terms for each database are specified in Appendix Section A.2. The search strategy was adapted from previous work [78], which aimed to assess the agreement between wrist-worn accelerometers and polysomnography using parameters related to sleep and wake patterns. To focus more on the performance of sleep staging, two changes were made to the search terms: (1) included more research-grade and consumer-grade devices to capture a broader set of studies using different wrist-worn accelerometers (2) added "sleep stage" to the search terms to make the search results relevant to sleep staging. In developing the search syntax for each database, the Ovid syntax was first determined for Embase and MEDLINE, which was then translated into the search terms to other databases.

2.2.3 Study selection and data extraction

Literature search results were uploaded to Covidence (Covidence, Melbourne, Australia), an Internet-based systematic review program that assists reviewers in streamlining the review process. Covidence automatically flagged any possible reference duplication. The duplication removal was manually verified. I also conducted the literature search and abstract screening. A second reviewer (EAH) independently verified that all the included abstracts meet the selection criteria. Full-text screening and data extraction were performed by both the second reviewer

and me independently in duplicate. The second reviewer and I undertook initial reconciliation, with a third reviewer resolving any outstanding disagreement.

The key data fields for extraction included sleep staging outcomes, device specifications (device type, data modality and sampling frequencies), data acquisition environment, and population characteristics (sample size, age, sex, and existing comorbidities). Information related to the sleep staging algorithms, including algorithmic techniques and validation methods, was also extracted. When multiple algorithms were reported in a study, data were extracted only for the algorithm with the best performance. If different datasets were used for algorithm development and validation, then the performance of the validation dataset was extracted. Appendix Section A.1 lists detailed data items. The data extraction form was piloted on a subset of studies during the screening process.

2.2.4 Data evaluation

The risk of bias was evaluated independently by a second reviewer and I using a modified QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies tool)[86]. Although QUADAS-2 has mainly been designed for primary diagnostic tools, the four key domains in QUADAS-2 are pertinent to the focus of the review: 1) patient selection, 2) index test, 3) reference standard, and 4) flow and timing. The key domains reflect whether all the participants received the same reference standard and participant inclusions throughout the analyses.

Domains 2 and 3 also reflected concerns around algorithmic validity. The following irrelevant signalling questions were removed from the original QUADAS-2:

q2 in patient selection, q1 and q2 in index test, q2 in reference standard and q1 in flow and timing. Two additional questions were included to better assess algorithmic bias: q1 in the index test and q2 in the reference standard. The risk of bias is judged as “low”, “high”, and “unclear.” The risk of bias is judged as “low” if answers to all signalling questions are “yes”; otherwise, the risk of bias is judged as “high.” When the information available does not permit answering the signalling questions, then the risk is judged as “unclear.” Table 2.2 entails the modified QUADAS-2 in full.

Table 2.2: Modified QUADAS-2 form: Bold text questions are newly added.

Domain	Patient selection	Index test	Reference standard	Flow and Timing
Description	Describe methods of patient selection: Describe included patients (prior testing, presentation, intended use of index test and setting):	Describe the index test and how it was conducted and interpreted:	Describe the reference standard and how it was conducted and interpreted:	Describe any patients who did not receive the index test(s) and/or reference standard or who were excluded from the 2x2 table (refer to flow diagram): Describe the time interval and any interventions between index test(s) and reference standard:
Signalling questions (yes/no/unclear)	1. Was a consecutive or random sample of patients enrolled? 2. Did the study avoid inappropriate exclusions?	**1. Was the index test obtained using a known algorithm?**	1. Is the reference standard likely to correctly classify the target condition? **2. Was the reference standard interpreted using all participants as the test sample?**	1. Did all patients receive a reference standard? 2. Did all patients receive the same reference standard? 3. Were all patients included in the analysis?
Risk of bias: High/low/unclear	Could the selection of patients have introduced bias?	Could the conduct or interpretation of the index test have introduced bias?	Could the reference standard, its conduct, or its interpretation have introduced bias?	Could the patient flow have introduced bias?
Concerns regarding applicability: High/low/unclear	Are there concerns that the included patients do not match the review question?	Are there concerns that the index test, its conduct, or interpretation differ from the review question?	Are there concerns that the target condition as defined by the reference standard does not match the review question?	

2.3 Review results

The literature search identified 5,180 records from the databases, 2,052 of which were removed due to duplication by Covidence (Figure 2.1). Another 3,128 records were removed after the abstract screening. Among the remaining 79 studies for full-text screening, wrong model input (n=25), conference abstract only (n=14), and wrong sleep parameters (n=14) were the top three reasons for exclusion before data extraction. In total, seven studies met the inclusion criteria and were included in this review. Appendix Table A.3 summarises the exact reasons for removal after the full-text screening.

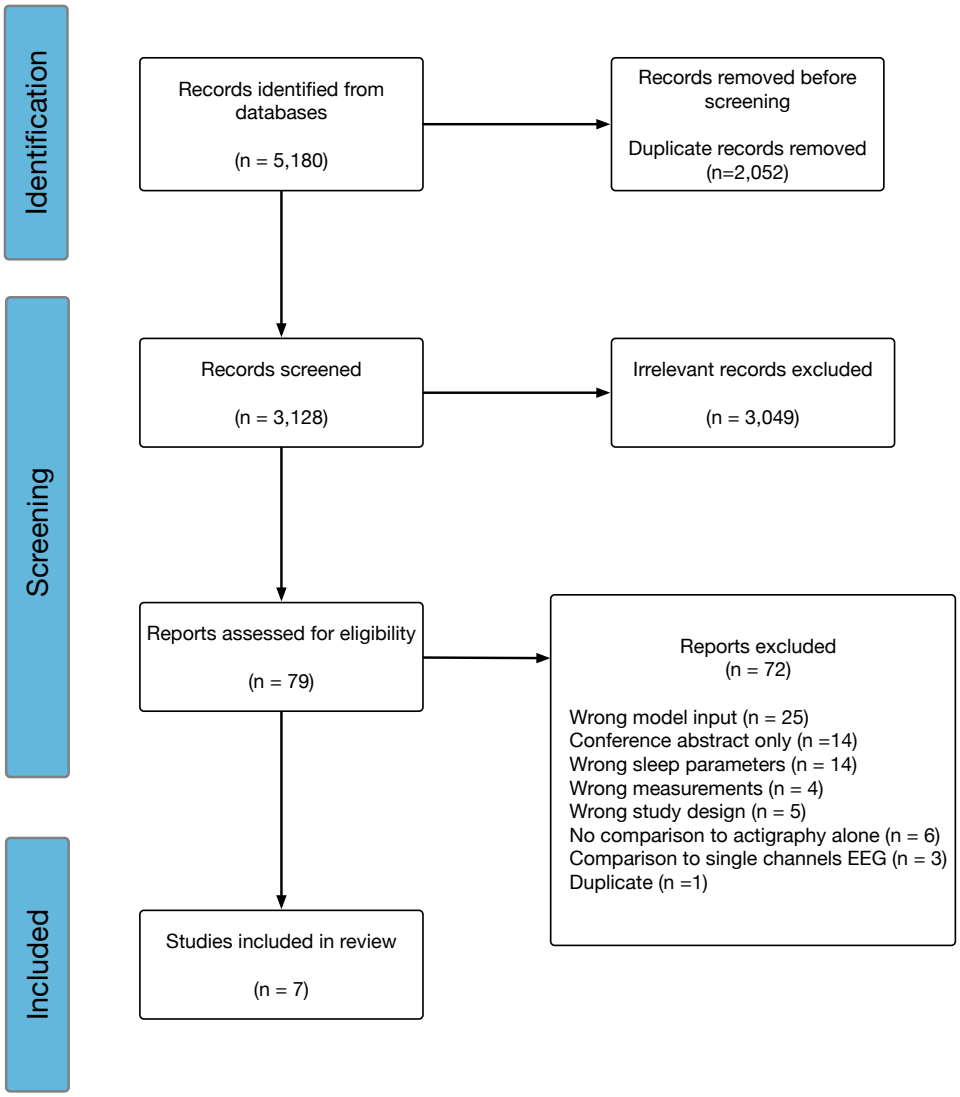
2.3.1 Study populations

The population characteristics are listed in Table 2.3. The majority of studies used data from healthy participants [64, 71, 87–90], whereas one study assessed the sleep staging performance in participants with sleep disorders [91]. The number of participants included in each study tended to be small. All but two studies included fewer than 100 participants in their assessment [71, 91], both of which reported the sleep classification performance using cross-fold validation without external validation.

2.3.2 Accelerometer devices

The full specifications of the accelerometer devices can be found in Table 2.3. Each study used a distinct type of device. Both research-grade and consumer-grade devices were used. There was no consensus on the device placement. Three studies placed the accelerometer on the non-dominant wrist; one study placed the device on

Figure 2.1: PRISMA flow diagram of studies screened and included to review the validity of wrist-worn accelerometers for the measurement of sleep stages.



the left wrist; one study placed the device on both wrists, and two studies did not report the device placement. Both raw data and activity count were used for sleep staging inference: three studies used the raw data, two studies used activity count, and two studies did not report the data type used. The accelerometer sampling rate for different devices ranged from 30 Hz to 62.5 Hz.

Table 2.3: Study characteristics.

Reference	Participant	Age	Sex	Polysomnography		Accelerometer			
	N	Range/mean (SD)	N (%)	Level and scoring guidelines	Setting	Location	Data type	Sampling rate	Device
Nguyen et al., 2021	22	18-36	14 (63.6%) F	Attended AASM 2012 [92] 30 nights in total	Lab	Left wrist	NA	NA	Unknown
Walch et al., 2019	31 Healthy	29.4 (8.5)	21 (67.7%) F	Attended AASM 2016 [93] One night per person	Lab	NA	Raw data	50 Hz	Apple Watch (Series 2 and 3, Apple Inc., Cupertino, USA)
Trevenen et al., 2019	816 Healthy	22.1 (.6)	413 (50.0%) F	Attended AASM 2012 [92] One night per person	Lab	Non-dominant wrist	Raw data	30 Hz	ActiGraph GT3X+ Monitor (ActiGraph, Pensacola, USA)
Sundararajan et al., 2021	158 (70 sleep disorders and the rest healthy)	20-72	Not stated	Attended Not stated	Lab	NA	Raw data	50 Hz	Mostly GENEActiv (ActivInsights, Kimbolton, UK)
Mantua, Gravel, and Spencer, 2016	40 Healthy	22.4 (4.9)	19 (47.5%) F	Attended AASM 2007 [94] One night per person	Home	Both wrists	Activity count	NA	Misfit Shine (Misfit Wearables, San Francisco, USA)
Devine et al., 2020	8 Healthy	30.4 (3.2)	4 (50%) F	Attended AASM 2012 [92] Three nights per person	Lab	Non-dominant wrist	NA	NA	Zulu watch (Institutes for Behavior Resources, Baltimore, USA)
Boe et al., 2019	12 Healthy	27.4 (3.9)	5 (41.7%) F	Attended AASM 2007 [94] One night per person	Lab	Non-dominant wrist	Activity count	62.5 Hz	ActiWatch Spectrum (Philips Respironics; Murrysville, USA)

2.3.3 Sleep staging algorithms

Table 2.4 shows the algorithms used and the corresponding performance from each study. Seven studies reported algorithm details. Studies that used commercial-grade devices relied on the proprietary algorithms from the vendors which do not reveal the methods for sleep stage classification [88, 89]. Heuristic-based methods were not used by any study. Statistical learning methods were the predominant method class, including logistic regression [64], Hidden Markov Model [71], and tree-based methods [90, 91]. Deep learning techniques were also used in one study [91].

2.3.4 Performance metrics and assessment

Several performance metrics were used to report the sleep stage outcome, including sensitivity, specificity, accuracy, F1 and the area under the receiver operating characteristic (AUROC). Studies tended to report performance with different groupings of sleep stagings (Table 2.4). Hence, it is not possible to synthesise the evidence quantitatively. Only two studies reported the performance for all five AASM sleep stages [71, 91]. The highest average five-class performance was 43.2% in F1 score [71]. As the number of classes decreases, the reported performance increases. For instance, an accuracy of 66.4% was reported in a three-class setting (W vs N1+N2 vs N3+R) [87]. Studies using consumer-grade devices reported sleep staging performance by "light sleep" and "deep sleep" that were defined by the device vendors, making it impossible to compare sleep staging performance with other devices.

Method development studies used N-fold/leave-one-out cross-validation [64, 71, 90, 91]. Device validation studies used statistic testing or directly calculated the performance metrics [87–89]. Method validation studies aimed to evaluate the performance of different algorithms using data collected from one device, whereas device validation studies aimed to assess the performance of different devices against the reference standard using the default algorithm from each device.

Table 2.4: Sleep staging algorithms and performance. Wake: W; Non-rapid-eye-movement sleep 1, 2, and 3: NREM, N1, N2, and N3; rapid-eye-movement sleep: R; polysomnography: polysomnography.

Reference	Algorithm	Validation	Sensitivity (%)	Specificity (%)	Accuracy (%)	F1 (%)	AUROC
Nguyen et al., 2021	Not stated	Paired sample t-test	W: 27.9 N1+N2: 41.4 N3+R: 65.8	W: 98.6 N1+N2: 62.1 N3+R: 51.6	Average: 66.4	NA	NA
Walch et al., 2019	Logistic regression	Leave one out cross-validation	NA	NA	W: 60 NREM: 50.6 R: 33.2	NA	NA
Trevenen et al., 2019	Hidden Markov model	Leave one out cross-validation	NA	NA	NA	W: 61.0 N1: 10.0 N2: 52.2 N3: 63.8 R 6.4 Average: 43.2 [a]	NA
Sundararajan et al., 2021	Random forest	N-fold cross-validation	NA	NA	W: 60.7 N1: 1.0 N2: 57.9 N3: 27.9 R: 17.5	W: 55.1 N1: 4.2 N2: 53.2 N3: 20.5 R: 12.3	NA
Mantua, Gravel, and Spencer, 2016	Proprietary algorithm	Not stated	Light: 23.4 Deep: 69.6 [b]	NA	NA	NA	NA
Devine et al., 2020	Proprietary algorithm	Not stated	Light: 10.7 Deep: 84.2	Light: 88.2 Deep: 30.0	Light: 51.7 Deep: 48.7 [c]	NA	NA
Boe et al., 2019	Bagged trees	Leave one out cross-validation	NA	NA	NA	NA	W: .82 N1&N2: .58 N3: .64 R: .47 [d]

[a] The standard deviation for each class is: W (28.1), N1 (8.7), N2 (46.0), N (65.8), R (36.7).
[b] The average percentage errors are reported for the duration. Percentage errors: (Device duration - polysomnography duration)/polysomnography duration. Light: N1+N2; Deep sleep: N3+R.
[c] Each AASM polysomnography score is mapped from 0 (wake) to 3 (N3+R). Light sleep: any two-minutes bins with an average score between 1 and 2.55. Deep sleep: any two-minute bin with an average score greater than 2.5.
[d] The standard deviation for each class is: W (.18), N1&N2 (.05), N3 (.1), R (.09).

Table 2.5: Risk of bias using assessment Quality Assessment of Diagnostic Accuracy Studies tool (QUADAS-2).

Study	Risk of bias				Applicability concerns		
	Patient selection	Index test	Reference standard	Flow and timing	Patient selection	Index test	Reference standard
Nguyen et al., 2021	(-)	?	(-)	(-)	?	(-)	(+)
Walch et al., 2019	(-)	(+)	(+)	(+)	(+)	(+)	(+)
Trevenen et al., 2019	?	(+)	(+)	(+)	(+)	(+)	(+)
Sundararajan et al., 2021	?	(+)	(+)	?	?	(+)	(+)
Mantua, Gravel, and Spencer, 2016	(-)	(-)	(+)	(+)	(+)	(+)	(+)
Devine et al., 2020	(-)	(-)	(+)	(+)	(+)	(+)	(+)
Boe et al., 2019	(+)	(+)	(+)	(+)	(+)	(+)	(+)

2.3.5 Risk of bias of included studies

Overall, the risk of bias for the included studies based on the modified QUADAS-2 was limited (Table 2.5). The greatest risk of bias came from the participant selection: just one study with low risk, four studies with high risk and two studies with unclear risk. The next most common risk of bias was the index test having two studies with a high risk of bias, one with an unclear risk of bias and four with a low risk of bias. Several studies aimed to develop sleep staging algorithms, which included limited information about participant inclusion and did not consider appropriate participant inclusion for clinical use. Two studies that carried a high risk of bias for the index test domain used consumer-grade devices [88, 89] that relied on proprietary staging algorithms, which could introduce bias. All but one study carried a low risk of bias for the reference standard polysomnography.

2.4 Discussions

This review summarised the literature on the use of wrist-worn accelerometers in sleep staging. Seven studies were included, and only one study tested the validity of sleep staging in sleep patients. A large heterogeneity was found in the reporting of the performance in accelerometer-based sleep staging, especially around the sleep stage groupings and evaluation metrics, thus making it challenging to summarise existing evidence quantitatively. Nonetheless, based on the available studies, accelerometer-based sleep staging appears to have some ability to classify sleep stages.

Most of the performance evaluation has been conducted in young, healthy populations of limited size ($n < 100$). Only two of the included studies evaluated the sleep staging performance in more than 100 participants [71, 91]. Sundararajan et al. [91] was also the only study that evaluated the performance of accelerometers in patients with sleep disorders. Further evaluation of sleep staging performance in older adults and chronic conditions will help assess the clinical and scientific utility of accelerometers for sleep staging. However, the lack of open-access anonymous datasets limits the participant number and diversity for each study. Only one of the included studies used a dataset that was publicly available [91]. As a community, more effort should be put into making the data open access to allow for secondary data analysis.

The reporting of the sleep staging performance has been heterogeneous. Even though five sleep stages are defined by AASM, because of the difficulty in the sleep stage identification from accelerometers, several sleep stages are combined into one to improve the agreement with polysomnography. For example, Walch

et al. [64] conducted the evaluation in a three-class setting (W/N1+N2/N3+R). While sleep stage grouping will help to obtain better performance, five-class staging performance needs to be reported to aid the comparison across different studies. Furthermore, studies included used a range of evaluation metrics for multi-class classification. Future research will benefit from adopting existing reporting guidelines to increase performance comparison across studies [95]. In particular, researchers should aim to report the staging performance at the individual level despite the grouping preference. In addition to reporting the typical performance metrics like sensitivity and specificity, metrics robust to class imbalances, like F1 and Kappa scores, are encouraged.

Unlike sleep-wake classification for which heuristic-based methods are often used [52, 68, 96], all sleep staging methods reported here are based on statistical learning partly due to the difficulty of relying on pre-defined rules to infer sleep stages directly from wrist movement. One study leveraged recent advances in artificial intelligence (AI), such as deep neural networks, but, the model performance was limited [91]. Data-driven methods like deep neural networks are particularly data-hungry and it would probably require thousands of nights worth of polysomnography data for deep neural networks-based methods to reach their potential. Because of the costs of the polysomnography, it will not be feasible to collect concurrent polysomnography and accelerometers at scale. Nonetheless, large volumes of accelerometer data are available, allowing the use of recent advances in deep learning to obtain high-quality feature extractors for sleep staging.

Multi-modal methods for sleep staging predictions could help understand the value of each modality. Twenty-five references were excluded because of the wrong model input, and six studies were excluded because of no comparison to accelerometers alone. These groups of studies often used additional modalities such as heart rate or ECG to obtain better performance [97–99]. Reporting modality-specific performance in a multi-modal setting will inform the best modalities to use.

Limitations

These results are likely to suffer from publication bias. It is possible that studies that could not classify sleep stages from accelerometers have not been published, particularly in individuals with chronic conditions for whom the classification will be more complex. Records were screened by a single reviewer in the initial abstract screening stage instead of two, as outlined in the study protocol, which may have inadvertently introduced some selection bias. However, all further steps in the review process were completed by two independent reviewers. Finally, this review did not consider the performance difference for studies with using cross-fold validation or external validation using held-out set due to the limited number of studies.

2.5 Summary

Even though the included studies carry a limited risk of bias for the validity of sleep staging from wrist-worn accelerometers, clear methodological gaps remain, including:

- The reliance on hand-crafted features used by the sleep detection algorithms which may not be sufficient to capture the association between accelerometry and sleep stages;

- The small sample size of polysomnography studies which prevents using deep learning techniques that are data-hungry but could potentially capture the complex relationship between accelerometry and sleep stages.

Hence, Chapter 4 aims to develop a feature extractor based on self-supervised learning to remove the need for hand-crafted accelerometry features. And Chapter 5 aims to establish the upper limit of accelerometry-based sleep staging using one of the largest datasets with multicentre accelerometers and polysomnography data using a deep recurrent neural network. Finally, Chapter 6 will highlight how the sleep stage classifier could be used in an epidemiological analysis using the UK Biobank to investigate the association between sleep and all-cause mortality.

3

Raine Study and UK Biobank

Contents

3.1 Raine Study 36
3.1.1 Recruitment 36
3.1.2 Accelerometer and polysomnography follow up 37
3.1.3 Population characteristics 37
3.2 UK Biobank 40
3.2.1 Study protocol 40
3.2.2 Accelerometry in the UK Biobank 42
3.2.3 Population characteristics 43
3.3 Summary 47

The systematic review in Chapter 2 highlights the need to obtain datasets with concurrent accelerometers and polysomnography of large volumes to develop accurate sleep measurement models using recent advances in deep learning. Thus, this Chapter starts by describing one of the largest datasets with concurrent accelerometers and laboratory-based polysomnography data, the Raine Study, which is used to develop the sleep stage classifier. Then, this Chapter introduces the UK Biobank study, which is used throughout this thesis.

3.1 Raine Study

The Raine Study started off as a randomised controlled trial to investigate the influence of serial fetal ultrasounds on birth outcomes [100]. The Raine Study embarked in 1989 with an enrolment of 2,900 pregnant women (Generation 1) in Perth, Western Australia. A range of phenotypes and genomics data have been measured in Generation I and their children (Generation 2). The initial study aimed to understand the developmental origins of health and diseases, which then evolved into a life-course investigation of multiple interactions between genetics, behaviours and the environment of West Australians [101, 102].

3.1.1 Recruitment

Between May 1989 and November 1991, pregnant women at King Edward Memorial Hospital and its surrounding private clinics were invited to participate. To be eligible, one needed to be between 16 and 20 weeks of pregnancy, have fluency in English and intend to deliver at the hospital and stay in Western Australia [101]. A total of 2,900 women in Generation 1 were eventually enrolled. The cohort was regularly followed up pre-birth and post-birth: at week 18, week 38 during pregnancy and year 1, 2, 3, 5, 8, 10, 14, 20, 22, and 23 post birth. The Raine Study participants were shown to be representative of the populations in Western Australia at the recruitment sites after a comparison of demographic information between the Raine Study participants and the population census data.

3.1.2 Accelerometer and polysomnography follow up

During the year 23 follow-up, both Generation 1 and Generation 2 were invited to participate in an overnight sleep assessment using laboratory-based polysomnography. When the participants arrived at the facility between 5 pm and 7 pm, they first had to undergo a range of physical tests. During the period before sleep, the participants were instructed not to drink caffeinated drinks or use screens. Smoking was permitted under special permission. They were instructed to follow good sleep hygiene guidelines by settling down, maintaining a calm atmosphere and turning the lights off. The polysomnography setup began by 8:30–9:00 pm, and the recording was finished by 7 am the following day. The participants wore a concurrent wrist-worn accelerometer device (GT3X) throughout the night (Figure 3.1). In total, 865 nights of concurrent laboratory-based polysomnography and raw accelerometry data were obtained from Generation 1 and another 795 nights of data were obtained from Generation 2.

The polysomnography data was scored by certified sleep technologies following the American Academy of Sleep Medicine protocol [103]. The sleep scorers were asked to score the same set of participants to harmonize their scoring criteria before scoring the collected polysomnography from the Raine Study.

3.1.3 Population characteristics

The population characteristics of the participants from the Raine study are described in Table 3.1. Overall, generation 1 participants from the Raine study received more years of education (62.7% vs 43.3% in higher education), had a higher Body



Figure 3.1: Accelerometer used in the Raine study: The image is taken from [104].

Mass Index (BMI, 31.4% vs 15.1% in the obese category) and a higher apnea hypopnea index (AHI, 28.4% vs 2% with severe sleep apnea). Even though the Generation 1 participants had a higher BMI and AHI, generation 1 participants enjoyed greater self-reported health (12.3% vs 8.6% in the excellent category) compared to Generation 2. The sleep classifier developed will be trained using data from both generation 1 and generation 2 to ensure its generalisability. Generation 1 participants from the Raine Study has comparable demographic to the participants from the UK Biobank, making it an ideal dataset to train on in preparation of the health association analysis later on.

In addition to the Raine study, several other smaller studies have also collected concurrent laboratory-based polysomnography and raw accelerometry [105–108]. Most existing datasets are not publicly available due to the high costs of data collection and the concerns around privacy and data governance. Chapter 5 will elaborate on the demographic and polysomnography specifications for the additional datasets made available to this Thesis. The next section will provide an overview of the UK Biobank that is being used for both the development of a feature-extractor and the subsequent health association analysis.

Table 3.1: Population characteristics for both generation 1 and generation 2 from the Raine study. The income listed is in Australian dollar.

Characteristics	Generation 1	Generation 2
	n=1023	n=952
Age, year		
Mean (SD)	56.7 (5.68)	21.6 (0.71)
Median [Q1, Q3]	56.7 [52.9, 60.3]	22.0 [21.0, 22.0]
Sex		
Female	587 (57.4%)	480 (50.4%)
Male	436 (42.6%)	472 (49.6%)
Education		
School leaver	7 (0.7%)	0 (0.0%)
Further education	328 (32.1%)	440 (46.2%)
Higher education	641 (62.7%)	412 (43.3%)
Missing	47 (4.6%)	100 (10.5%)
Income (per week)		
< $ 1,000	343 (33.5%)	-
$ 1,000 - $ 1,999	338 (33.0%)	-
$ 2,000+	196 (19.2%)	-
Not stated	83 (8.1%)	-
Missing	47 (4.6%)	-
BMI		
<18.5, underweight	7 (0.7%)	26 (2.7%)
18.5-24.9, normal	289 (28.3%)	556 (58.4%)
25-29.9, overweight	403 (39.4%)	225 (23.6%)
30+, obese	321 (31.4%)	144 (15.1%)
Missing	3 (0.3%)	1 (0.1%)
Self-rated health		
Poor	17 (1.7%)	20 (2.1%)
Fair	109 (10.7%)	111 (11.7%)
Good	361 (35.3%)	351 (36.9%)
Very good	360 (35.2%)	285 (29.9%)
Excellent	126 (12.3%)	82 (8.6%)
Missing	50 (4.9%)	103 (10.8%)
Apnea Hypopnea Index (AHI)		
< 5	248 (24.2%)	660 (69.3%)
5 - 14.9	232 (22.7%)	219 (23.0%)
15 - 29.9	197 (19.3%)	54 (5.7%)
≥ 30	291 (28.4%)	19 (2.0%)
Missing	55 (5.4%)	0 (0.0%)

3.2 UK Biobank

Large-scale accelerometer datasets with longitudinal follow-up have also only been made available recently, allowing for the investigation of the dose-response relationship between device-measured behaviour traits and diseases [57, 58].

The UK Biobank project is a prospective cohort study with longitudinal follow-up tracking 500,000 participants in middle and older age adults in England, Scotland and Wales to study the risk factors of major diseases [109]. The UK Biobank contains rich information about each participant, ranging from questionnaires and physical measurements to biological samples, including genetics, proteomics and metabolomics. The participants are being followed up in further enrichments and health linked electronic healthcare records.

3.2.1 Study protocol

Recruitment

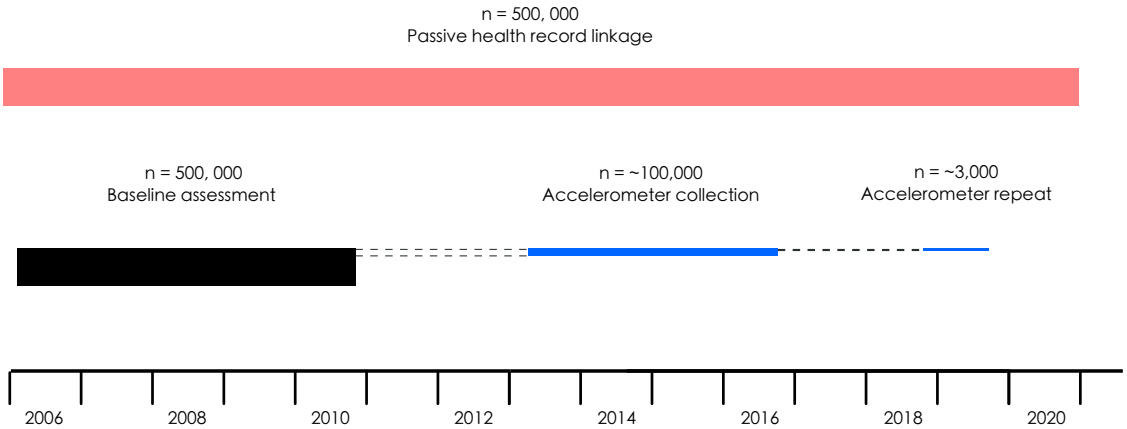


Figure 3.2: The timeline for the UK Biobank accelerometery dataset

The UK Biobank participants were recruited among residents who lived up to 25 miles away from the 22 study assessment centres across the UK [110]. In total,

9.2 million invitations were sent via mail to reach a total participant number of 503,325, resulting in a response rate of 5.5%. Figure 3.2 showcases the timeline for the UK Biobank accelerometer dataset collection. All the participants (n $\approx$ 500,000) underwent baseline assessment between 2006 and 2010 to complete questionnaires, physical measurements and sample collection of blood, urine and saliva [40]

Additional data enrichments were conducted in a subset of the participants; about 100,000 had worn wrist-worn accelerometers [57], 20,000 had obtained repeat baseline measurements, and 75,000 took part in an imaging visit to have different parts of their bodies scanned [111]. At the same time, online questionnaires regarding one's lifestyle, mental health and pain were sent to about 330,000 participants whose email addresses were available [112, 113].

Participant health outcome data

The UK Biobank participants provided their informed consent to allow follow-up of their health status via health record linkage from the UK National Health Service (NHS). The health outcomes are collected from various data sources, including death and disease registry, hospitalisation and primary care [110]. Some outcome definitions require cross-referencing evidence from the low-cost electronic record to more resource-intensive primary data collections to help reduce misclassification and identify disease subtypes.

Ethical framework

The UK Biobank received ethical approval from the National Health Service National Research Service (Ref 21/NW/0157). An independent UK Biobank Ethics and

Governance Council was established to serve as the oversight body to monitor and advise on the UK Biobank Ethics and Governance Framework, which laid out the general principles of project ethics and participant enrolment [114]. During the baseline data collection, the participants gave a broad consent to allow for their involvement in future research studies. No feedback is provided to the participants from the UK Biobank at any time.

3.2.2 Accelerometry in the UK Biobank

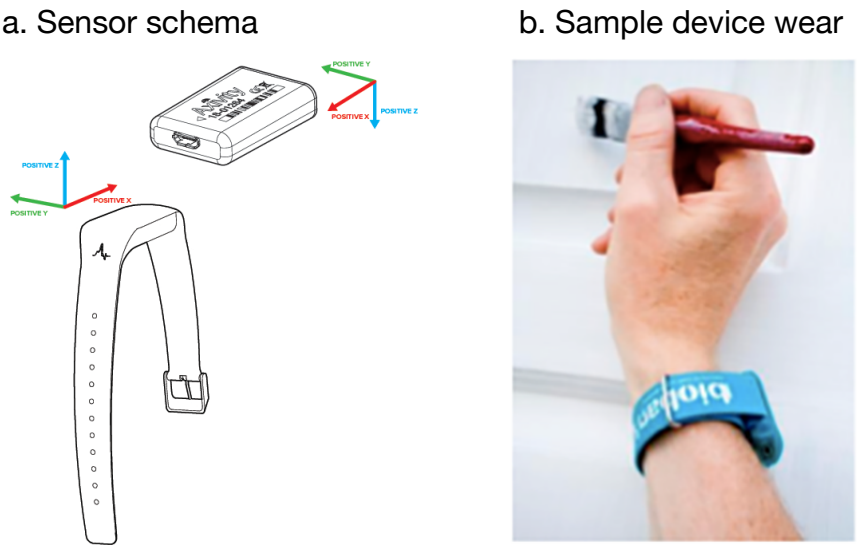


Figure 3.3: Accelerometer used in the UK Biobank: (a) device orientation (b) sample device wear. The images are taken from [57] and [115].

Data collection

Participants who provided an email address were contacted to be part of the accelerometry collection in the UK Biobank between February 2013 and December 2015. The list of participants was determined at random with an exclusion of residents in the North West region, which had already been burdened with new trials at the time [57]. The participants were asked to wear a tri-axial wrist-worn

accelerometer, Axivity AX3 (Newcastle, the UK), for up to seven days on their dominant wrist (Figure 3.3). After consenting, the participants were mailed the accelerometer devices via post and instructed to wear the device as soon as the device arrived. In total, 236,519 participants were invited, 106,053 participants consented to take part, and the data for 103,712 participants were received. A subset of about 3,000 participants had four repeated seven-day accelerometer measurements between 2017 and 2018, which is not used in this thesis.

Data processing

The accelerometers were configured to record the raw accelerometry at 100 Hz with a range of ± 8g. The processing pipeline of the raw data consists of the following steps: (1) Parsing: convert the compressed binary format into a plain text format via a deserialiser; (2) Calibration: harmonise the accelerometry signals collected by different devices to ensure that the recorded signal is similar during stationarity [116]; (3) Filtering: remove machine noise using a low pass filter with a frequency of 20 Hz; (4) Quality control: check for device wear duration and extreme values. All the processing described here can be found in the biobankAccelerometerAnalysis package that I have contributed to on GitHub, `https://github.com/OxWearables/biobankAccelerometerAnalysis`.

3.2.3 Population characteristics

Cohort profile

Table 3.2 describes the baseline characteristics for the overall UK Biobank participants, participants who were invited to the accelerometry collection but declined and

participants who completed the accelerometry collection. No apparent difference exists between the UK Biobank population and those who were invited but declined to participate. When comparing overall UK Biobank population with the accelerometer cohort, participants who completed the accelerometry study also tend to spend more time in education (43.1% vs 32.1% in higher education), consume more alcohol (48.9% vs 43.2% in 3+ times alcohol consumption per week), have a lower Body Mass Index (19.6% vs 24.% in obesity) and enjoy greater self-reported health (21.6% vs 16.3% in excellent self-report health). Similar demographic differences were also observed between the accelerometer cohort and those invited but declined to participate in the accelerometer enrichment.

Table 3.2: Comparison of the population characteristics for all the UK Biobank participants, participants invited to the accelerometry collection but declined and participants who completed the accelerometry collection.

Characteristics	Overall UK Biobank cohort	Invited but declined	Accelerometer cohort
	n=502414	n=132745	n=103672
Age, year			
Mean (SD)	56 (8.10)	55 (8.09)	56 (7.84)
Median [Q1, Q3]	58.0 [50.0, 63.0]	57.0 [49.0, 63.0]	57.0 [50.0, 62.0]
Sex			
Female	273329 (54.4%)	70170 (52.9%)	58275 (56.2%)
Male	229085 (45.6%)	62575 (47.1%)	45397 (43.8%)
Ethnicity			
Non-white	27021 (5.4%)	7610 (5.7%)	3315 (3.2%)
White	472617 (94.1%)	124581 (93.8%)	99990 (96.4%)
Missing	2776 (0.6%)	554 (0.4%)	367 (0.4%)
Smoking status			
Never smoker	273479 (54.4%)	73248 (55.2%)	58958 (56.9%)
Ex-smoker	173025 (34.4%)	46403 (35.0%)	37162 (35.8%)
Current smoker	52962 (10.5%)	12540 (9.4%)	7275 (7.0%)
Missing	891 (0.2%)	137 (0.1%)	47 (0.0%)
Alcohol consumption			
Never drinker	40628 (8.1%)	8922 (6.7%)	5870 (5.7%)
< 3 times per week	243108 (48.4%)	62517 (47.1%)	47011 (45.3%)
3+ times per week	217178 (43.2%)	61073 (46.0%)	50704 (48.9%)
Missing	1500 (0.3%)	233 (0.2%)	87 (0.1%)
Education			
School leaver	168536 (33.5%)	38269 (28.8%)	23714 (22.9%)
Further education	162618 (32.4%)	45220 (34.1%)	34279 (33.1%)
Higher education	161131 (32.1%)	47513 (35.8%)	44635 (43.1%)
Missing	10129 (2.0%)	1743 (1.3%)	1044 (1.0%)
Townsend Deprivation Index			
Mean (SD)	-1 (3.09)	-1 (2.92)	-1 (2.83)
Median [Q1, Q3]	-2.1 [-3.6, 0.5]	-2.4 [-3.8, 0.1]	-2.4 [-3.8, -0.2]
Missing	623 (0.12%)	167 (0.13%)	123 (0.12%)
BMI			
<18.5, underweight	2626 (0.5%)	634 (0.5%)	583 (0.6%)
18.5-24.9, normal	162374 (32.3%)	42285 (31.9%)	39987 (38.6%)
25-29.9, overweight	212079 (42.2%)	57365 (43.2%)	42572 (41.1%)
30+, obese	122231 (24.3%)	31749 (23.9%)	20287 (19.6%)
Missing	3104 (0.62%)	712 (0.54%)	243 (0.23%)
Self-rated health			
Poor	22769 (4.5%)	4965 (3.7%)	2684 (2.6%)
Fair	105345 (21.0%)	26658 (20.1%)	16404 (15.8%)
Good	288970 (57.5%)	78940 (59.5%)	61889 (59.7%)
Excellent	81844 (16.3%)	21571 (16.2%)	22443 (21.6%)
Missing	918 (0.2%)	139 (0.1%)	48 (0.0%)

Selection bias

The UK Biobank was designed to identify preventable causes of diverse diseases that can generalise across populations. Despite the best efforts, differences exist between the study population and the general UK population on several levels. Compared to the general UK population, the UK Biobank participants tend to be healthier, leaner and less likely to smoke. Importantly, the all-cause mortality rate for the UK Biobank participants is almost half the mortality rate compared to the general population, 46.2% lower in men and 55.5% lower in women [117]. In view of the healthy volunteer bias, Fry et al. 2017 have argued that that valid exposure-outcome associations could still be drawn from the UK Biobank.

Nonetheless, larger data volume cannot overcome bias in the data collection efficiently, the Big Data Paradox [118], which has been shown to influence large surveys that already take steps to deal with selection bias [119]. Additional enrichments in the UK Biobank might present a greater challenge regarding selection bias. Table 3.2 from the previous subsection indicates that the participants who took part in the accelerometer study tend to be healthier and receive more years in education compared to the overall UK Biobank cohort. A complete elimination of selection bias is likely to be not possible [120].

Because of the selection bias, pre-trained model developed in the next Chapter will also be biased towards wealthy and healthy populations in the UK, limiting its generalisability to other populations.

3.3 Summary

The Raine Study contains one of the largest sets of concurrent raw accelerometry and polysomnography, an essential requirement of a deep learning-based sleep staging model. The UK Biobank accelerometer dataset and its longitudinal follow-up is a rich resource with phenotypic and genetic information for investigating the association between human sensing data and health. My thesis uses the UK Biobank accelerometer dataset to (1) obtain a pre-trained network to improve sleep staging performance from accelerometry and (2) investigate the association between device-measured sleep and mortality. The next Chapter will outline how the UK Biobank accelerometer dataset can be used to obtain a self-supervised neural network that could improve the classification of human activities and sleep stages.

4

Self-supervised representation learning for accelerometers

Contents

4.1 Introduction 50
4.2 Methods 53
4.2.1 Datasets 54
4.2.2 Multi-task self-supervised-learning 55
4.2.3 Network training 57
4.2.4 Evaluation – human activity recognition 59
4.3 Results 61
4.3.1 Weighted single task training 61
4.3.2 Multi-task self-supervised learning 64
4.3.3 Downstream performance - human activity recognition . 65
4.3.4 Transfer learning using labelled pre-training 68
4.3.5 Ablation studies 68
4.3.6 Understanding the representation 71
4.4 Discussions 73
4.5 Summary 74

The materials for this chapter have been made available online under the following citation: **Hang Yuan***, Shing Chan*, Andrew P Creagh, Catherine Tong, Aidan Acquah, David A Clifton, and Aiden Doherty. "Self-supervised Learning for Human Activity Recognition Using 700,000 Person-days of Wearable Data". In:

npj Digital Medicine (2024)

Shing Chan and I conceptualised the project together, and I conducted the majority of the experiments. All the data shown in the Chapter was produced by myself unless specified otherwise.

4.1 Introduction

The systematic review in Chapter 2 identified the need to move away from hand-crafted features employed by existing sleep detection algorithms. This Chapter addresses how to develop a feature extractor by learning directly from the data using one of the largest collections of wrist-worn accelerometry data, the UK Biobank. This Chapter will first describe how such a feature extractor can be obtained using self-supervised learning to learn human motion dynamics from the unlabelled data in the UK Biobank. Then, the pre-trained model will be evaluated in a wide range of downstream benchmarks to test its generalisability. Because of the lack of sleep classification benchmarks, the downstream evaluation for the pre-trained network will be focused on its utility for human activity recognition first, an area with far more established performance baselines. The pre-trained model will learn features related to human motion dynamics in a task-agnostic way. The obtained model should also generalise sleep stage classification, whose performance will be assessed in the next Chapter.

Human activity recognition from on-body inertial sensors is a fundamental recognition problem in wearable computing. The typical standard in building such

models is the Activity Recognition Chain [121], which describes the pre-processing and segmentation of raw time series into fixed-length frames, from which features are extracted for frame-based classification. Prior works have successfully used tree-based approaches with hand-crafted features [122]. More recently, neural networks have been studied using shallow architectures [123–125]. A reason for this is the small size and diversity of benchmark datasets, which limits the performance of complex deep models with many parameters. Some studies have found that earlier approaches, such as random forests, still outperform deep learning-based models in certain small data regimes [126, 127].

Recently, a growing interest was seen in using self-supervised learning (SSL) for human activity recognition. Some explored temporal variants of the pretext tasks, including forecasting [128], masked reconstruction [129], and contrastive learning [130, 131]. It was shown that multi-task SSL using tasks adapted from image recognition yields better performance with fewer labelled data [65, 132]. However, these findings are only true as far as their dataset size allows. All previous works have only used a small dataset ($<$ 10K samples). Conversely, this Chapter leverages the largest ever collection of wearable data for SSL (6B samples).

Three simple SSL tasks were trained independently and jointly to assess how they could facilitate human activity recognition across a variety of environments using the UK Biobank dataset, which contains terabytes of wearable sensor data collected in the real world. To investigate the potential benefit of SSL in human activity recognition, three tasks were chosen to prioritise the temporal dependencies of human motion, namely, *arrow of time* (AoT), *permutation* and *time-warping*

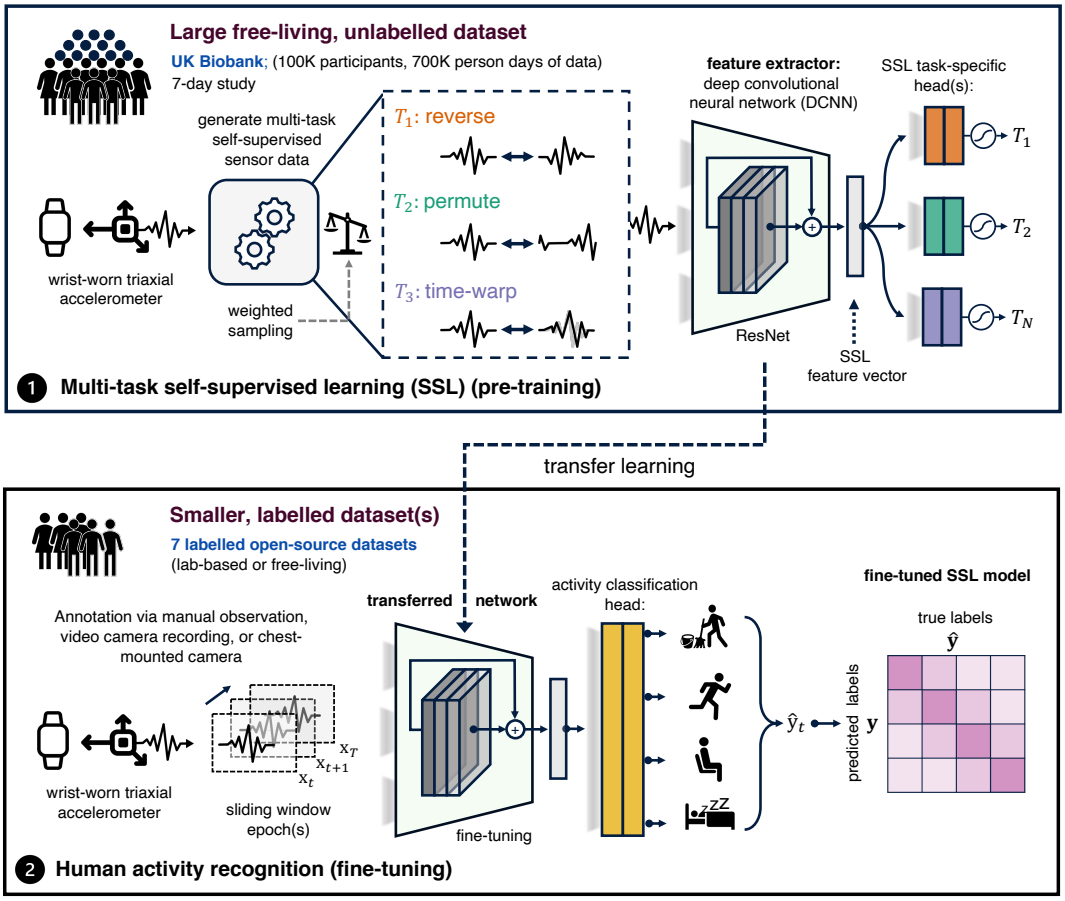


Figure 4.1: Overview of the proposed self-supervised learning pipeline: step 1 involves multi-task self-supervised learning on 700,000 person-days of data from the UK Biobank. Then, utility of the pre-trained network is evaluated in seven benchmark human activity recognition baselines via transfer learning. Reproduced and modified with permissions from [133]. Figure produced by Andrew P. Creagh.

(TW) [134]. Figure 4.1 provides a schematic overview of my study design: first, multi-task SSL is used to pre-train a deep convolutional neural network on 700,000 person-days of free-living accelerometer data from the UK Biobank; second, the pre-trained network is evaluated via transfer learning on seven benchmark human activity recognition datasets to assess representation quality on various activity types and populations. The main contributions of this Chapter are as follows:

- The first demonstration that multi-task SSL on a large free-living accelerometry

can train a human activity recognition model that generalises well across seven external datasets that differ in activity classes, devices, populations and recording environments.

- In contrast to previous works, results in this Chapter provide a more realistic evaluation of the utility of self-supervision for human activity recognition by factoring in common issues seen in the practical use cases of pre-trained models such as domain shift and task shift [135].

- Publicly available models to enable the research community to build high-performing activity recognition models even in a resource-restricted environment.

4.2 Methods

Tri-axial accelerometer data were collected from wrist-worn activity trackers, which record acceleration on three orthogonal axes at a high sampling rate (e.g. 100 Hz). Following [121], signals were into windows of equal duration, effectively treating them as independent inputs to the human activity recognition models. Each window was labelled with an activity class. Throughout this study, all data was linearly resampled to 30 Hz resolution and used ten-second-long windows to compare the downstream benchmarks fairly. The 30 Hz sampling rate was used because most human activities have a frequency less than 10 Hz. A sampling rate higher than the presumed Nyquist rate was used (20 Hz) to ensure that no useful signal was lost.

Table 4.1: Wrist-worn accelerometer datasets used to evaluate the utility of self-supervised learning for human activity recognition tasks.

Dataset	#Subjects	#Samples	#Classes	Environment	References
UK Biobank	∼100K	6 B	Unlabelled	Free-living	[57]
Capture-24	152	910K	4-10	Free-living	[126]
Rowlands	55	36K	13	Lab	[136]
WISDM	46	28K	18	Semi free-living	[137]
REALWORLD	14	12K	8	Lab	[138]
Opportunity	4	3.9K	4	Semi free-living	[139]
PAMAP2	8	2.9K	8	Lab	[140]
ADL	7	0.6K	5	Lab	[141]

Table 4.2: Dataset characteristics

Dataset	Age Mean (± SD)	Sex Female (Male)	Device	Placement	Country
UK Biobank	56.0 ± 7.8	58,275 (45,397)	Axitivity (AX3)	Dominant	United Kingdom
Capture-24	18 - 91	99 (53)	Axitivity (AX3)	Dominant	United Kingdom
Rowlands	40 - 63	37 (23)	Genea	Both	United Kingdom
WISDM	-	-	LG G Watch	Dominant	USA
REALWORLD	18 - 25	LG G Watch R	Dominant	USA	
Opportunity	-	-	Custom inertial measurement units (IMUs)	Dominant	Switzerland
PAMAP2	27.2 ± 3.3	1 (8)	Colibri inertial measurement units (IMUs)	Dominant	Germany
ADL	-	-	Non-specified	Right	Italty

4.2.1 Datasets

The proposed multi-task SSL training relied on the unlabelled UK Biobank dataset, which contains roughly 700,000 person-days of free-living activity data (>100,000 participants, seven days of wear). The free-living aspect is important because the data can contain all sorts of activities, as opposed to lab data that are constrained to scripted activities only.

For the subsequent activity recognition benchmarks, seven external labelled

datasets were considered, which vary in size (600 to 600,000 samples), activity classes (4 to 18 classes), devices used (five different brands), and collection protocol (free-living, scripted, and lab settings). See Table 4.1 and Table 4.2 for additional details. Three datasets had license information, and four datasets reported informed consent information (Appendix Table B.1). However, several open benchmark datasets used did not have suitable licencing or consent information, possibly due to the lack of data governance awareness at the time of their studies. In small datasets with less than 10 individuals, activity classes not present in all the subjects were removed during data cleaning.

4.2.2 Multi-task self-supervised-learning

Three self-supervised tasks from works by Saeed et al. (2019) [65] were considered, which were first used as data augmentation techniques [134] :

- Arrow of time (**AoT**) flips the signal along the time axis, effectively playing the signal in reverse.
- **Permutation** breaks the signal into chunks and shuffles them. The number of chunks was set to four, and the minimum length of each chunk was greater than 10 timestamps.
- Time warping (**TW**) stretches and compresses arbitrary segments of the signal, effectively slowing down and speeding up the signal randomly.

The original multi-task self-supervised learning for human activity recognition included a list of eight transformations: noised, scaled, rotated, negated, horizontally

flipped, permuted, time-warped, and channel-shuffled [65]. However, the success of multi-task learning requires a good synergy between the tasks included, especially if the training is done using free-living data, which contains more heterogeneity than lab-based data. The reasons why other transformations were not included were:

- Noised: A noised signal tends to produce poor quality features leading to performance worse than random chance level [65];
- Scaled: Different devices likely have different absolute values, and thus scaling the value of the signals is less relevant for human motion dynamics;
- Rotated: We used rotation in data augmentation to produce a direction-invariant network;
- Negated: Negation is a form of rotation, making it redundant;
- Channel-shuffling: We used channel-shuffling in data augmentation to produce a direction-invariant network.

Following [65], each of the included tasks was treated as a binary problem predicting whether a transformation has been applied. In the multi-task learning (MTL) setting, not all tasks might benefit human activity recognition when trained jointly. The influence of different task combinations on the downstream performance was also assessed (Section 4.3.2). Given the computational constraints, all the tasks had a 0.5 ratio of whether the transformation was applied.

The cross-entropy loss was computed for each task, and each task was weighted equally in the loss calculation. The weight of each task could potentially explored to further improve the representation quality.

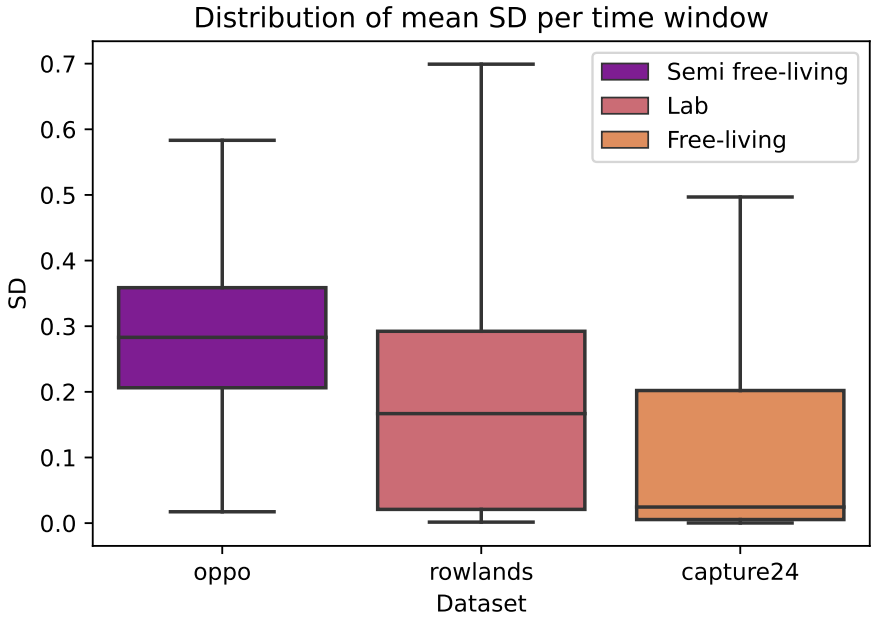


Figure 4.2: Data collected from free-living environment tends to have more stationary periods than the data collected in lab environment. SD: standard deviation; oppo: opportunity dataset.

Weighted sampling Motion data collected in the real world contains large portions of low movement periods that are less informative (Figure 4.2), an issue for SSL tasks as static signals remain virtually unchanged after the transformations. Experiments showed that it was crucial to perform weighted sampling for improved training stability and convergence (Subsection 4.3.1). Weighted sampling involved a random selection of data windows in proportion to their standard deviation to give more weight to high-movement periods.

4.2.3 Network training

Shallow networks with only a few layers tend to be used in existing works for SSL [67, 132]. However, the model with a limited capacity will saturate quickly if the data volume increases. Therefore, a ResNet-V2 with 18 layers and 1D convolutions [142] was adapted for the main trunk (feature extractor), totalling 10M parameters.

The learned feature vector was of size 1024. All the tasks shared the same feature extractor. Then, a softmax layer was attached for each of the self-supervised tasks. In the downstream evaluation, a fully-connected (FC) layer of size 512 was added between the feature extractor and softmax readout. The network structure was fixed for all the downstream evaluations.

For SSL, four subjects were loaded from the UK Biobank at each iteration. For each subject, one day out of the week-long data was sampled, from which 1,500 10-second windows were further sampled to make up a training batch. Self-supervised transformations were then applied to the batch of data. Since the axis orientation differs between device manufacturers, random axis swaps and rotations were used to augment the training data to embed this invariance into the models. For optimisation, Adam [143] with a learning rate of 1e-3 was used. Linear scaling for the learning rate with five epochs as burn-in was used to account for large batch sizes, $1500 \times 4 = 6000$ [144]. Thus, a total of $100,000 \times 6,000 = 60$ million time windows were fed to the network in each training iteration when the training was done on 100,000 participants. The network training was distributed over four Tesla V100-SXM2 with 32GB of memory. It took about 420 GPU hours to train the MTL model (about 20 epochs). An 8:2 ratio was used for the train/test split for all the self-supervised experiments. For fine-tuning, the same training setup as the pre-training was used where possible except for the batch size, but was re-adjusted depending on the size of each dataset.

4.2.4 Evaluation – human activity recognition

Held-one-subject-out cross-validation was used to evaluate the downstream activity recognition performance for the datasets that had < 10 subjects. Additionally, activity classes not done by all the subjects were removed in these small datasets. For datasets with ≥ 10 subjects, five-fold subject-wise cross-validation was used instead. Each cross-validation had a 7:1:2 split ratio for train/validation/test sets. Early-stopping with a patience of five was applied to avoid over-fitting.

After the network was trained on the UK Biobank using $\sim$100,000 participants, the network was fine-tuned on the seven labelled downstream datasets to perform human activity detection using two approaches: (1) fine-tuning all the layers (2) freezing the trunk (feature extractor) and fine-tuning only the FC layers in the end. Additional baselines included a network of the same architecture but fully trained from scratch and a random forest model with tried-and-tested time series features, which has often been neglected in baseline model comparisons [126, 145–147]. The list of hand-crafted features extracted for random forest were:

- Mean, standard deviation, and range for each axis;
- Correlation between each axis pair;
- Euclidean norm, its mean, standard deviation, range, median absolute deviation, kurtosis, and skew;
- The top two dominant frequencies in the power spectrum.

In addition, a **shared implementation** was introduced for the network training, model evaluation and preprocessing. Differences in experiment setup, such as training rates, regularisation and data augmentation, can lead to inconsistent results [148]. A unified evaluation framework would ensure a fair comparison between different baseline models. The proposed evaluation framework contrasts with previous work, where there is no fixed evaluation protocol across the benchmark datasets, making it hard to compare model performance with the current state-of-the-art. The results produced in this Chapter can serve as a good baseline for future human activity recognition research.

Transfer learning

Transfer learning involves pre-training on a larger labelled dataset and fine-tuning on a smaller dataset. However, transfer learning has not been commonly reported as a baseline for human activity recognition. The success of transfer learning, however, depends on how similar the source and target datasets are. Hence, additional experiments used the two largest labelled datasets, Capture-24 [126] and Rowlands [136], for pre-training, which were then fine-tuned on other smaller labelled datasets.

The benefits of data volume

Deep learning methods are data intensive. The learning efficiency depends on the needed unlabelled data for pre-training. The representation quality depends on the needed labelled data for fine-tuning. Therefore, the ablation studies investigated

how the downstream performance differs on two axes: the amount of labelled data and the amount of unlabelled data.

Concretely, the fine-tuning step gradually included more labelled subjects in both Capture-24 and Rowlands in the downstream evaluation to assess the performance of pre-trained models in a limited-data regime. Regarding unlabelled data, pre-training used data from 100 to 100,0000 participants with one order of magnitude increment. The amount of unlabelled data per subject was also investigated using values from 0.25 to 1 for 10,000 participants. A data ratio of 0.25 means that if one day of data per subject was used previously, only six hours of data per subject would now be used for training. Investigating how unlabelled data influences downstream performance will guide how much data one needs to obtain an effective SSL model for human activity recognition.

4.3 Results

4.3.1 Weighted single task training

When training individual pretext tasks, Figure 4.3 showed that all the tasks had worse convergence behaviour without weighted sampling. The performance degradation was most pronounced for the AoT and permutation tasks. The test performance for the AoT stayed at the random chance level, and the test performance for permutation dropped roughly 10 percentage points without weighted sampling. After the pre-trained models were fine-tuned on the downstream human activity recognition datasets, permutation had the best performance in six out of seven

baselines except for REALWORD, for which the performance for time-warping was better (Table 4.3).

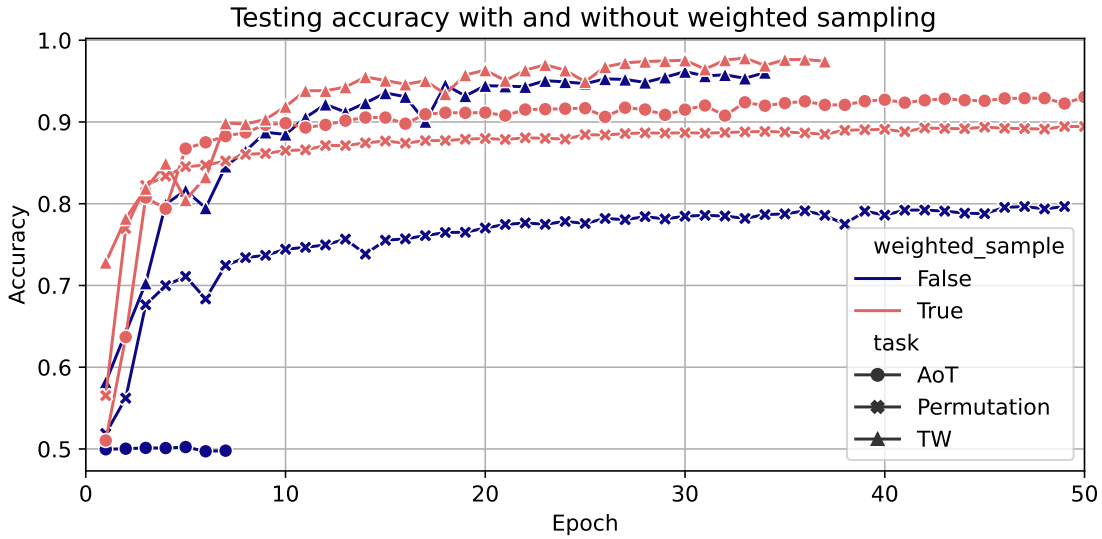


Figure 4.3: Accuracy test curves for training four self-supervised tasks individually using 1,000 randomly sampled participants from the UK Biobank with and without weighted sampling. The performance shown was on the training data. The patience for early-stopping was five. AoT: arrow of time; TW: time-warping.

Table 4.3: Downstream human activity recognition performance (subject-wise F1 (±SD)) when using different self-supervised learning tasks after fine-tuning: Arrow of the Time (AoT), Permutation (P), Time warping (TW). The tasks were trained using the same 1,000 UK-Biobank participants.

Task	Capture-24	Rowlands	WISDM	REALWORLD	Opportunity	PAMAP2	ADL
#Subjects	152	55	46	14	4	8	7
#Samples	573K	36K	28k	12k	3882	2869	635
AoT	.671 ± .094	.565 ± .120	.767 ± .124	.750 ± .084	.582 ± .054	.715 ± .036	.754 ± .157
Permutation	**.721 ± .093**	**.783 ± .099**	**.778 ± .109**	.766 ± .063	**.588 ± .076**	**.750 ± .057**	**.810 ± .097**
TW	.715 ± .093	.776 ± .110	.767 ± .120	**.772 ± .073**	.584 ± .064	.737 ± .079	.765 ± .117

Table 4.4: Downstream human activity recognition performance (subject-wise F1 (±SD)) for different self-supervised task combinations using 1,000 randomly sampled UK Biobank participants. N is the number of samples.

AoT	Permutation	TW	Capture-24 n=573K	Rowlands n=36k	Opportunity n=3.9K
Single task					
✓	✗	✗	.671 ± .094	.565 ± .120	.582 ± .054
✗	✓	✗	.721 ± .093	.783 ± .099	**.588 ± .076**
✗	✗	✓	.715 ± .093	.776 ± .110	.584 ± .064
Multi-task					
✗	✓	✓	.714 ± .094	.755 ± .103	.587 ± .070
✓	✗	✓	**.719 ± .094**	.762 ± .102	.530 ± .071
✓	✓	✗	.718 ± .092	**.781 ± .101**	.502 ± .081
✓	✓	✓	.718 ± .095	.770 ± .102	.482 ± .078

4.3.2 Multi-task self-supervised learning

Three downstream datasets, one large (Capture-24), medium (Rowlands) and small (Opportunity) dataset were chosen to investigate how different SSL configurations performed. Different tasks were trained both individually and jointly using 1000 participants from the UK Biobank. Then, the model was fine-tuned on the subsequent human activity recognition benchmarks (Table 4.4).

The differences between different SSL combinations on large datasets (Capture-24 and Rowlands) was smaller than that of the smaller dataset (Opportunity). There was no clear best performing configuration, and thus, for ease of comparison, we chose to use all tasks in pre-training for the remaining experiments. In addition, training more tasks together might yield the most general representation for different downstream datasets.

4.3.3 Downstream performance - human activity recognition

Table 4.5 summarises the subject-wise F1 and Kappa scores for seven human activity recognition datasets. The random forest models outperformed the deep learning models trained from scratch for all except the Capture-24 dataset, which is the largest dataset in the evaluations (Table 4.1). The performance gap between random forest and training from scratch was the largest in smaller datasets. Meanwhile, pre-trained models outperformed the models trained from scratch and random forests in all seven datasets. Fine-tuning all layers was better than fine-tuning just the fully connected (FC) layers after the ConV layers.

The most significant improvement using pre-training was seen on the small datasets. Conversely, the SSL benefit was more modest for larger datasets. In Capture-24, the F1 improvement was 2.5% when comparing the model with and without SSL pre-training. Nonetheless, with SSL pre-training, the median relative F1 improvement was 18.4% when compared against the same network trained from scratch and 8.5% when compared against the random forest model.

Table 4.5: Subject-wise F1 and Kappa (κ) for downstream human activity recognition tasks (mean ± SD) using 100,000 UK Biobank participants for pre-training. The relative improvement compares the performance between the model that is trained from scratch and fine-tuning using all the layers.

Data		Random forest	ResNet			Improvement
			Trained from scratch	Fine-tune self-supervised		
				After ConV layers	All layers	
Capture-24	F1	.694 ± .099	.708 ± .094	.723 ± .097	**.726 ± .093**	2.5 %
	κ	.683 ± .101	.703 ± .092	.718 ± .090	**.737 ± .087**	4.8%
Rowlands	F1	.700 ± .090	.696 ± .106	.724 ± .081	**.796 ± .093**	14.4%
	κ	.830 ± .086	.810 ± .098	.850 ± .062	**.874 ± .073**	7.9%
WISDM	F1	.711 ± .149	.684 ± .123	.759 ± .121	**.810 ± .127**	18.4%
	κ	.715 ± .153	.685 ± .124	.758 ± .121	**.809 ± .126**	18.1%
REALWORLD	F1	.731 ± .119	.705 ± .062	.764 ± .052	**.792 ± .075**	12.3%
	κ	.680 ± .142	.638 ± .079	.703 ± .063	**.739 ± .086**	15.8%
Opportunity	F1	.416 ± .185	.383 ± .124	.570 ± .078	**.595 ± .085**	55.4%
	κ	.318 ± .206	.238 ± .154	.435 ± .092	**.471 ± .104**	97.9%
PAMAP2	F1	.753 ± .093	.605 ± .086	.725 ± .054	**.789 ± .054**	30.4%
	κ	.744 ± .101	.596 ± .086	.717 ± .057	**.769 ± .059**	29.0 %
ADL	F1	.764 ± .180	.414 ± .179	.645 ± .107	**.829 ± .101**	100.0 %
	κ	.720 ± .199	.368 ± .198	.654 ± .123	**.849 ± .113**	130.7%

Table 4.6: Transfer learning (subject-wise F1) performance comparison between supervised pre-training with self-supervised pre-training.

Target data	Source data				
	Rowlands		Capture-24		UK Biobank
	Supervised	Self-supervised	Supervised	Self-supervised	Self-supervised
Capture-24	.707 ± .094	.709 ± .094	-	.707 ± .094	**.726 ± .093**
Rowlands	-	.734 ± .082	.728 ± .094	.730 ± .084	**.796 ± .093**
WISDM	.680 ± .109	.702 ± .123	.715 ± .119	.723 ± .121	**.810 ± .127**
REALWORLD	.712 ± .086	.737 ± .105	.759 ± .070	.771 ± .061	**.792 ± .075**
Opportunity	.536 ± .019	.539 ± .018	.547 ± .043	.547 ± .042	**.595 ± .085**
PAMAP2	.677 ± .082	.689 ± .078	.678 ± .118	.725 ± .725	**.789 ± .054**
ADL	.634 ± .182	.701 ± .111	.768 ± .169	.754 ± .159	**.829 ± .101**

4.3.4 Transfer learning using labelled pre-training

Supervised pre-training can boost the learning outcome substantially than training from scratch (Table 4.5 vs Table 4.6). However, self-supervised pre-training could outperform supervised pre-training when using Rowlands and Capture-24 as the source data, implying that we might not need to rely on labelled large-scale data for pre-training after all. Table 4.6 also underpins the importance of data volume in self-supervision as the SSL representation from the UK Biobank consistently outperformed the SSL baselines trained using Rowlands and Capture-24.

4.3.5 Ablation studies

Varying labelled data in the downstream

The pre-trained models did well regardless of the number of labelled subjects in two downstream datasets (Figure 4.5). However, fully-supervised and random forest models were more susceptible to the number of labelled subjects. The performance gain for having more labelled subjects was roughly linear with respect to the number of subjects included, with a greater increase when fewer labelled subjects were available.

Varying unlabelled data in the pre-training

The downstream human activity recognition performance appeared to increase linearly with respect to the number of unlabelled subjects on a log scale (Figure 4.4 left). This conforms with the previous finding in a semi-supervised setting [148]. The SSL performance boost with more SSL unlabelled subjects was most significant in the smallest dataset, Opportunity. Furthermore, if the number of participants

is fixed at 10,000 in pre-training, the data ratio included per subject did not significantly influence the downstream performance. For example, the downstream performance did not degrade too much even when the subject data ratio decreased from 1.0 to .25 (Figure 4.4 right).

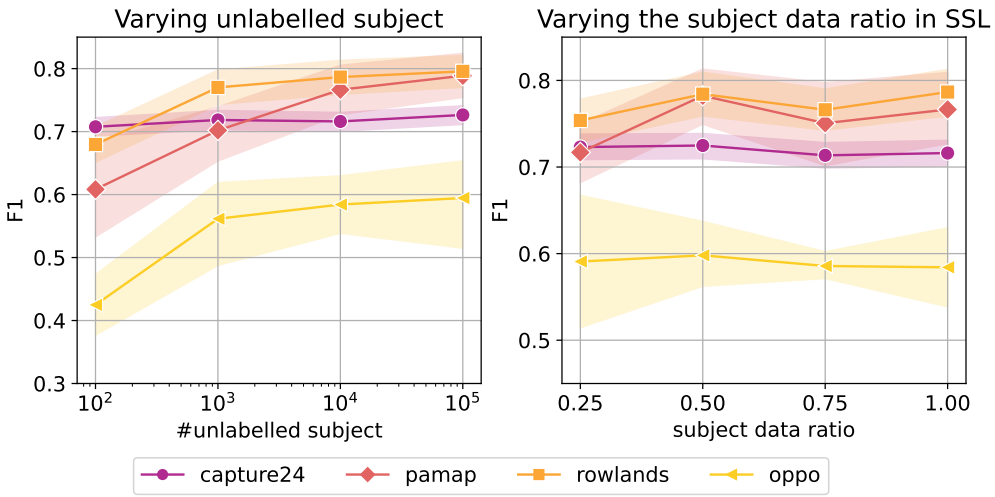


Figure 4.4: Left: More #unlabelled subjects in the self-supervision leads to better downstream performance. **Right**: Unlabelled data per subject does not improve downstream performance when using 10,000 subjects for pre-training. Oppo: opportunity dataset.

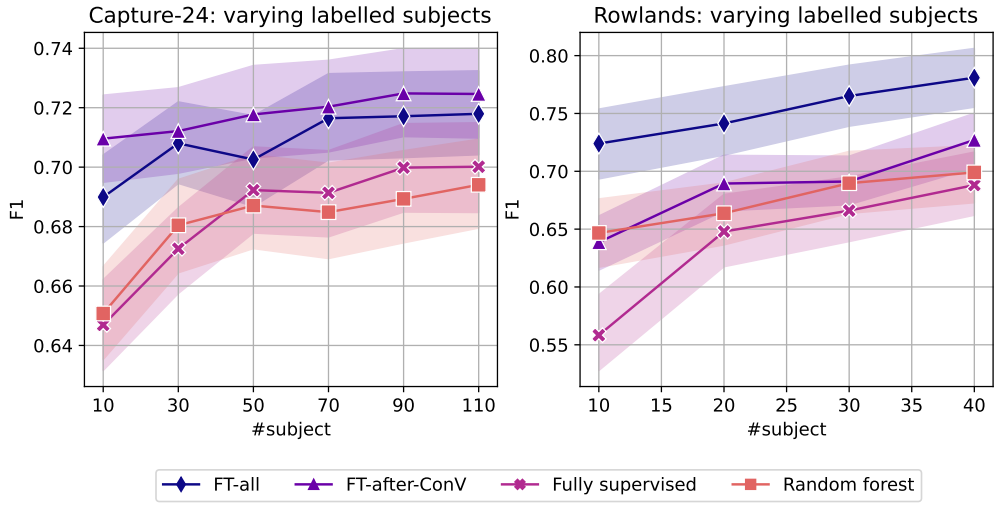
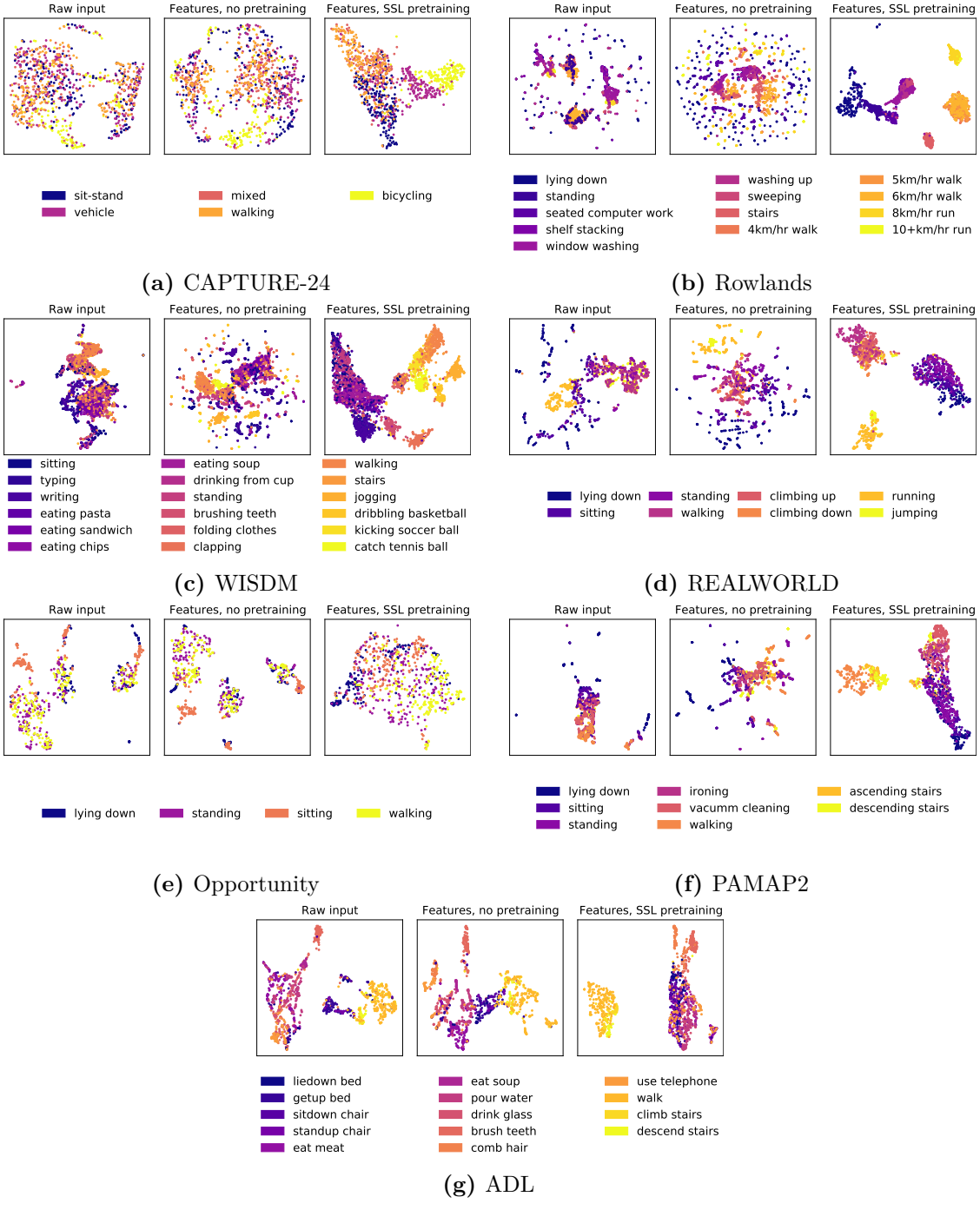


Figure 4.5: How different models perform in the downstream tasks when we change the number of labelled subjects: Capture-24 (**left**) and Rowlands (**right**)

4.3.6 Understanding the representation

Cluster analysis Uniform Manifold Approximation and Projection (UMAP) [149] was applied with default parameters to project the raw data input, features with no pre-training and features with SSL pre-training into a lower-dimensional space for visualisation. UMAP was chosen over other cluster analysis techniques because of its computational efficiency and better preservation of both the local and global structure in the data. The UAMP results are shown in Figure 4.6. Features with SSL pre-training tended to group activities by intensity and frequency, which was not apparent from the raw input or features without pre-training. Furthermoroe, samples of the same class exhibited more compactness with the pre-trained features.



(a) CAPTURE-24 (b) Rowlands

(c) WISDM (d) REALWORLD

(e) Opportunity (f) PAMAP2

(g) ADL

Figure 4.6: Cluster analysis on raw inputs, untrained features and self-supervised pretrained features: The colour gradients denote activity intensities. Results suggest that self-supervised features were better at clustering similar activities (e.g. walking, stair climbing vs. sitting, writing, typing) as well as their intensities (e.g. lying down, sitting, standing vs. jogging, sports). Figure produced by Shing Chan.

4.4 Discussions

Results from this Chapter have shown that self-supervised pre-training consistently improved downstream human activity recognition (HAR), especially in small datasets, reducing the need for labelled data. Self-supervised representations generalise well across a range of external datasets, tasks, devices and populations. In contrast, previous studies applied SSL training and fine-tuning only on the same data sources [65, 130, 150], making it necessary to pre-train on every new dataset in practical applications. My pre-trained network can serve as a foundational human activity recognition model that removes the need to pre-train on every unseen dataset.

This Chapter represents the most robust human activity recognition foundational model to date as it is trained on a much larger and more diverse dataset than previous efforts in this space. The scale of UK Biobank is several orders of magnitude greater than the largest datasets used by the current state-of-the-art such as the Fenland data [151] ($n = 160,000$). In addition, not just big in size, but UK Biobank is much more diverse as it contains hundreds, if not thousands, of natural human activities – a crucial aspect regarding the input distribution and generalisation. The representation quality from self-supervision was found to be superior than that of supervised learning in an apple-to-apple comparison when using Rowlands as the source of pre-training. SSL using other modalities has also found that self-supervised pre-training can outperform supervised pre-training [152, 153]. Pre-training on the UK Biobank yields the most significant improvement among all the SSL representations, as the HAR performance is uniformly better than other

pre-trained baselines (Table 4.6). The performance boost could be attributed to the large data volume, diverse activity classes and rich population characteristics of the UK Biobank over other alternative pre-training datasets.

Several directions of research could be further explored in the future. For example, more work is needed to quantify the trade-off between the number of participants and the amount of data per participant in SSL for human activity recognition. When more pre-training data became available, the representation quality increased to only a certain point (Figure 4.4). Perhaps, inter-person variability plays a larger role in the representation quality than intra-person variability. In addition, it is unknown how much performance might be boosted with the addition of data from more diverse demographic groups and sensing modalities. Due to a current lack of raw accelerometer datasets in different regions of the world, a limitation of this work is that the pre-training data (UK Biobank dataset) consists mostly of Caucasians from the UK. A multi-modal representation that also includes electrocardiogram and other time-series wearable sensor data sources will also be important to consider when such datasets are collected in future.

4.5 Summary

In summary, this Chapter developed and evaluated a self-supervised deep neural network on large-scale activity tracking data. The self-supervised model improved on prior state-of-the-art performance across seven benchmark human activity recognition datasets. The open-sourced model represents a foundation model that others can build upon for state-of-the-art human activity recognition applications.

Even though the evaluation has only been limited to how the pre-trained network can be used in the domain of activity recognition, multi-task self-supervision was designed to learn generic features related to human motion dynamics. Chapter 5 will investigate how the self-supervised pre-trained network could benefit accelerometer-based sleep classification.

5

Deep learning for sleep stage detection using accelerometers

Contents

5.1 Introduction 78
5.2 Methods 79
5.2.1 Study design and participants 79
5.2.2 Accelerometer devices and data preprocessing 82
5.2.3 Ascertainment of sleep stages via polysomnography . . . 83
5.2.4 Sleep stage classification using wrist-worn accelerometers 84
5.3 Results 90
5.3.1 Baseline SleepNet performance 90
5.3.2 Subject-wise performance external evaluation 92
5.3.3 Epoch by epoch performance evaluation 96
5.3.4 The performance of sleep parameters estimation 101
5.4 Discussion 103
5.4.1 Comparison to existing sleep staging methods 103
5.4.2 Sleep parameters agreement between SleepNet and polysomnography 104
5.4.3 Strengths and limitations 105
5.5 Summary 107

The materials for this chapter have been made available online under the following citation: **Hang Yuan**, Tatiana Plekhanova, Rosemary Walmsley, Amy C. Reynolds,

Kathleen J. Maddison, Maja Bucan, Philip Gehrman, Alex Rowlands, David W. Ray, Derrick Bennett, Joanne McVeigh, Leon Straker, Peter Eastwood, Simon D. Kyle, and Aiden Doherty. "Self-supervised learning of accelerometer data provides new insights for sleep and its association with mortality". In: *npj Digital Medicine* (2024)

5.1 Introduction

In Chapter 4, a self-supervised deep neural network was developed to learn features related to human motion dynamics by using 700,000 person-days of unlabelled accelerometer data from the UK Biobank. The obtained self-supervised network showed generalisability in enhancing human activity recognition models across devices, activities, and populations. This Chapter builds upon the obtained self-supervised pre-trained network as a feature extractor to develop a sleep stage classification model using concurrent laboratory-based polysomnography as the ground truth.

Despite the popularity of sleep monitoring in consumer and research-grade wrist-worn devices, sleep assessment algorithms are frequently proprietary and validated in small populations, making their measurement validity unclear [70, 88–90]. Methods for sleep classification (i.e. defining periods of wake, NREM and REM sleep) primarily rely on hand-crafted spatiotemporal features such as device angle, which may not make full use of all the information in wearable sensing signals. Hence, data-driven methods like deep learning could be advantageous. The Systematic Review in Chapter 2 on accelerometer-based sleep classification

also identified the lack of sufficient training data as a key limiting factor for deep learning approaches that are data-intensive. This could potentially be mitigated by using my self-supervised pre-trained network as a feature extractor.

This Chapter, therefore, sets out to (1) develop and benchmark a deep recurrent neural network for sleep stage classification using wrist-worn accelerometers and (2) validate the model performance on unseen datasets to assess its generalisability.

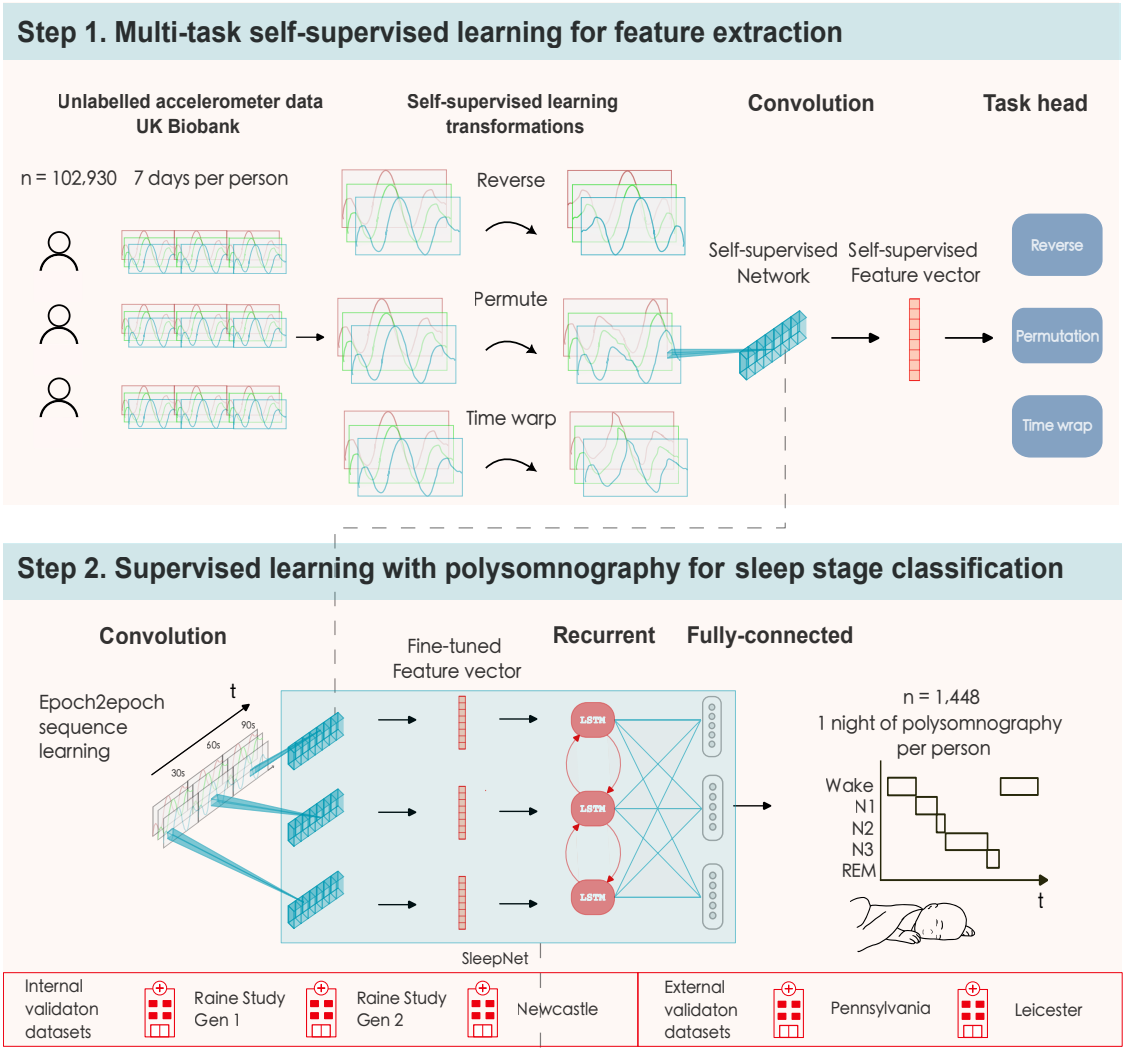


Figure 5.1: The SleepNet development pipeline: 1. The multi-task self-supervised learning was used to obtain a feature extractor by learning from 700,000 person-days of tri-axial accelerometry data in the UK Biobank. 2. The pre-trained feature extractor was then fine-tuned with a deep recurrent network to train a sleep-stage classifier using polysomnography as the ground truth.

5.2 Methods

5.2.1 Study design and participants

The model development and validation were done using data from multiple study centres. The model was designed to classify each 30-second window of accelerometry data into one of the three sleep stages: wake, rapid-eye-movement sleep (REM), and non-rapid-eye movement sleep (NREM). The sleep grouping was done based on conceptual simplicity, as REM and NREM sleep are two distinct groups that are easy to understand in population health studies. Alternative sleep groupings based on movement profile can be considered to improve the classification performance, such as wake, NREM I, REM and deep sleep (NREM II+NREM III) [154]. Figure 5.1 illustrates the two main steps of the analysis: (1) feature extraction from unlabelled free-living data and (2) sleep staging model development.

The self-supervised model pre-training was done on the unlabelled UK Biobank dataset. For sleep staging model development, internal validation consisted of two generations of participants from the Raine Study [101, 102] and patients who were assessed for sleep disorders as part of their routine clinical assessment in the Newcastle cohort [105]. The external validation consisted of two general populations from Leicester [106] and Pennsylvania [107]. The rest of this subsection describes each dataset and inclusion criteria. Table 5.1 shows the dataset characteristics for all the included datasets.

Raine Study As mentioned in Chapter 3, the Raine Study has followed up roughly 2900 children since 1989 in Australia. A subset of children (Raine Gen2, 50%

Table 5.1: Characteristics of the datasets used for internal validation, external validation and health association analyses "Patient" indicates whether a cohort consists of sleep patients in a clinic.

Name	n	Age	Placement	Device	Patient	Publication
UK Biobank	103,561	62.3 ± 7.9	Dom wrist	Axivity	✗	[57]
Raine Gen1	865	56.7 ± 5.6	Dom wrist	GT3X	✗	[101]
Raine Gen2	795	22.1 ± 0.6	Dom wrist	GT3X	✗	[101]
Newcastle	28	44.9 ± 14.9	Both wrists	GENEActiv	✓	[105]
Leicester	30	30.8 ± 6.7	Both wrists	Axivity	✗	[106]
Pennsylvania	22	22.8 ± 4.5	Non-dom wrist	Axivity	✗	[107]

females) at the age of 22 and their parents (Raine Gen1, 57% females) were invited to undergo one night of laboratory-based polysomnography at Western Australia's Center for Sleep Science [101, 102]. Every participant was instructed to wear an ActiGraph GT3X device on the dominant wrist. Earlier GT3X firmware would enter an idle mode to save the battery when no sufficient movement was detected, so only participants with no missing data were included for the Raine Gen2 cohort.

Newcastle The Newcastle dataset recruited 28 adult patients (39% females) for a one-night laboratory-based polysomnography assessment in Newcastle upon Tyne, UK, as part of their routine clinical visit [105]. During the polysomnography recording, the participants wore two GENEActive devices, one on each wrist. The sampling frequency for the wristbands was set to 85.7 Hz.

Leicester Thirty healthy volunteers (63% females and 73% white) wore three devices: GENEActive, Axivity AX3, and ActiGraph GT9X on each wrist during one night of laboratory-based polysomnography assessment [106]. The relative position of the devices was randomly allocated for each participant. The devices were set to record at 100 Hz. During the lab visit, when the participants wished to go to bed, the

recording was started. The sleep episodes usually ended between 6 a.m. and 7 a.m. the following morning. The recording sessions were pre-processed such that every recording would start with "light off" and end with "light off" to ensure comparability.

Pennsylvania The Pennsylvania dataset consists of 22 healthy sleepers who had one night of laboratory-based polysomnography assessment at the University of Pennsylvania Center for sleep [107]. The participants were asked to wear an Axivity device on the non-dominant wrist during the polysomnography session.

UK Biobank The UK Biobank accelerometer dataset has been described in Chapter 3, which will not be repeated here.

5.2.2 Accelerometer devices and data preprocessing

Three different devices were used to collect the accelerometry for the included datasets: ActiGraph GT3X, Axivity AX3 and GENEActive Original accelerometers. The devices used have been shown to have a high inter-instrument agreement ($> 80\%$) in derived sedentary and sleep-related time estimates in free-living environments [155]. As for device placement, data from the dominant wrist was selected where possible to be consistent with the UK Biobank protocol.

The Biobank Accelerometer Analysis Tool [61, 126] was used to preprocess all the data. The raw tri-axial accelerometry was first resampled into 30 Hz and clipped to $\pm$ 3 g. The accelerometry sequence was then divided into consecutive 30-second windows. Stationary periods had(x/y/z sd $<$ 13 mg) with a duration greater than 60 minutes as non-wear [57]. Additional data was excluded if it could

not be parsed, had unrealistic high values (> 200 mg), was poorly calibrated, or did not pass visual inspection.

5.2.3 Ascertainment of sleep stages via polysomnography

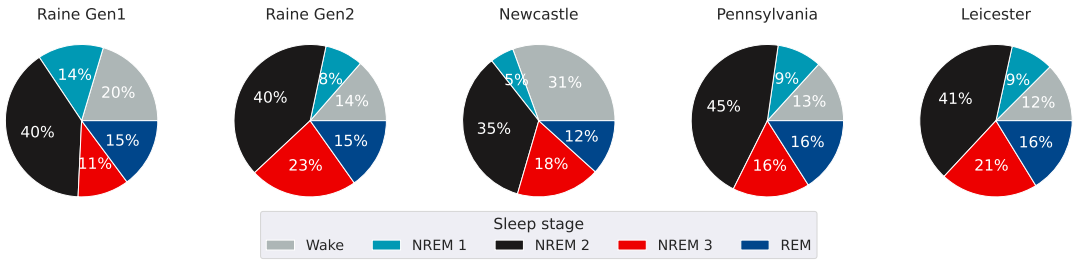


Figure 5.2: Sleep stage distribution for all the datasets used.

The gold-standard, laboratory-based polysomnography sleep label was aligned with its concurrent accelerometer data as the model ground truth. During laboratory-based polysomnography, the participants were instructed to wear an array of electrophysiological sensors to measure brain activity, eye movement and muscle tone. A trained polysomnographic technician would score every 30-second window of data into one of one of the five sleep categories: wake, REM, and NREM I, II, and III.

The scoring in this study was conducted according to the American Academy of Sleep Medicine (AASM) protocol [103], in which scoring criteria for each sleep stage were outlined concerning the frequency, shape, and activity patterns of recorded signals [156]. The inter-rater agreement between polysomnography scorers is around 0.7-0.8 in Kappa [157]. In total, 1,157,913 (∼10,000 hours) sleep windows were used to train the network. The sleep stage distributions were similar across all the datasets except for the Newcastle cohort, which had a greater proportion of wakefulness than the others (Figure 5.2).

5.2.4 Sleep stage classification using wrist-worn accelerometers

Model setup The sleep stage classifier (SleepNet) was built on top of the self-supervised pre-trained ResNet, described in Chapter 4. The ResNet was appended with two layers of Bi-directional Long-Short-Term-Memory (LSTM) layers of 1024 units to learn the temporal dependencies of the model [158]. Accelerometer-based sleep staging is particularly challenging only using a 30-second window of data. As the aim of SleepNet is not to be used for real-time applications, it is beneficial to consider a longer temporal horizon of data to infer specific sleep stages, making LSTM an excellent candidate. In the end, two fully-connected layers of 512 units were used to generate the sleep stages. The model was trained to discriminate five sleep stages directly (wake, N1, N2, N3 and REM). The inter-rater agreement between N1, N2, and N3 is only moderate, and thus, a three-class output was also obtained by merging all the NREM stages. Finally, NREM I, II, III and NREM were merged into the sleep class to obtain the two-class output.

SleepNet training The learning rate was set to be 1e-3. The gradient clipping was set to 1 to avoid the exploding gradient for LSTM. The weighted Cross-Entropy was used as the objective function, and each class was weighted with the inverse of its frequency to account for the imbalanced dataset. Rotation and axis swap were also used to augment the input data to obtain a direction-invariant model. Each training mini-batch consisted of five participants. For each individual, four 1.5-hour sequences were selected with random starting points to avoid overfitting to

the study protocol, where the beginning and the end of the sequence are always the "wake" class. The model was trained on a Tesla V100-SXM2 with 32GB of memory. It took about 12 hours for the model to converge. The model performance was reported using five-fold subject-wise cross-validation. The data was first split into train/test with a ratio of 4:1. The train set was further split into train/validation with a ratio of 4:1. Early stopping with a patience of ten steps was used to avoid overfitting on the validation set in each cross-validation fold.

For internal validation, subject-wise five-fold cross-validation was used on the Raine Gen2, Raine Gen1, and Newcastle cohorts. For external validation, the SleepNet was trained on all the internal datasets and then evaluated its performance on the Leicester and Pennsylvania cohorts. The SleepNet performance was compared with the current state-of-the-art, a random forest model that used the hand-crafted spatiotemporal features [61, 91]. The random forest feature definitions are listed in Table 5.2.

Table 5.2: Hand-crafted features

Handcrafted features	Notes
Sleep features [91]	
	All sleep features have 12 derived variables:
ENMO	mean, std, min, max, entropy 20 bins (low resolution),
Angle Z	entropy 200 bins (high resolution), median absolute derivation,
Locomotor inactivity during sleep	and mean difference between neighbouring windows.
Axis features [61]	
Mean	1 per axis
Standard deviation	1 per axis
Range	1 per axis
Inter-quantile-range	1 per axis
Correlation of variations	1 per axis
Features on the vector norm [61]	norm $= \sqrt{x^2 + y^2 + z^2}$
Mean	
Standard deviation	
Inter-quantile-range	
Median absolute derivation	
Kurtosis	
Skew	
Truncated ENMO	
Absolute value of ENMO	
Entropy	
Dominant Frequency	
Total power	
Dominant frequencies	3 features: 0.3-5 Hz, 0.3-15 Hz, and 0.6-2.5 Hz
Dominant frequency power	3 features: 0.3-5 Hz, 0.3-15 Hz, and 0.6-2.5 Hz
Second dominant frequency	1 feature: 0.3-15 Hz
Fourier transform coefficients	11 features: 1 Hz-11 Hz
Fourier coefficients	12 features: 1st-12th coefficient

The staging performance was reported in both subject-wise and epoch-to-epoch fashion. Three-class and five-class confusion matrices were plotted for both internal and external validation. Since Cohen Kappa, F1 scores, and balanced accuracies (Table 5.3) are less influenced by class imbalance, they were used to evaluate the overall model. To assess the relationship between the model performance and population characteristics, subject-wise sleep staging performance was stratified by age, sex, employment status, income level, body mass index (BMI), presence and severity of sleep apnea using the apnea-hypopnea index (AHI), existing sleep disorders, and neurological disorders where available.

Finally, the agreement between derived summary sleep parameters and polysomnography was assessed via Bland-Altman plots for the following sleep parameters: total sleep duration, sleep efficiency (proportion of total time asleep when in bed), wake-time after sleep onset (WASO), REM duration, NREM duration, REM ratio, NREM ratio. Table 5.4 details the sleep parameter definitions and their calculations.

Table 5.3: Model performance metric definitions (TP: true positive; TN: true negative; FP: false positive; FN: false negative)

Metric	Definition
Precision	$\frac{\text{TP}}{\text{TP+FP}}$
Sensitivity/Recall	$\frac{\text{TP}}{\text{TP+FN}}$
Specificity	$\frac{\text{TN}}{\text{TN+FP}}$
Accuracy	$\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$
F1	$2 \times \frac{\text{precision} \cdot \text{recall}}{\text{precision+recall}}$
Kappa	$1 - \frac{1-p_o}{1-p_e}$ p_o: relative observed agreement p_e: expected agreement probability
Balanced accuracy	$\frac{1}{n} \sum_i \text{Accuracy}_{class_i}$

Table 5.4: Sleep parameter definitions: total sleep duration (TSD), rapid-eye-movement (REM), non-rapid-eye-movement (NREM), sleep onset latency (SOL), wake after sleep onset (WASO), and sleep efficiency (SE).

Parameter	Definition
Total sleep duration (TSD)	The total time spent in sleep during the recording period per day.
Overnight sleep duration	The longest sleep window duration (max one hour of sleep discontinuity allowed) over a noon-to-noon interval.
Time in bed	The amount of time spent in bed: A person might not be asleep during this period. The time in bed was estimated using a random forest model that was trained using data from sleep diaries.
Sleep onset latency (SOL)	The time difference between when one gets in bed and the sleep onset. The sleep onset (SOL) is defined as the first occurrence of three consecutive 30-sec sleep windows.
Wake after sleep onset (WASO)	The amount of wake time spent after the sleep onset during the longest sleep window.
Sleep efficiency (SE)	SE for sleep window after device-detected sleep onset: $\frac{\text{Overnight sleep duration}}{\text{time in bed}}$
REM duration	The total time spent in the REM stage.
REM ratio	$\frac{\text{REM duration}}{\text{TSD}}$
NREM duration	The total time spent in the NREM I, II, and III stages.
NREM ratio	$\frac{\text{NREM duration}}{\text{TSD}}$

5.3 Results

5.3.1 Baseline SleepNet performance

After preprocessing, 1,395 participants were included in the internal validation, and 53 participants were included in the external validation. The proposed deep recurrent neural network (SleepNet) pre-trained with self-supervision achieved the best performance when compared with other baseline models that used hand-crafted features (Table 5.5). Even though for sleep-wake classification, the performance between the random forest with hand-crafted features and the SleepNet without pre-training was comparable, SleepNet with pre-training outperformed the random forest model (F1: 0.75 vs 0.73). The performance difference for the three-stage classification was larger between the SleepNet with pre-training (F1: .57) and the random forest model (F1: .46). Finally, the self-supervised SleepNet achieved an area under the receiver operating characteristics (ROC) curve of 0.88 for sleep-wake classification and an area under the ROC curve of 0.81 for the wake-REM-NREM classification (Figure 5.3).

Table 5.5: Subject-wise sleep stage classification for benchmark models using internal validation datasets with the Raine Study and the Newcastle cohort: The random forest model was trained using hand-crafted features. SleepNet is the deep recurrent network without pre-training. SleepNet-SSL is the network pre-trained using self-supervision. Five-fold subject-wise performance metrics (mean ± SD) are reported using the internal validation data. REM: rapid-eye-movement sleep, NREM: non-rapid-eye-movement sleep, Kappa score: κ.

Model	Sleep versus Wake			Wake versus REM versus NREM			Wake versus REM versus NREM I, II, III		
	κ	Accuracy	F1	κ	Accuracy	F1			
Random forest [61, 91]	0.472±0.192	0.756±0.102	0.729±0.104	0.290±0.149	0.507±0.072	0.464±0.072	0.202±0.103	0.336±0.056	0.287±0.060
SleepNet	0.468±0.193	0.757±0.103	0.727±0.105	0.313±0.162	0.576±0.112	0.535±0.106	0.247±0.105	0.441±0.090	0.356±0.084
SleepNet-SSL	**0.511±0.196**	**0.775±0.105**	**0.750±0.107**	**0.375±0.163**	**0.625±0.116**	**0.573±0.116**	**0.297±0.115**	**0.474±0.098**	**0.407±0.094**

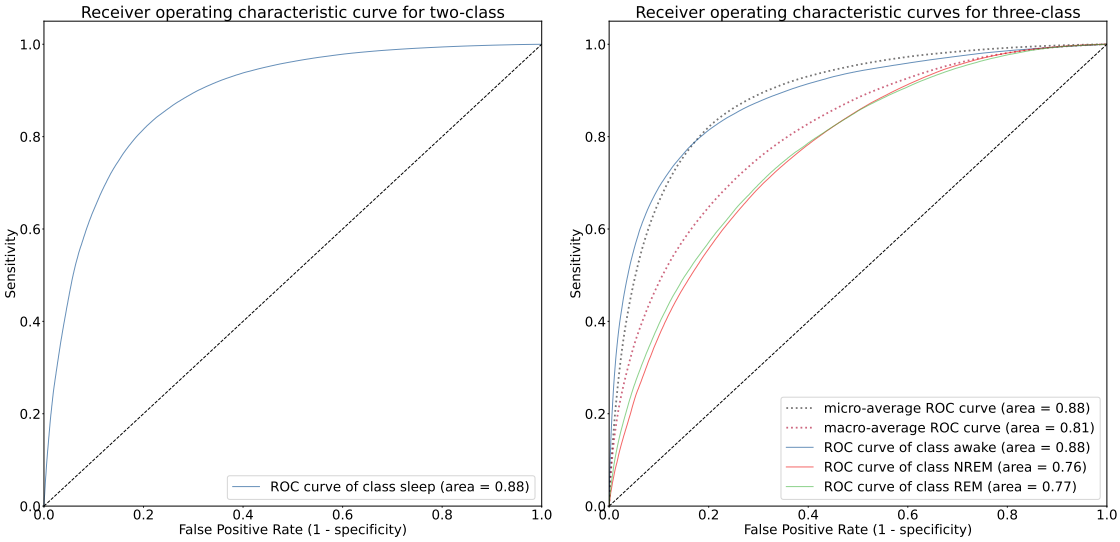


Figure 5.3: Receiver operating characteristics curves for two-class (wake/sleep) and three-class (wake/REM/NREM) settings on the internal validation dataset using the best performing model self-supervised SleepNet. REM: rapid-eye-movement sleep, NREM: non-rapid-eye-movement sleep.

5.3.2 Subject-wise performance external evaluation

The cohort-specific subject-wise performance for both the internal and external validation using the pre-trained SleepNet is shown in Table 5.6. The agreement decreased slightly on the external validation with an F1 decrease from 0.75 ± 0.1 to 0.67 ± 0.11 in the two-class setting (sleep/wake) and an F1 decrease from 0.57 ± 0.11 to 0.52 ± 0.10 in the three-class setting (wake/REM/NREM). In the Newcastle cohort, for the sleep/wake classification, sensitivity decreased and specificity increased in participants with sleep disorders. No obvious difference was observed in both Raine Gen1 (age: 56.7 ± 5.6) and Gen2 (age: 22.1 ± 0.6) cohorts when the participants were stratified by sex, BMI, AHI, and sleep disorder conditions.(Table 5.7-5.9).

Table 5.6: Subject-wise performance sleep classification validation using the best-performing model: All the performance is reported within period in bed. Cohort-specific and pooled performance (Kappa (κ), balanced accuracy, and F1) are shown for both internal and external validation. The pooled performance is calculated by combining all the participants from different datasets. REM: rapid-eye-movement sleep; NREM: non-rapid-eye-movement sleep.

Dataset	Sleep versus Wake			Wake versus REM versus NREM		
	κ	Accuracy	F1	κ	Accuracy	F1
Internal validation						
Raine Gen1	0.561±0.161	0.791±0.091	0.775±0.089	0.389±0.152	0.623±0.108	0.586±0.105
Raine Gen2	0.437±0.189	0.758±0.101	0.712±0.100	0.344±0.161	0.603±0.115	0.552±0.108
Newcastle	0.394±0.189	0.715±0.091	0.686±0.103	0.285±0.151	0.513±0.078	0.467±0.085
Pooled internal	0.509±0.184	0.777±0.097	0.748±0.099	0.369±0.158	0.613±0.112	0.571±0.108
External Validation						
Leicester	0.278±0.141	0.678±0.072	0.633±0.075	0.253±0.122	0.527±0.086	0.488±0.082
Pennsylvania	0.468±0.225	0.807±0.117	0.725±0.118	0.374±0.172	0.626±0.092	0.565±0.097
Pooled external	0.360±0.205	0.734±0.114	0.673±0.106	0.306±0.157	0.570±0.101	0.521±0.097

Table 5.7: Model characteristics on the internal validation datasets (wake versus sleep): subject-wise performance metrics (mean ± SD) are reported using the internal validation data. Sen: sensitivity, Spe: specificity. Wake is the negative class and the sleep is the positive class when calculating model performance.

Wake versus Sleep									
Subgroups	Raine Gen1			Raine Gen2			Newcastle		
	n	Sen (%)	Spe (%)	n	Sen (%)	Spe (%)	n	Sen (%)	Spe (%)
Sex									
Male	357	92.0 ± 9.1	63.7 ± 20.4	264	87.6 ± 10.0	62.7 ± 21.8	15	79.6 ± 21.9	59.1 ± 24.9
Femal	459	91.5 ± 9.2	68.6 ± 21.3	273	88.8 ± 9.3	63.3 ± 22.3	7	82.9 ± 14.9	69.3 ± 14.0
Body Mass Index (BMI)									
< 25	232	92.9 ± 8.5	63.7 ± 22.1	338	88.5 ± 9.3	62.8 ± 21.6	-	-	-
25 - 29.9	318	92.1 ± 8.8	65.8 ± 21.0	120	89.7 ± 8.7	62.8 ± 22.7	-	-	-
>30	265	90.1 ± 9.9	69.7 ± 19.7	79	84.9 ± 11.8	64.0 ± 23.1	-	-	-
Apnea Hypopnea Index (AHI)									
< 5	199	93.7 ± 7.2	67.0 ± 20.7	338	88.0 ± 10.0	65.2 ± 21.6	-	-	-
5 - 14.9	349	91.5 ± 8.4	67.6 ± 21.2	146	88.9 ± 9.4	58.4 ± 23.0	-	-	-
15 - 29.9	150	90.9 ± 9.3	66.5 ± 20.8	39	88.5 ± 8.1	61.7 ± 20.5	-	-	-
≥ 30	114	89.9 ± 12.2	62.5 ± 20.4	14	84.9 ± 8.8	62.4 ± 21.5	-	-	-
Has sleep disorder(s)?									
Yes	155	90.6 ± 10.1	64.5 ± 22.3	106	87.6 ± 10.0	65.2 ± 22.2	15	75.3 ± 21.5	66.0 ± 23.4
No	661	91.9 ± 8.9	66.9 ± 20.7	431	88.4 ± 9.6	62.5 ± 22.0	7	92.3 ± 6.0	54.4 ± 18.3

Table 5.8: Model characteristics on the internal validation datasets (wake versus REM versus NREM): subject-wise performance metrics (mean ± SD) are reported using the internal validation data. REM: rapid-eye-movement, NREM: non-rapid-eye-movement, Kappa score: κ.

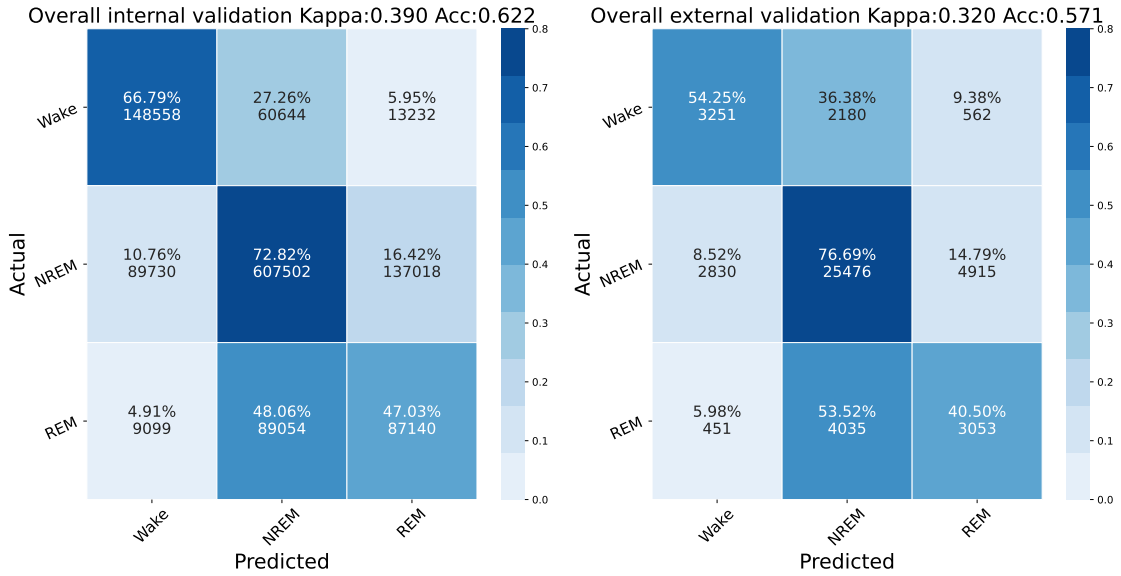
	Wake versus REM versus NREM					
Subgroups	Raine Gen1		Raine Gen2		Newcastle	
	n	κ	n	κ	n	κ
Sex						
Male	357	0.293 ± 0.100	264	0.286 ± 0.120	15	0.200 ± 0.137
Female	459	0.313 ± 0.114	273	0.284 ± 0.117	7	0.258 ± 0.084
Body Mass Index (BMI)						
< 25	232	0.375 ± 0.162	338	0.342 ± 0.148	-	-
25 - 29.9	318	0.390 ± 0.152	120	0.335 ± 0.170	-	-
>30	265	0.401 ± 0.143	79	0.329 ± 0.178	-	-
Apnea Hypopnea Index (AHI)						
< 5	199	0.397 ± 0.163	338	0.349 ± 0.156	-	-
5 - 14.9	349	0.390 ± 0.148	146	0.317 ± 0.158	-	-
15 - 29.9	150	0.395 ± 0.153	39	0.355 ± 0.166	-	-
≥ 30	114	0.369 ± 0.143	14	0.273 ± 0.139	-	-
Has sleep disorder(s)?						
Yes	155	0.386 ± 0.149	106	0.354 ± 0.162	15	0.277 ± 0.148
No	661	0.390 ± 0.153	431	0.335 ± 0.157	7	0.303 ± 0.179

Table 5.9: Model characteristics on the internal validation datasets (wake versus REM versus NREM I, II, III): subject-wise performance metrics (mean ± SD) are reported using the internal validation data. REM: rapid-eye-movement, NREM: non-rapid-eye-movement, Kappa score: κ.

	Wake versus REM versus NREM I, II, III					
Subgroups	Raine Gen1		Raine Gen2		Newcastle	
	n	κ	n	κ	n	κ
Sex						
Male	357	0.279 ± 0.103	264	0.287 ± 0.120	16	0.014 ± 0.102
Female	459	0.307 ± 0.111	273	0.285 ± 0.113	9	0.125 ± 0.106
Body Mass Index (BMI)						
< 25	232	0.295 ± 0.117	338	0.286 ± 0.110	-	-
25 - 29.9	318	0.309 ± 0.107	120	0.292 ± 0.127	-	-
>30	265	0.307 ± 0.102	79	0.273 ± 0.140	-	-
Apnea Hypopnea Index (AHI)						
< 5	199	0.307 ± 0.114	338	0.293 ± 0.116	-	-
5 - 14.9	349	0.309 ± 0.108	146	0.264 ± 0.118	-	-
15 - 29.9	150	0.309 ± 0.104	39	0.299 ± 0.127	-	-
≥ 30	114	0.283 ± 0.104	14	0.274 ± 0.131	-	-
Has sleep disorder(s)?						
Yes	155	0.286 ± 0.107	106	0.297 ± 0.131	15	0.213 ± 0.136
No	661	0.309 ± 0.108	431	0.283 ± 0.115	7	0.230 ± 0.099

5.3.3 Epoch by epoch performance evaluation

By classifying each given window in an epoch-by-epoch fashion, the SleepNet achieved a Kappa score of 0.39 on the internal validation set and a Kappa score of 0.32 on the external validation set in the three-class setting (Figure 5.4). Cohort-specific confusion matrices can be found in Figures 5.5-5.8. Figure 5.9 visualizes a one-night sample actigram, its ground-truth polysomnography labels, and SleepNet predictions.



(a) Overall internal validation **(b)** Overall external validation

Figure 5.4: Three class classification (wake/REM/NREM) confusion matrix: epoch-to-epoch Kappa and balanced accuracies are shown. The number of predictions and proportion ratios are shown for each pair of ground-truth and prediction class. REM: rapid-eye-movement sleep; NREM: non-rapid-eye-movement sleep.

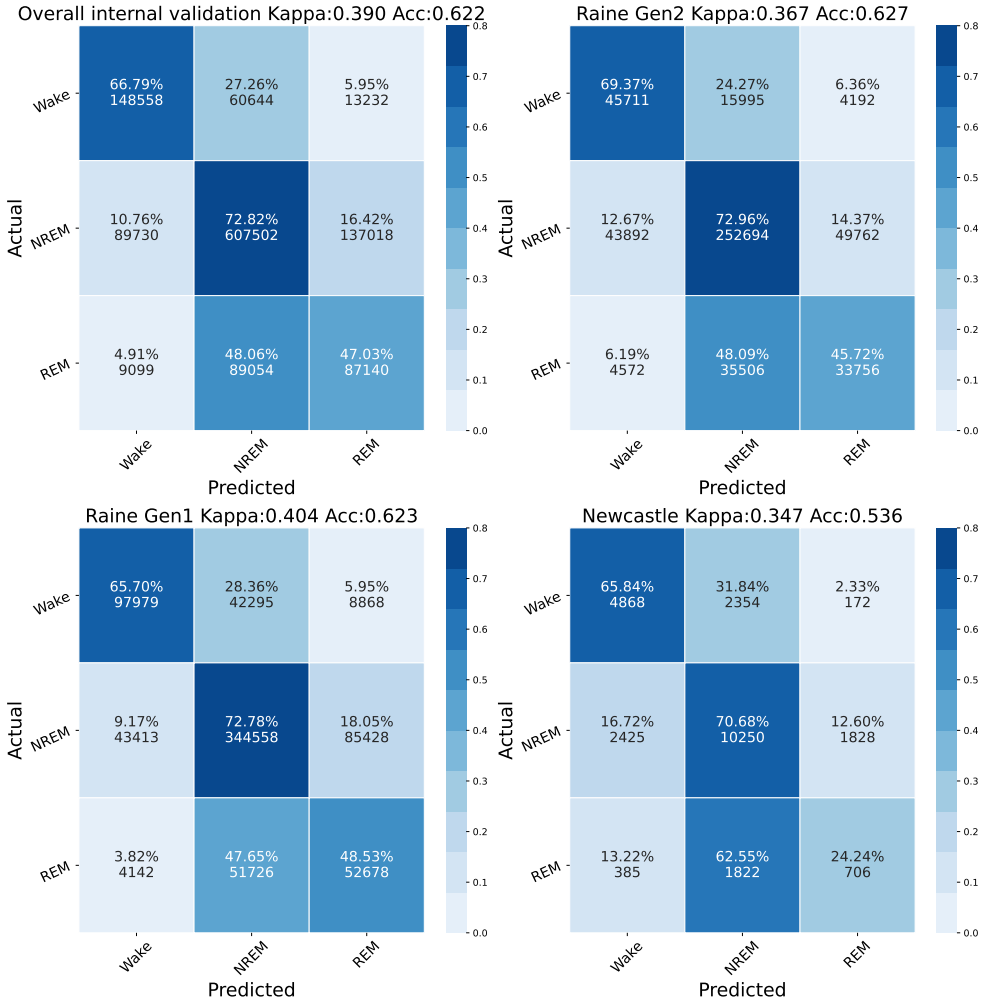


Figure 5.5: Three-class sleep staging (wake/REM/NREM) for internal validation: epoch-to-epoch Kappa and balanced accuracies are shown. The number of predictions and proportion ratios are shown for each pair of ground-truth and prediction class. REM: rapid-eye-movement sleep; NREM: non-rapid-eye-movement sleep.

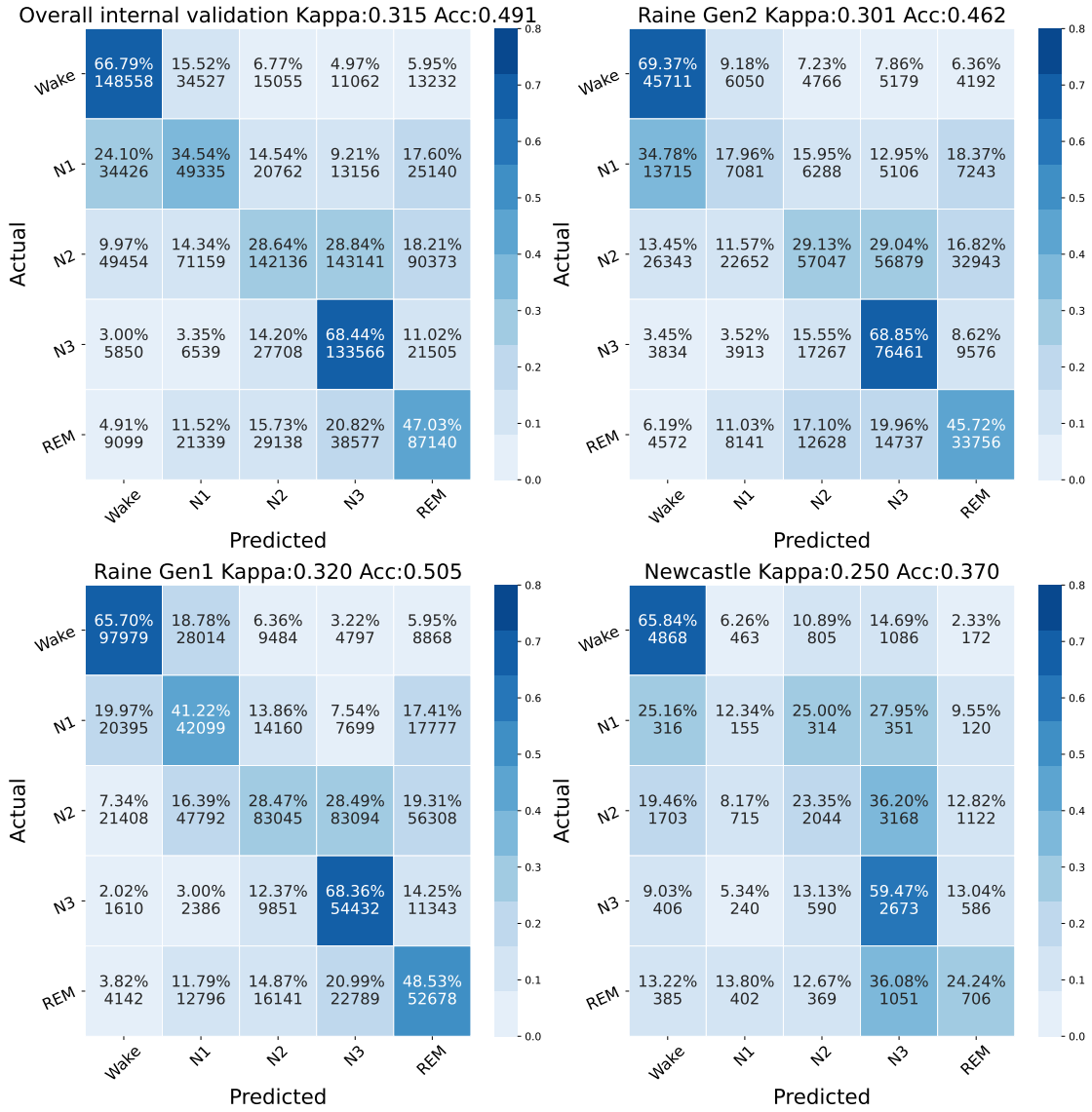


Figure 5.6: Five-class sleep staging (wake/REM/N1/N2/N3) for internal validation: epoch-to-epoch kappa and balanced accuracies are shown. The number of predictions and proportion ratios are shown for each pair of ground-truth and prediction class. REM: rapid-eye-movement sleep, N1, N2, N3: non-rapid-eye-movement sleep 1, 2, 3.

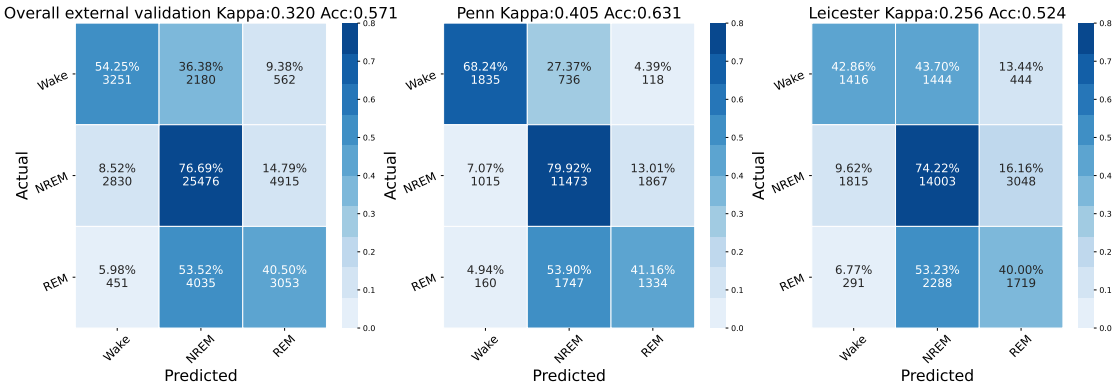


Figure 5.7: Three-class sleep staging (wake/REM/NREM) for external validation: epoch-to-epoch kappa and balanced accuracies are shown. The number of predictions and proportion ratios are shown for each pair of ground-truth and prediction class. REM: rapid-eye-movement sleep; NREM: non-rapid-eye-movement sleep.

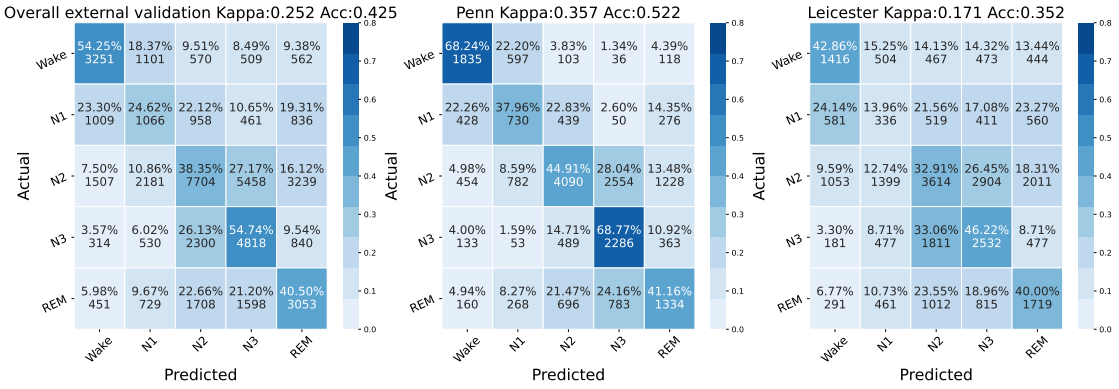


Figure 5.8: Five-class sleep staging (wake/REM/N1/N2/N3) for external validation: epoch-to-epoch kappa and balanced accuracies are shown. The number of predictions and proportion ratios are shown for each pair of ground-truth and prediction class. REM: rapid-eye-movement sleep, N1, N2, N3: non-rapid-eye-movement sleep 1, 2, 3.

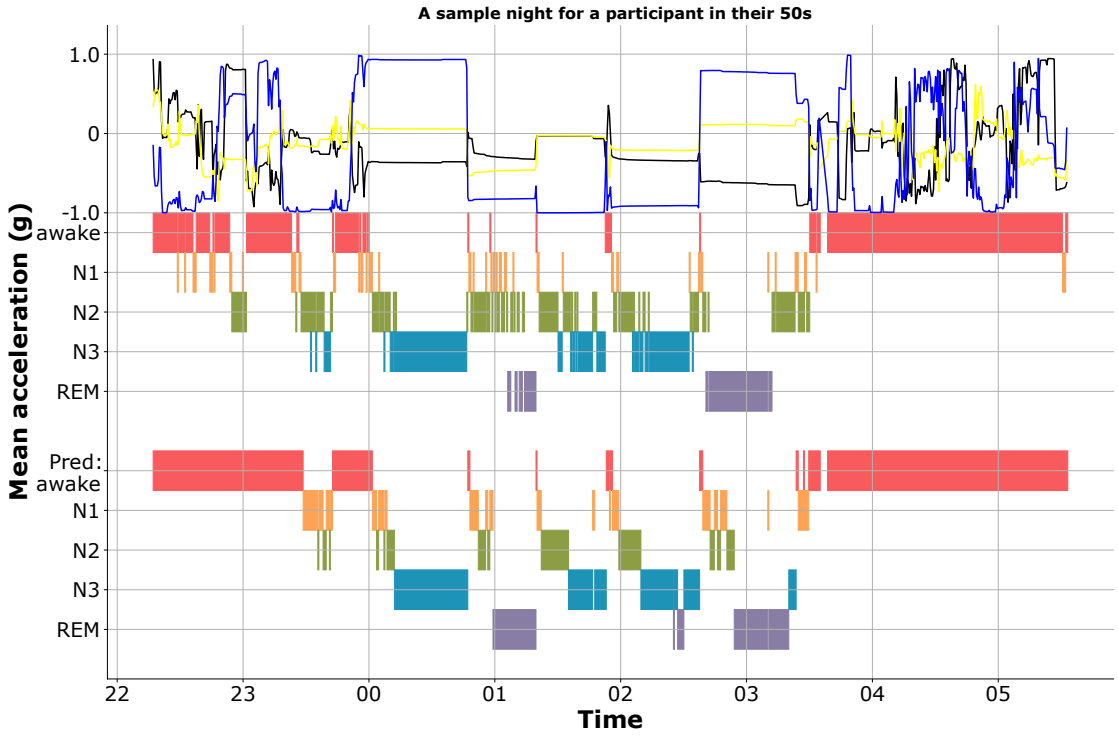


Figure 5.9: A sample actigram, hypnogram ground truth and prediction for a participant whose sleep stages are well captured: the **top** hypnogram is the ground-truth and the **bottom** hypnogram is the prediction generated by SleepNet based on the actigram. REM: rapid-eye-movement sleep, N1, N2, N3: non-rapid-eye-movement sleep 1, 2, 3.

5.3.4 The performance of sleep parameters estimation

On the internal validation, SleepNet had a mean bias of 8.9 minutes (95% limits of agreement (LoA): -89.0 to 106.9 minutes) for total sleep duration, -18.7 minutes (95% LoA: -130.9 to 93.6 minutes) for REM duration, and 27.6 minutes (95% LoA: -100.6 to 155.8 minutes) for NREM duration (Figure 5.10). In comparison, on the external validation, the mean bias was 34.7 minutes (95% LoA: -37.8 to 107.2 minutes) for total sleep duration, -2.6 minutes (95% LoA: -68.4 to 73.6 minutes) for REM duration, and 32.1 minutes (95% LoA: -54.4 to 118.5 minutes) for NREM duration. Overall, the model tended to underestimate REM and short sleep and overestimate NREM and long sleep. Appendix Figures C.1 to C.6 depict the agreement assessments for REM ratio, NREM ratio, waken after sleep onset (WASO) and sleep efficiency on the individual cohorts. Based on the results from external validation, accelerometer-based sleep parameters tended to underestimate total sleep duration, NREM duration, NREM ratio, and sleep efficiency and overestimate REM duration, REM ratio, and WASO.

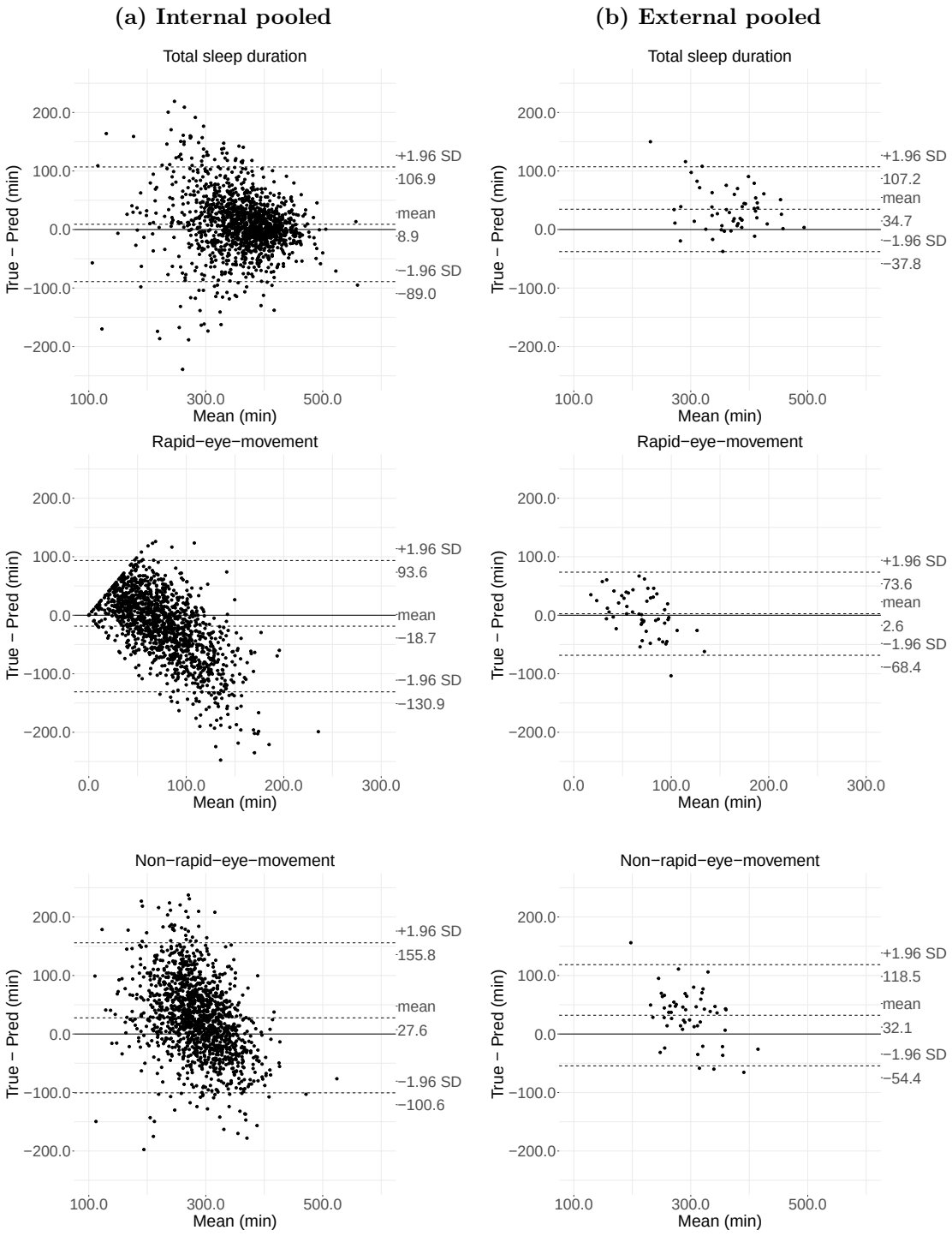


Figure 5.10: Agreement assessment via Bland-Atman plot for total sleep duration, rapid eye movement sleep (REM) duration, and non-rapid eye movement sleep (NREM) duration on internal and external validation. The internal validation consists of 1,373 polysomnography nights from the Raine Study and the Newcastle cohort, whereas the external validation consists of 53 polysomnography nights from the Leicester and Pennsylvania cohorts.

5.4 Discussion

This Chapter has developed, and internally and externally validated a deep-learning method to characterise sleep stages from a wrist-worn accelerometer with competitive performance against 1,448 nights of laboratory-based polysomnography recordings. The open-source algorithm will facilitate future studies on sleep and sleep stages using large-scale accelerometer databases.

5.4.1 Comparison to existing sleep staging methods

The presented self-supervised deep learning sleep staging method outperformed existing baseline methods that rely on hand-crafted features. The inferred sleep architecture estimates had a fair agreement ($\kappa = 0.39$) with the polysomnography ground truth on the internal validation [159]. Unlike previous work in sleep classification methods that depended on hand-crafted features [71, 91], the proposed method automatically extracted the features using self-supervision, hence removing the need for manual engineering. Even for sleep-wake classification, SleepNet achieved comparable results to a systematic evaluation of eight state-of-the-art sleep algorithms [70] in the Newcastle dataset. However, this work offers a more robust evaluation and identifies the upper limit of using accelerometry for sleep classification by developing a model with one of the largest multicentre datasets with polysomnography ground truth, at least ten times the size of existing studies.

5.4.2 Sleep parameters agreement between SleepNet and polysomnography

The derived sleep parameters from SleepNet exhibit varying levels of agreement with polysomnography. SleepNet can be used for population-scale inference to understand the broad association between device-measured sleep and important health outcomes. At the same time SleepNet can be used in smaller clinical studies to investigate the differences between groups of varying clinical characteristics. Inevitability, the levels of precision required for population inference and clinical studies are different.

A task force of experts commissioned by the American Academy of Sleep Medicine performed a meta-analysis on using accelerometers to assess sleep disorders and circadian rhythm sleep-wake disorders [76], and concluded that the accelerometers often provide consistent sleep parameters that are unique from the self-reported sleep for some parameters in sleep patients. This meta-analysis also outlined clinical significance thresholds for the maximum allowable 95% CI of the mean difference between device-measured sleep and polysomnography with a breakdown of 40 minutes for total sleep duration, 30 minutes for WASO, and 5% for sleep efficiency. These clinical thresholds are based on the judgement of the expert task force. The proposed SleepNet met the pre-determined clinical thresholds for the listed three sleep parameters: on the external validation, the mean bias was 32.1 minutes (95% LoA: -37.8 to 107.2 minutes) for total sleep duration, -17.3 minutes (95% LoA: -77.9 to 43.2 minutes) for WASO, and 7.8% (95% LoA: -8.9% to 24.5%) for sleep efficiency.

Despite the overlap between the algorithmic precision and clinical guidelines, accelerometer-based sleep classification could still be improved to reduce variations

across individuals. Future guidelines might factor into classification variance and potentially provide gradings for the agreement (low/mid/high) to differentiate higher quality and lower quality measurement tools within the acceptable clinical threshold.

Our error analysis further reveals the sources of observed errors can be attributed to:

- Low signal-to-noise ratio: The participants can be fully awake with little movement or fully asleep with lots of movement. Without additional data modality, these types of errors cannot be reduced.
- Unstable network dynamics: For some of the participants, the SleepNet is stuck in a state that only produces the same sleep stage class, likely because the internal network is in some local extrema, making the predictions unrealistic. Proper quality control is required during the model deployment.
- Poor sleep quality: SleepNet also had poor performance on participants whose sleep stages transitioned quickly, making a restless night of sleep. More data on poor sleepers is needed for the model to perform well.

5.4.3 Strengths and limitations

The sleep stage classification model presented has several strengths. This Section shows for the first time that the use of a self-supervised pre-trained model can improve the classification of both sleep and stages of sleep. The extensive multicentre evaluation of the sleep classification enabled a detailed characterisation of the measurement uncertainty and a less biased interpretation to inform future health association analysis. Sleep stage identification from actigraphy is highly challenging,

especially for wake periods in bed that are not characterised by wrist movement. Future work might improve sleep staging performance by incorporating additional physiological signals, such as the electrocardiogram, to improve sleep staging performance [64]. At the moment, multi-modal sensor signals are not yet available for population-scale studies with longitudinal follow-up beyond a few years [160]. Despite the best efforts to include diverse validation cohorts from different centres, the included datasets mainly consist of healthy populations from a Caucasian ethnic background. Validation in populations with chronic diseases and different ethnic backgrounds would aid in quantifying the measurement uncertainty.

Finally, sleep recorded during a laboratory-based polysomnography session is unlikely to resemble daytime napping that could occur in older adults, a key limiting factor that prevents the model from capturing daytime napping. The dominant overnight sleep period and daytime napping differ in that adults could nap when lying in bed and sitting upright. Daytime napping also tends to be more discontinuous and exposes the sleepers to more light. Thus, a more suitable ground truth for napping detection could rely on ambulatory polysomnography to capture more realistic sleep episodes throughout the day.

5.5 Summary

This Chapter developed and validated an open-source sleep staging method that substantially improves the ability to measure sleep characteristics with wrist-worn accelerometers in large biomedical datasets. The next Chapter will showcase how the proposed sleep stage classification model can be applied in a large health dataset

with longitudinal follow-up, the UK Biobank, to investigate the association between device-measured sleep and mortality outcomes.

6

Associating device-measured sleep with all-cause mortality

Contents

6.1 Introduction 110
6.2 Methods 112
6.2.1 Reliability assessment for device wear time exclusion criterion 112
6.2.2 Measurements of sleep in 100,000 UK Biobank participants 114
6.2.3 Association analysis for mortality 117
6.3 Results 119
6.3.1 Population characteristics for the UK Biobank 119
6.3.2 Descriptive analysis of device-measured sleep 122
6.3.3 Association of device-measured sleep with all-cause mortality 128
6.4 Discussion 132
6.4.1 Descriptive analysis in the UK Biobank 132
6.4.2 Association analysis between device-measured sleep and mortality 133
6.4.3 Strengths and limitations 134
6.5 Summary 135

The materials for this chapter have been made available online under the following citation: **Hang Yuan**, Tatiana Plekhanova, Rosemary Walmsley, Amy C. Reynolds,

Kathleen J. Maddison, Maja Bucan, Philip Gehrman, Alex Rowlands, David W. Ray, Derrick Bennett, Joanne McVeigh, Leon Straker, Peter Eastwood, Simon D. Kyle, and Aiden Doherty. "Self-supervised learning of accelerometer data provides new insights for sleep and its association with mortality". In: *npj Digital Medicine* (2024)

6.1 Introduction

Current understanding of how sleep is associated with health and morbidity primarily draws on studies that use self-report sleep diaries, which capture the subjective experience [161]. However, sleep diaries have a low correlation with objective device-measured sleep parameters [162, 163]. The accepted standard for sleep measurement is laboratory-based polysomnography, which monitors sleep using a range of physical and physiological signals. However, polysomnography is not feasible for use at scale due to its high cost and technical complexity. Instead, wrist-worn accelerometers are more viable to deploy in large-scale epidemiological studies because of their portability and low user burden. Nonetheless, existing accelerometer-based sleep studies on large health datasets have only focused on the differentiation between sleep and wakefulness [60, 108, 163, 164] without evaluating variations in the stages of sleep.

Chapter 5 developed and validated a sleep stage classifier (SleepNet) using one of the largest sets of multi-centre polysomnography data with state-of-the-art performance for the classification of sleep and stages of sleep using wrist-worn accelerometers. This Chapter aims to deploy SleepNet on the UK Biobank

accelerometer dataset to investigate the association between device-measured overnight sleep duration and efficiency with all-cause mortality (Figure 6.1).

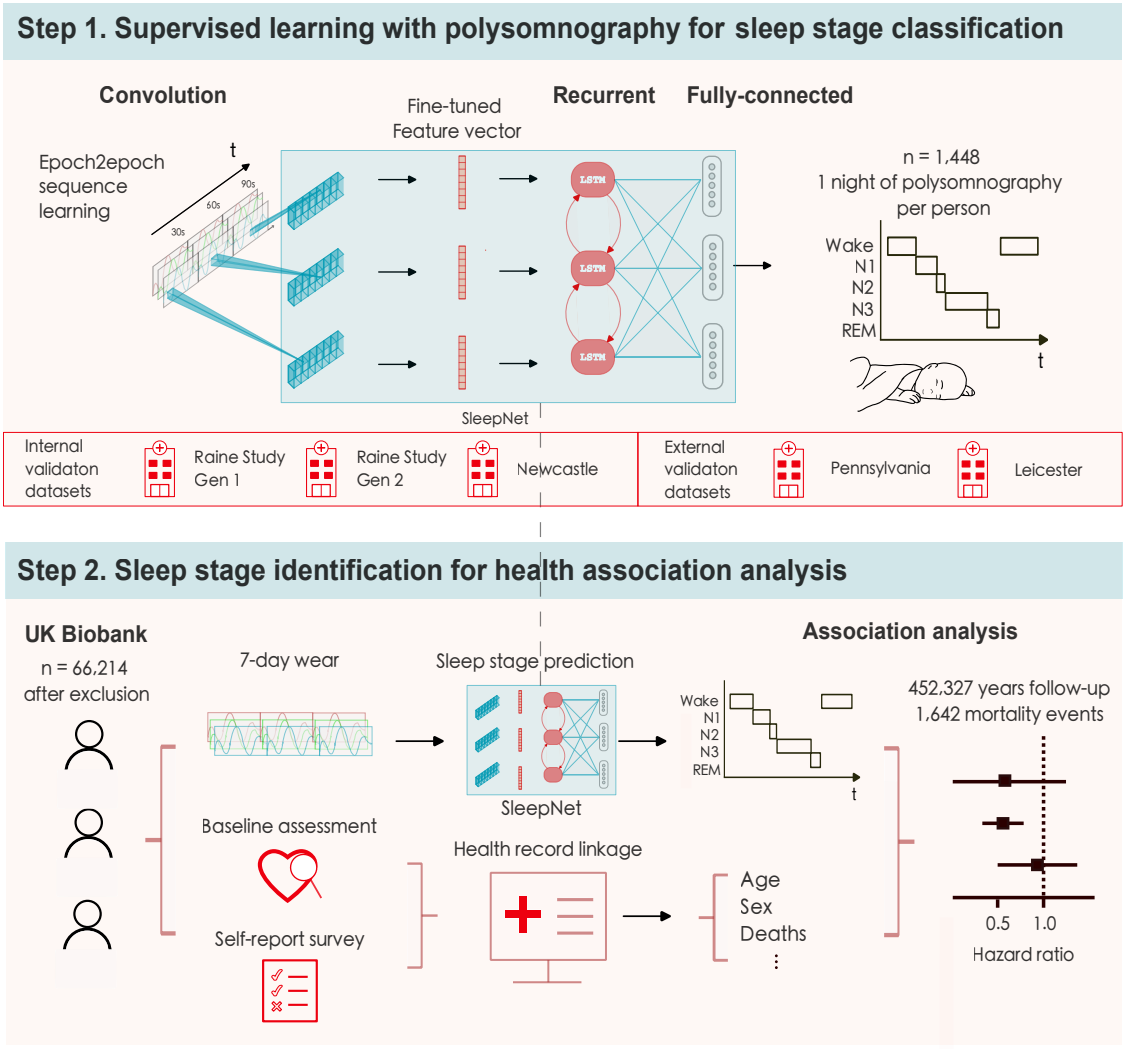


Figure 6.1: The SleepNet deployment pipeline: 1. The pre-trained feature extractor was fine-tuned with a deep recurrent network to train a sleep-stage classifier using polysomnography as the ground truth. 2. The sleep prediction model was deployed on the UK Biobank to investigate the association between device-measured sleep and mortality outcomes.

6.2 Methods

6.2.1 Reliability assessment for device wear time exclusion criterion

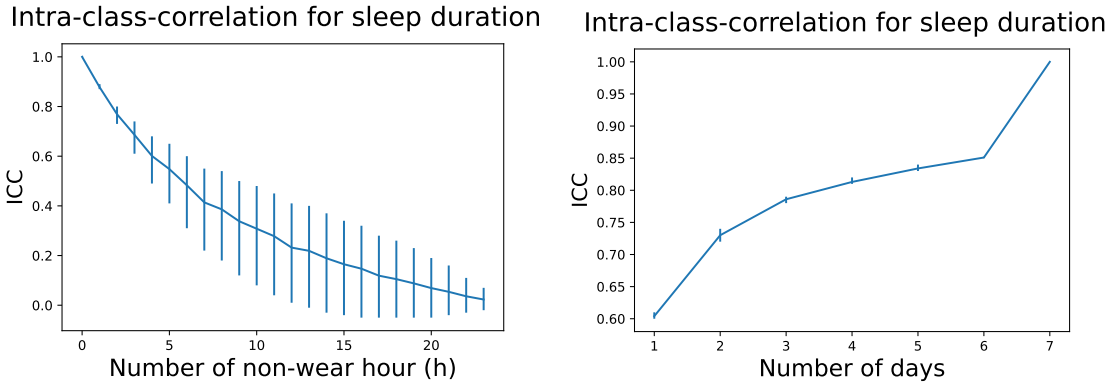


Figure 6.2: How the intraclass correlation coefficient (ICC) changes with respect to the non-wear hours (h) (left) and the number of wear days (right) in a reliability simulation using data from 27,870 participants that had zero non-wear time across a seven-day period. Mean and 95% confidence intervals are plotted.

To estimate the true effect of sleep on different health outcomes, sleep parameters that are representative of one's behaviour need to be obtained. Even though the UK Biobank accelerometry study instructed its participants to wear a wristband for seven days in a free-living environment, the participants could have taken off their devices at times. Therefore, participants with excessively long non-wear time must be discarded as their sleep estimates are unlikely to be stable. Ideally, all the participants would have perfect seven-day device wear, which was not the case. Thus, it was crucial to determine the minimum wear time for seven days so that there was a high agreement between the sleep duration computed for participants with perfect data and those computed for participants with missing data. A subset of 27,870 participants with perfect wear time during the seven-day window was

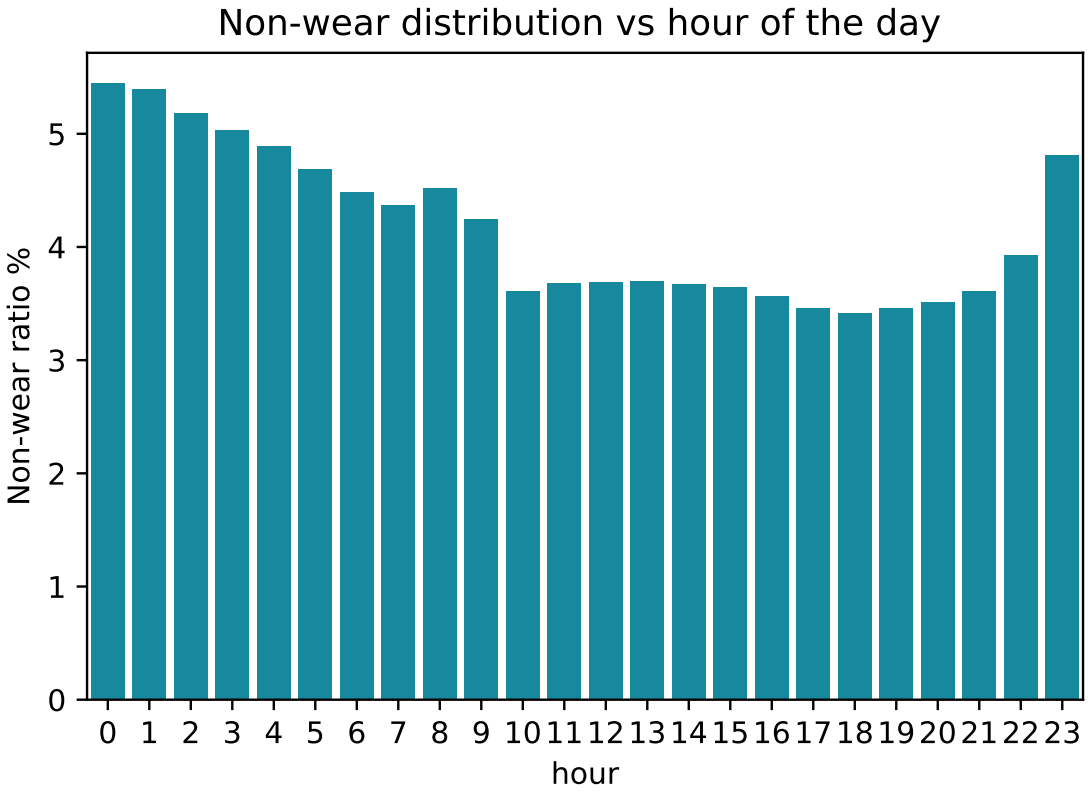


Figure 6.3: The distribution of non-wear time for all the participants from the UK Biobank.

selected. Then, missing data simulation was systematically performed by randomly removing the same hour from every day within a week or one whole day of data from each week from their recordings. The amount of simulated missing data increased in one hour or day increments until all the data was removed. Then, the agreement assessment was done on the weekly mean sleep durations computed on data before and after removing the simulated missing periods.

The intraclass correlation coefficient (ICC) was used to determine the acceptable missing time threshold. Two-way random-effects, single rater with an absolute agreement, ICC2, was used to reflect the reliability of the sleep duration measurement with missing data in the measurements [165]. Figure 6.2 depicts the ICC mean

and 95% confidence intervals for the missing non-wear hour (Figure 6.2 Left) and missing days (Figure 6.2 Right). An ICC threshold of 0.75 was used to decide the acceptable device wear range. According to the 0.75 cut-off, a maximum of two non-wear hours per day and a minimum of three days per week were suitable for obtaining stable measurements of sleep duration.

6.2.2 Measurements of sleep in 100,000 UK Biobank participants

The sleep stage estimates were obtained on the UK Biobank accelerometer dataset by applying SleepNet on the longest overnight sleep windows. Since no concurrent sleep diaries were collected in the UK Biobank, a random forest model was trained on sleep diaries with Hidden Markov Model smoothing to first obtain time in bed [61, 126]. The random forest model achieved 90%+ precision and recall for detecting sleep windows in 152 free-living participants with sleep diaries that asked two questions: "what time did you first fall asleep last night?" and "what time did you wake up (eyes open, ready to get up)?" [61]. The sleep window output from the random forest model was then used as a proxy for the time in bed. Time in bed windows within 60 minutes of one another were merged as one [69].

Finally, the SleepNet was applied on the longest window over each noon-to-noon interval to estimate the overnight sleep duration. The overnight sleep duration refers to the amount of sleep one obtains for a noon-to-noon interval in a free-living environment. The sleep window was detected using a random forest model, and the sleep stage identification on this window was then done by SleepNet. In a single night of laboratory-based polysomnography, the total sleep duration is the

same as overnight sleep. However, in a free-living environment, total sleep duration consists of napping and overnight sleep.

In addition to the minimal wear criterion, 22 hours of wear time per day for at least three days, the analysis counteracted the weekend effect by only including the participants who had at least one weekday and one weekend day during the device wear. Shift workers and participants whose data had daylight saving cross-overs were also excluded, as circadian disruption is not the focus of this Section.

Descriptive analyses were performed on the device-measured sleep parameters in the UK Biobank to check for expected variations by age, sex, device-measured physical activity level, self-reported chronotype and insomnia symptoms. Estimated marginal means, adjusted for age and sex, were also calculated for different self-rated health groups and self-reported insomnia symptoms. Finally, the correlation between self-reported sleep and device-measured sleep was assessed using a correlation matrix and a box plot.

Table 6.1: Code table for UK Biobank variables used in the study.

Variable	Source	UK Biobank fields	Coding
Month of birth	Baseline	52	
Year of birth	Baseline	34	
Device wear time	Accelerometry	90010	
Sex	Baseline	31	Female, Male
Ethnicity	Baseline	21000	White, Non-white
Smoking status	Baseline	20116	Never smoker, Ex-smoker, Current smoker
Alcohol consumption	Baseline	1558	Never, <3 times/week, 3+ times/week
Education qualification	Baseline	6138	School leaver, Further education (education beyond O-Levels/CSEs, excluding college/ university degree), Higher education (college/university degree)
Body mass index	Baseline	21001	Divided by quartile in the study population
Employment status	Baseline	6142	Employed, Not employed
Overall health rating	Baseline	2178	Poor, Fair, Good and Excellent
Self-reported total sleep duration	Baseline	1160	
Townsend Deprivation Index	Baseline	189	Divided by tertile in the study population
Overall accelerometry average	Accelerometry	90012	
Self-reported trouble falling/ staying asleep	Baseline	1200	Never/rarely, Sometimes, Usually

The UK Biobank variables are shown in Table 6.1. The age at wear time was computed using the month of birth (52) and year of birth (34) along with device wear time (90010). Participants were asked about their insomnia symptoms history (1200) by "Do you have trouble falling asleep at night or do you wake up in the middle of the night?". Four responses were possible: "never/rarely", "sometimes", "usually", and "prefer not to answer".

6.2.3 Association analysis for mortality

The primary method development of this Thesis improved the measurement of sleep and sleep stages as described in the last Chapter that the SleepNet only had a fair agreement with polysomnography, making it not a suitable candidate for health association analysis. Instead, both sleep duration and sleep efficiencies could be better measured. Hence, they are used in the subsequent epidemiological analysis.

The associations of overnight sleep duration and sleep efficiency with incident mortality were assessed using Cox proportional hazards regression. All-cause mortality was determined using death registry data (obtained by UK Biobank from NHS Digital for participants in England and Wales and from the NHS Central Register, National Records of Scotland, for participants in Scotland). Participants were censored at the earliest of UK Biobank's record censoring date for mortality data (2021-09-30 for participants in England and Wales and 2021-10-31 for participants in Scotland, with country assigned based on baseline assessment centre). Cox models used age as the timescale, and the main analysis was adjusted for sex, ethnicity, Townsend Deprivation Index, educational qualifications, smoking status,

alcohol consumption, and overall activity.

The main analysis split the participants into six groups stratified by sleep efficiency cut-off with clinical relevance. Then, five groups were created based on exact hour cut-offs in line with sleep recommendation guidelines for overnight sleep duration [166]. Four groups were created based on percentage cut-offs of clinical relevance for sleep efficiency [167]. In the sensitivity analysis, seven sleep groups were created on exact hour cut-offs to capture the variations in participants with lower and higher sleep durations. Statistical testing was not used to compare group differences because the P value was less informative when the sample size was large. The effect size was better illustrated by comparing the confidence intervals of the hazard ratios across groups [168].

In addition to the exclusions described for the analyses above, for prospective analyses for incident mortality, participants with a prior hospitalisation for restless syndrome, any cardiovascular disease or cancer (a hospital episode with primary diagnosis G473, I00-I99 or C00-C99) were also excluded.

Results are presented with their 95% confidence intervals. To enable direct comparison between groups, the Floating Absolute Risk approach was used to calculate confidence intervals for the estimate in each group, without contrast to a reference group [169–171].

6.3 Results

6.3.1 Population characteristics for the UK Biobank

Before deploying the SleepNet on the UK Biobank, participants were excluded if they had unusable accelerometer data (n=13,969), performed shift work or had their data collected over daylight saving crossovers (n=5,165), and if any missing covariate was missing (n=1,730). Participants with any prior hospitalisation for cardiovascular disease or cancer (n=16,535) were also excluded from the association analysis (Figure 6.4). In sum, 66,214 participants were included in the final analysis.

Table 6.2 describes the variations in overnight sleep duration, REM and NREM durations, and sleep efficiency across population subgroups in the UK Biobank. Older participants generally slept longer with higher sleep efficiency. Females had longer overnight sleep and NREM sleep but shorter REM sleep than males. Participants with better self-rated health had longer sleep duration and higher sleep efficiency than those with poor self-rated health. Sleep efficiency was relatively stable across different seasons and days of the week. Sleep duration was also longer for participants with employment versus those without, during the winter versus other seasons and over the weekend versus weekday average.

Figure 6.4: Participant flow diagram for the analysis of sleep and all-cause mortality in the UK Biobank. TDI: Townsend deprivation index, BMI: body mass index, SR_health: self-reported overall health, SR_insomnia: self-reported insomnia symptoms, CVD: Cardiovascular disease.

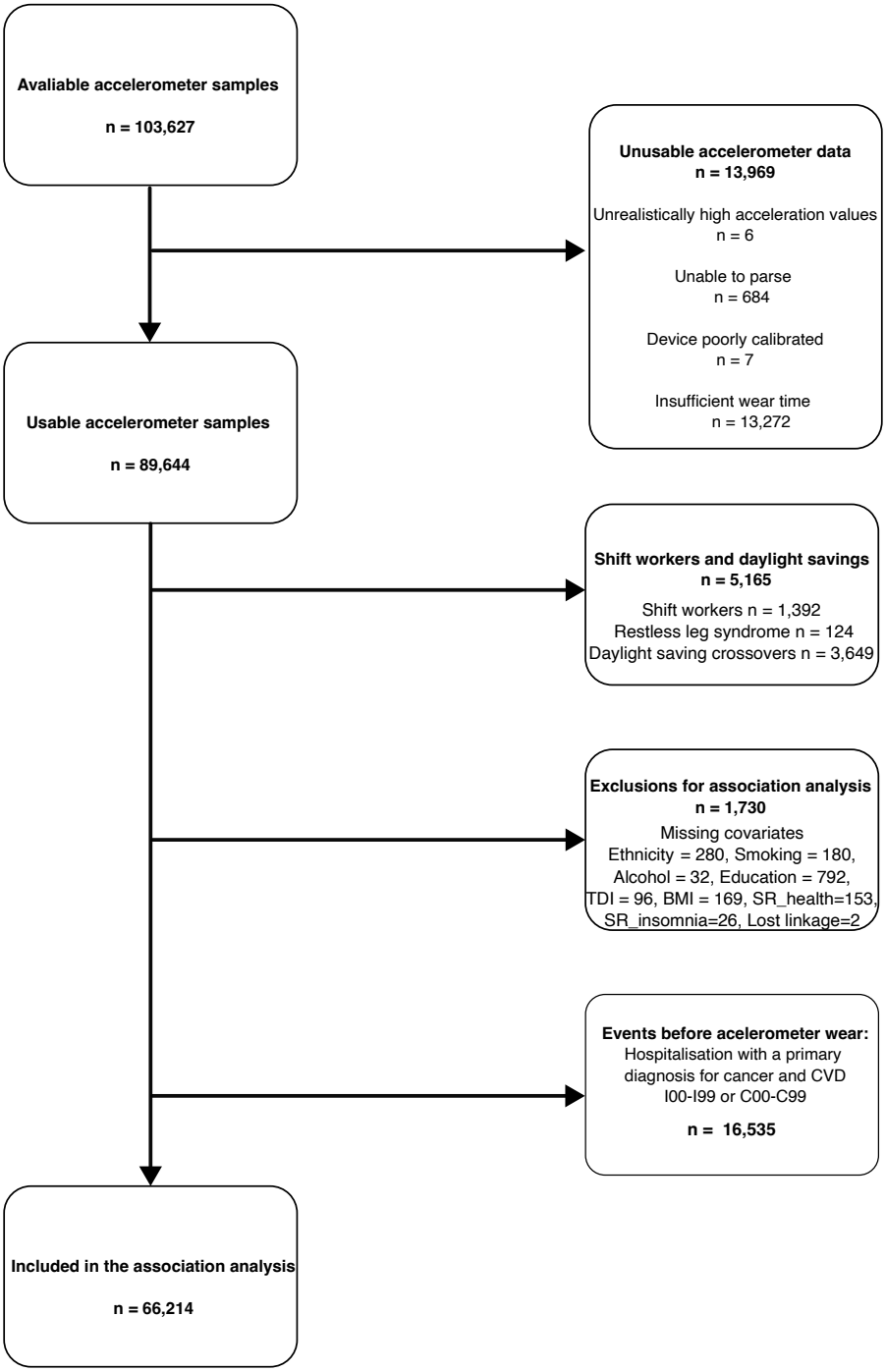


Table 6.2: Overall sleep parameters by participant characteristics in the UK Biobank (mean ± SD) for overnight sleep duration, non-rapid-eye-movement sleep (NREM), rapid-eye-movement sleep (REM), and sleep efficiency.

Characteristics		Overnight sleep	NREM	REM	Sleep efficiency
	n (%)	h/day	h/day	h/day	%
Overall	66214 (100.0)	7.5±1.0	5.0±1.0	2.5±0.9	87.9±4.9
Age, year					
40-49	6115 (9.2)	7.4±0.9	4.7±0.9	2.6±0.9	87.9±4.7
50-59	20130 (30.4)	7.4±0.9	4.9±1.0	2.5±0.9	87.7±4.9
60-69	29198 (44.1)	7.5±1.0	5.0±1.0	2.5±0.9	88.0±4.9
70-79	10771 (16.3)	7.5±1.0	5.0±1.0	2.5±0.9	88.2±5.0
Sex					
Female	38525 (58.2)	7.6±0.9	5.2±1.0	2.4±0.9	88.3±4.7
Male	27689 (41.8)	7.3±1.0	4.7±0.9	2.7±0.9	87.4±5.1
Ethnicity					
Non-white	2003 (3.0)	7.0±1.1	4.8±1.0	2.2±0.9	86.3±5.6
White	64211 (97.0)	7.5±0.9	5.0±1.0	2.5±0.9	88.0±4.9
Physical activity level					
low < 24.08 mg	22058 (33.3)	7.7±1.1	5.1±1.0	2.5±1.0	87.2±5.4
Medium 24.08-30.42 mg	22072 (33.3)	7.5±0.9	5.0±1.0	2.5±0.9	88.1±4.7
High > 30.42 mg	22084 (33.4)	7.3±0.9	4.8±0.9	2.5±0.9	88.5±4.5
Smoking status					
Never smoker	38930 (58.8)	7.5±0.9	5.0±1.0	2.5±0.9	88.0±4.8
Ex-smoker	22870 (34.5)	7.5±1.0	5.0±1.0	2.5±0.9	88.0±4.9
Current smoker	4414 (6.7)	7.3±1.0	5.0±1.0	2.3±0.9	87.4±5.5
Alcohol consumption					
Never drinker	3607 (5.4)	7.4±1.1	5.0±1.0	2.4±0.9	87.4±5.4
< 3 times per week	30074 (45.4)	7.5±1.0	5.0±1.0	2.5±0.9	87.7±5.0
3+ times per week	32533 (49.1)	7.5±0.9	4.9±1.0	2.5±0.9	88.2±4.7
Education					
School leaver	14648 (22.1)	7.6±1.0	5.1±1.0	2.5±0.9	87.5±5.1
Further education	21700 (32.8)	7.5±1.0	5.0±1.0	2.5±0.9	87.8±5.0
Higher education	29866 (45.1)	7.4±0.9	4.9±1.0	2.5±0.9	88.2±4.7
Townsend Deprivation Index					
Least deprived (<-3.8)	16552 (25.0)	7.5±0.9	5.0±1.0	2.6±0.9	88.1±4.8
Second least deprived (-3.8 to -2.5)	16554 (25.0)	7.5±0.9	5.0±1.0	2.6±0.9	88.0±4.8
Second most deprived (-2.5 to -0.2)	16552 (25.0)	7.5±1.0	5.0±1.0	2.5±0.9	87.9±4.9
Most deprived (> -0.2)	16556 (25.0)	7.4±1.0	5.0±1.0	2.4±0.9	87.8±5.1
BMI					
<18.5, underweight	397 (0.6)	7.5±1.0	5.1±1.0	2.5±0.9	89.1±4.7
18.5-24.9, normal	26759 (40.4)	7.6±0.9	5.0±1.0	2.6±0.9	88.4±4.6
25-29.9, overweight	26920 (40.7)	7.5±1.0	4.9±1.0	2.5±0.9	87.8±4.9
30+, obese	12138 (18.3)	7.3±1.1	5.0±1.0	2.3±0.9	87.1±5.4
Employment					
Employed	41640 (62.9)	7.4±0.9	4.9±1.0	2.5±0.9	87.9±4.8
Not employed	24574 (37.1)	7.6±1.0	5.1±1.0	2.5±0.9	88.0±5.0
Self-rated health					
Poor	1282 (1.9)	7.4±1.3	5.0±1.1	2.3±1.0	87.0±6.0
Fair	9162 (13.8)	7.4±1.1	5.0±1.0	2.4±0.9	87.3±5.3
Good	40120 (60.6)	7.5±0.9	5.0±1.0	2.5±0.9	87.9±4.9
Excellent	15650 (23.6)	7.5±0.9	4.9±1.0	2.6±0.9	88.4±4.6
Day					
Weekday	66214 (100.0)	7.4±1.0	4.9±1.0	2.5±0.9	88.0±5.2
Weekend	66214 (100.0)	7.7±1.2	5.1±1.2	2.6±1.1	87.8±6.2
Wear season					
Spring	14717 (22.2)	7.5±0.9	4.9±1.0	2.5±0.9	87.9±4.9
Summer	18203 (27.5)	7.4±0.9	4.9±1.0	2.4±0.9	88.2±4.8
Autumn	18682 (28.2)	7.5±1.0	5.0±1.0	2.5±0.9	87.9±4.9
Winter	14612 (22.1)	7.6±1.0	5.0±1.0	2.6±0.9	87.7±5.0

6.3.2 Descriptive analysis of device-measured sleep

The correlation coefficients between device-measured sleep parameters during accelerometer wear and self-reported total sleep duration at baseline assessment were not greater than 0.25 (Figure 6.5). Among all the correlations assessed between self-reported sleep and device-measured sleep, the highest correlation coefficient was observed between self-reported total sleep duration and device-measured overnight sleep duration, with a correlation coefficient of 0.25.

The distributions of device-measured overnight sleep duration tend to have a greater variability for participants who self-reported to have less than 5 or greater than 10 hours of total sleep duration (Figure 6.6). Overall, sleep stage distribution was similar for males and females aged between 45 and 75, with NREM sleep fluctuating around 5 hours and REM sleep fluctuating around 2.5 hours per night (Figure 6.7). No major differences were seen between females and males.

Expected sleep-wake patterns were found in population subgroups. For example, timing of the sleep opportunity for participants with a self-reported "morning" chronotype was about one hour earlier when compared with those that had a self-reported "evening" chronotype (Figure 6.8a). A similar but shorter sleep phase advance (∼30 mins) was observed in participants who were most physically active compared to the participants that were least physically active (Figure 6.8b). Between groups that had a history of self-reported insomnia symptoms versus those who did not, participants with a history of insomnia symptoms were less likely to be in REM sleep on average during the overnight sleep window (Figure 6.8c and Figure 6.8d). Participants with a history of self-reported insomnia symptoms

tended to have a longer overnight sleep duration but with a lower sleep efficiency (Figure 6.9). The sleep stage distributions for different population subgroups were similar between weekdays and weekends, with a slight delay sleep timing over the weekend (Figure 6.10).

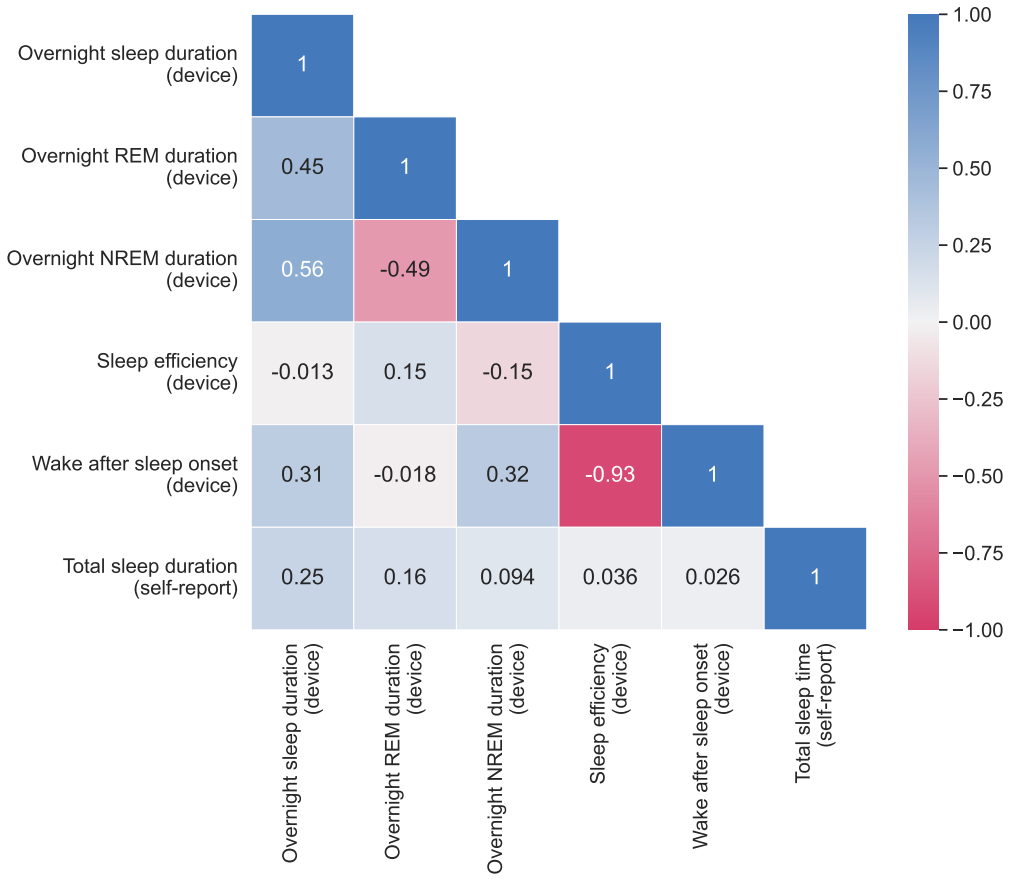


Figure 6.5: Correlation matrix for device-measured and self-reported sleep parameters on the UK Biobank. The self-reported total sleep duration was obtained via questionnaire at baseline assessment in the UK Biobank. REM: rapid-eye-movement sleep, NREM: non-rapid-eye-movement sleep.

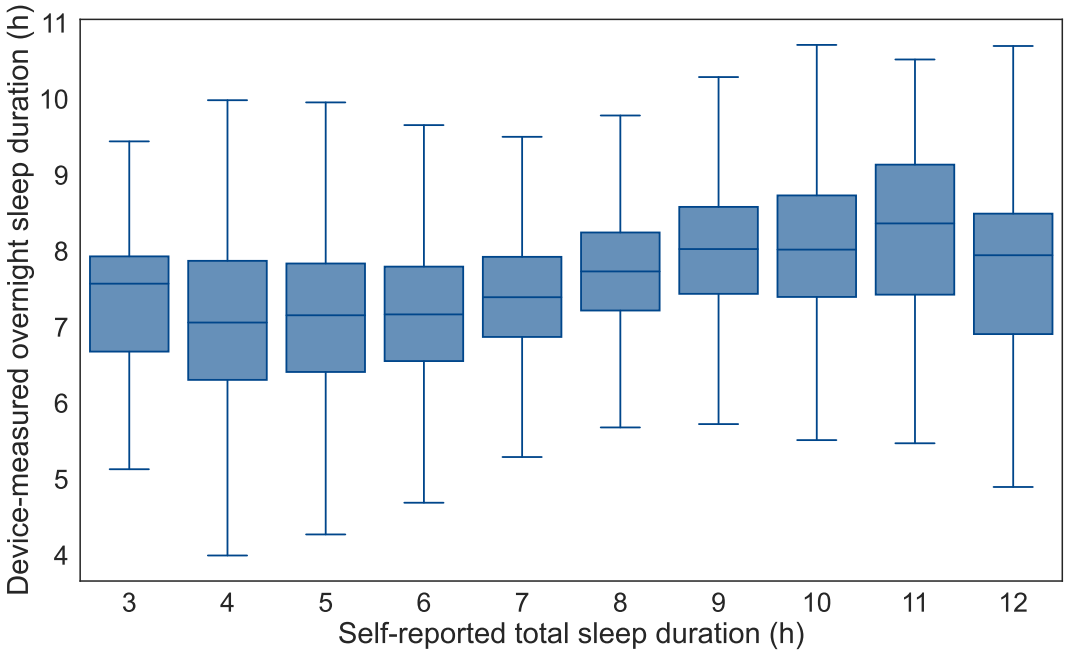


Figure 6.6: Box plots showing the distributions of device-measured overnight sleep duration against self-reported total sleep duration. The box whiskers reflect the lowest and highest data points that are 1.5 times of the inter-quartile-range from the median.

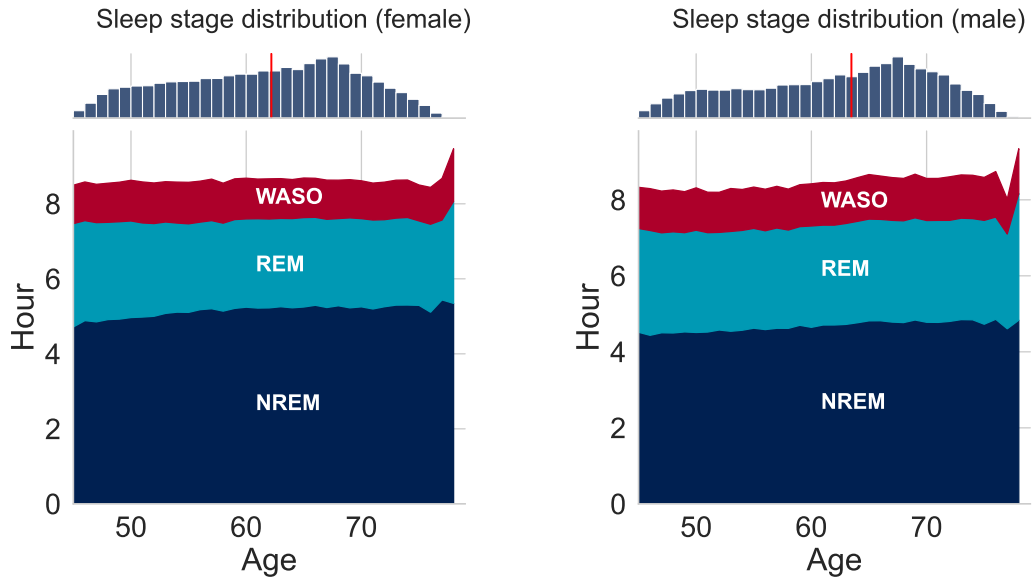
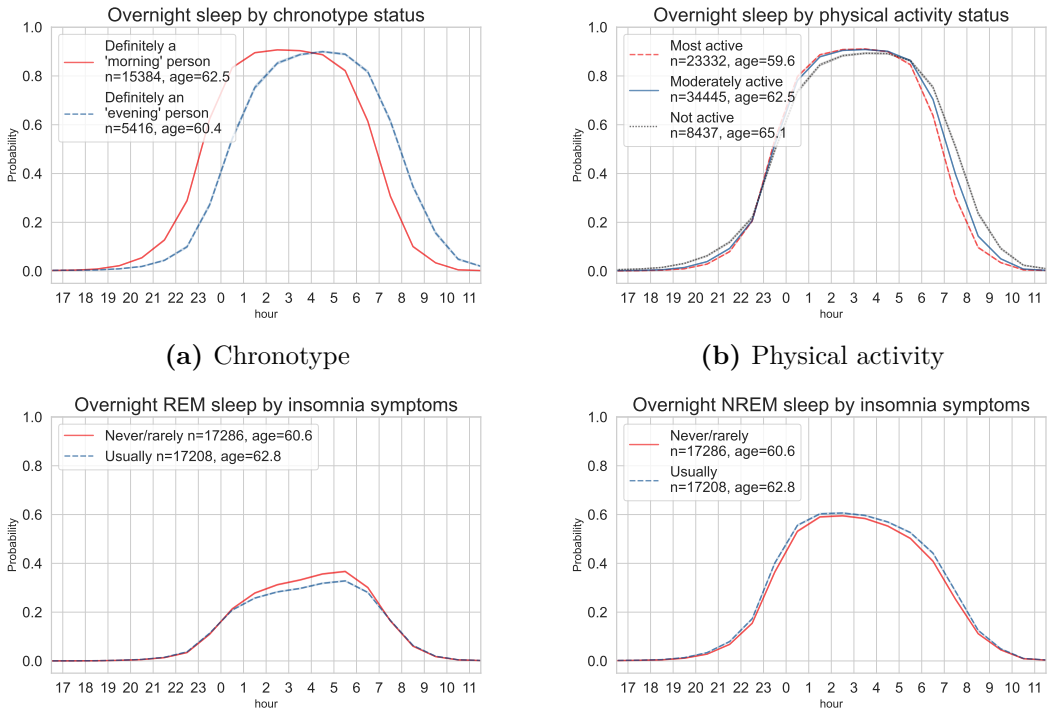


Figure 6.7: The average device-measured sleep stage distribution with respect to age for both females (left) and males (right) on the UK Biobank. The histograms on the top show the age distribution for the participants. The red vertical line denotes the median age for each sex. WASO: wake after sleep onset; REM: rapid-eye-movement sleep; NREM: non-rapid-eye-movement sleep.



(a) Chronotype

(b) Physical activity

(c) Insomnia symptoms REM

(d) Insomnia symptoms NREM

Figure 6.8: Device-measured sleep probability trajectories throughout the day for the UK Biobank participants. Top: variations of the average overnight sleep probability for the participants with self-reported "morning" and "evening" chronotype (a) and the overnight sleep distributions across thirds of device-measured physical activity level (b). Bottom: variations of the average REM (c) and NREM (d) probability in participants with a history of self-reported insomnia symptoms versus those without. REM: rapid-eye-movement sleep; NREM: non-rapid-eye-movement sleep.

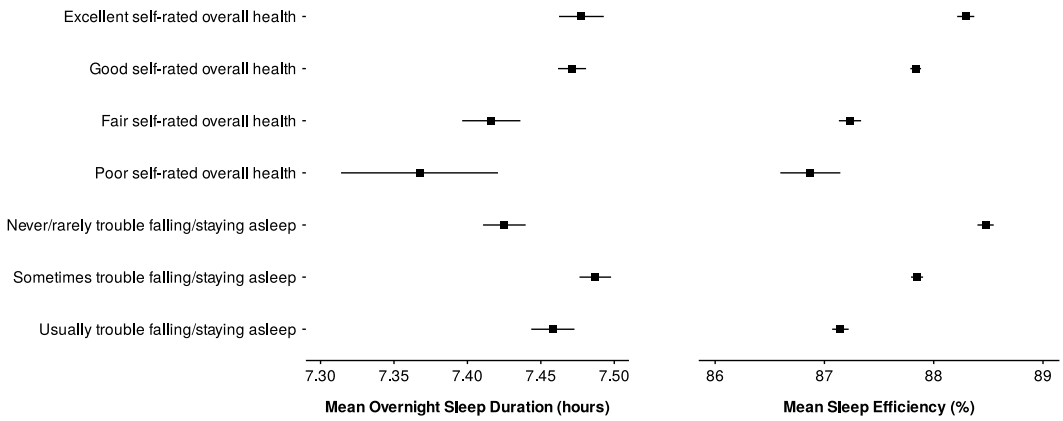
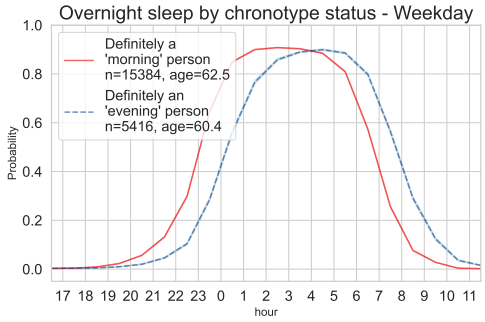
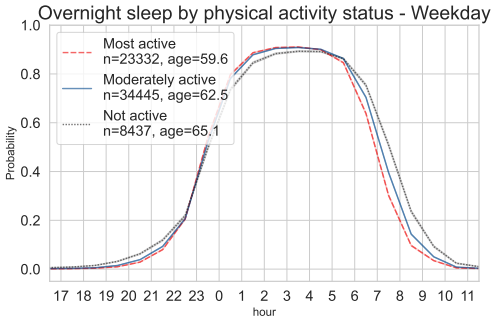
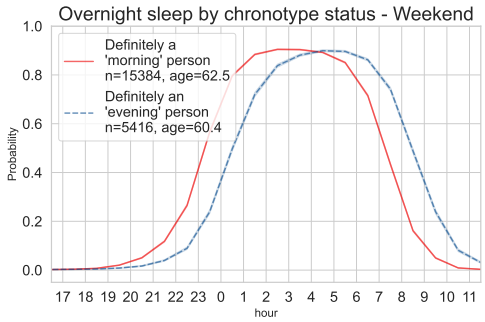
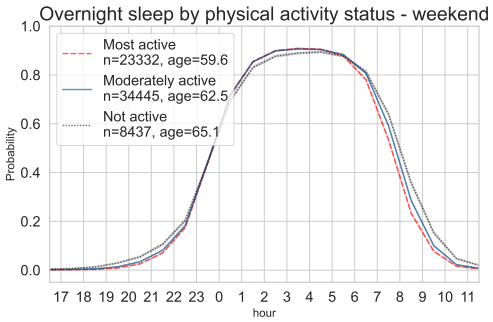


Figure 6.9: Adjusted marginal mean (95% confidence interval) device-measured mean overnight sleep duration and mean sleep efficiency by self-reported overall health status and insomnia history in the UK Biobank. Mean overnight sleep duration and sleep efficiency were adjusted for age and sex.

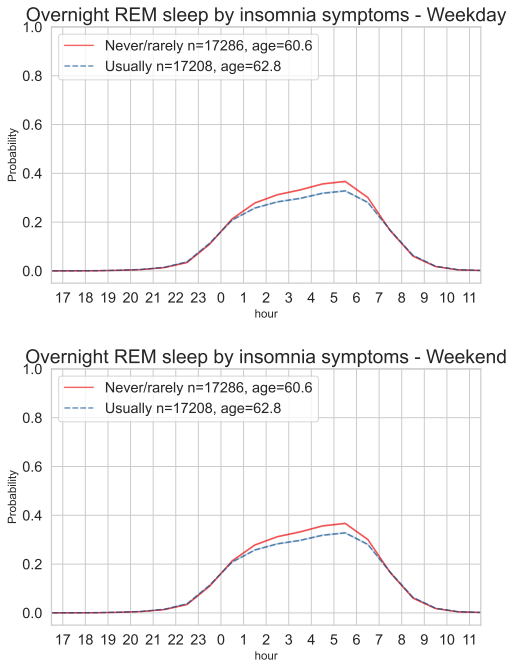


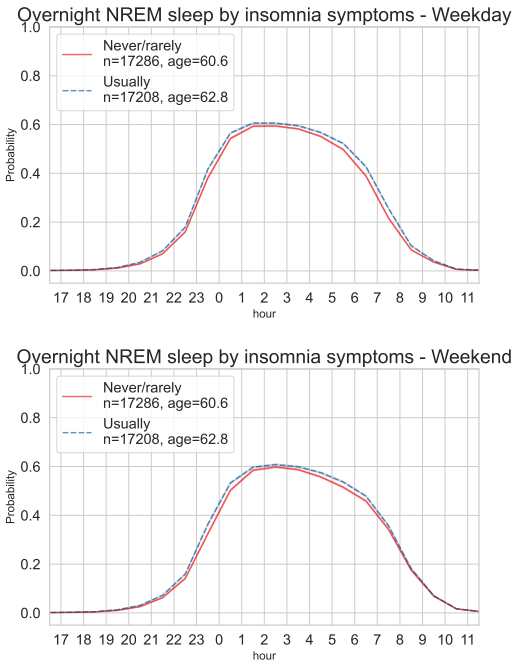






(a) Chronotype **(b)** Physical activity



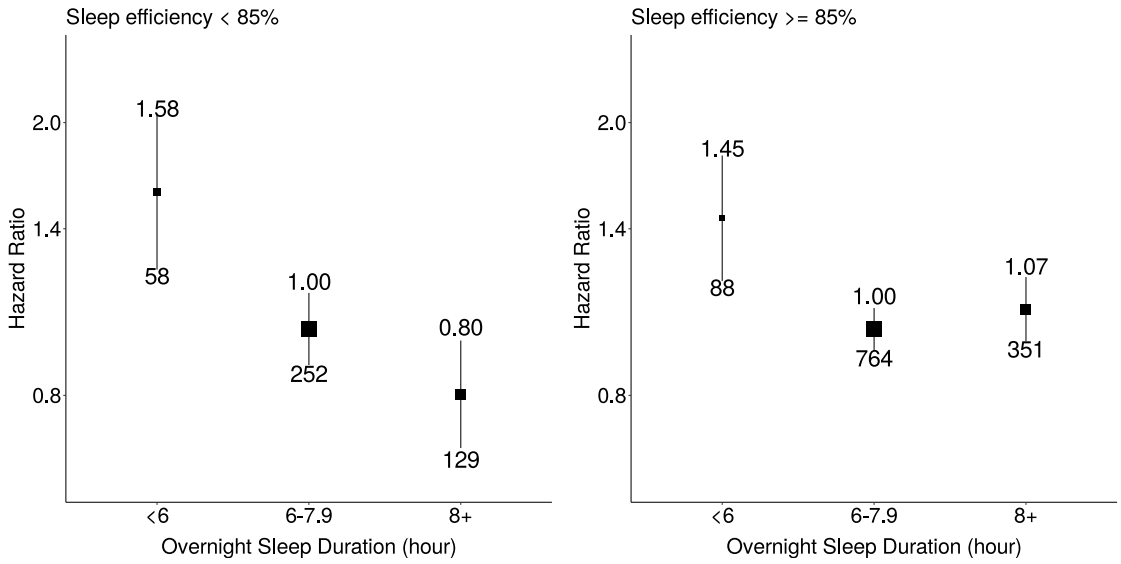


(c) Insomnia symptoms REM **(d)** Insomnia symptoms NREM

Figure 6.10: Device-measured sleep probability trajectories throughout the day for the UK Biobank participants (weekday vs weekend). Top: variations of the average overnight sleep probability for the participants with self-reported "morning" and "evening" chronotype (a) and the overnight sleep distributions across thirds of device-measured physical activity level (b). Bottom: variations of the average REM (c) and NREM (d) probability in participants with a history of self-reported insomnia symptoms versus those without. Rapid-eye-movement sleep (REM), and non-rapid-eye-movement sleep (NREM).

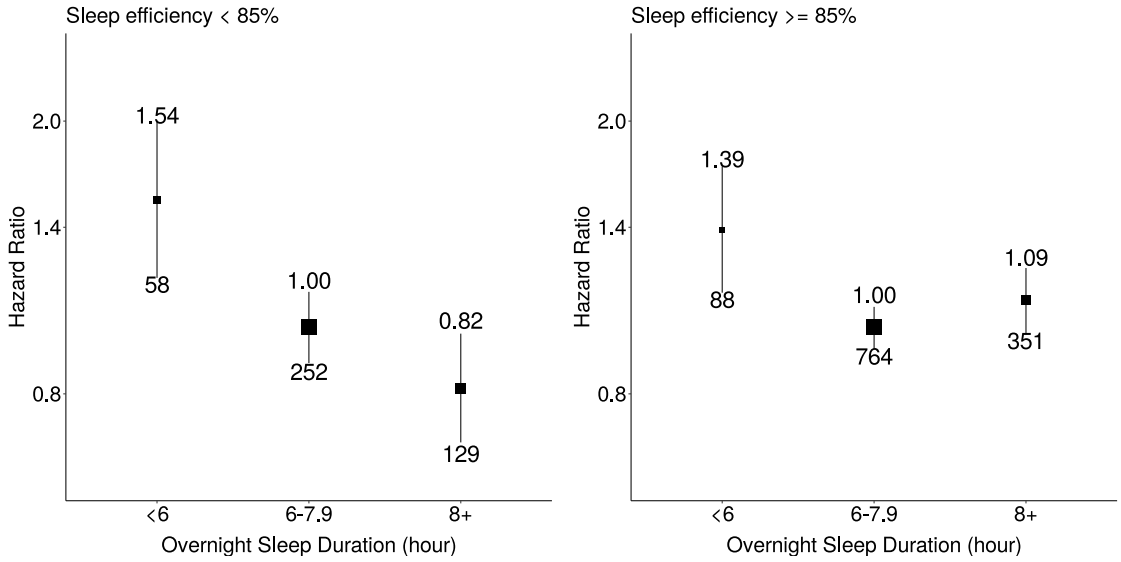
6.3.3 Association of device-measured sleep with all-cause mortality

Over 452,327 person-years of the follow-up (median: 6.9 years), 1,642 mortality events among 66,214 participants were observed. Short sleepers (<6 hours) had a higher risk of mortality in groups of low sleep efficiency (Hazard ratios (HRs): 1.58; 95% confidence intervals (CIs): 1.19 to 2.11) and high sleep efficiency (HRs: 1.45; 95% CIs: 1.16 to 1.181) compared to participants with normal sleep duration (6 to 7.9 hours, Figure 6.11 Top Row). The risk of all-cause mortality appeared to decrease linearly as sleep efficiency increased. However, a non-linear association was observed in the association for overnight sleep duration (Figure 6.12 Top Row). When further adjusted for BMI, associations of overnight sleep duration and sleep efficiency with all-cause mortality were slightly attenuated (Figure 6.11 Bottom Row and 6.12 Bottom Row). After removing the first two years of follow-up for the main analysis, Figure 6.13 showed weaker associations in both low and high sleep efficiency groups, indicating signs for reverse causality.



(a) Adjusted without body mass index

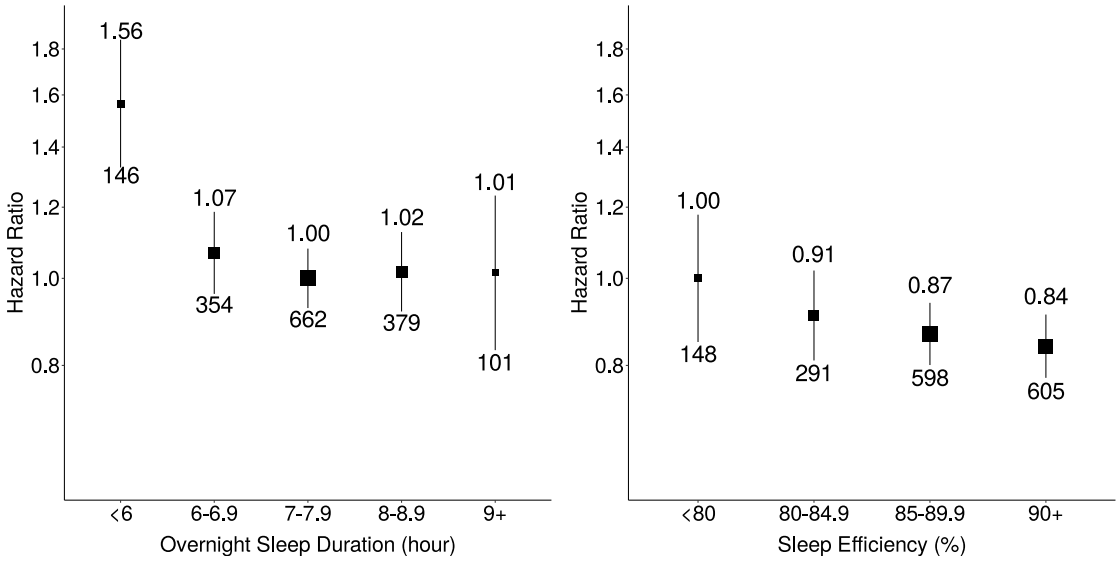
(b) Adjusted without body mass index



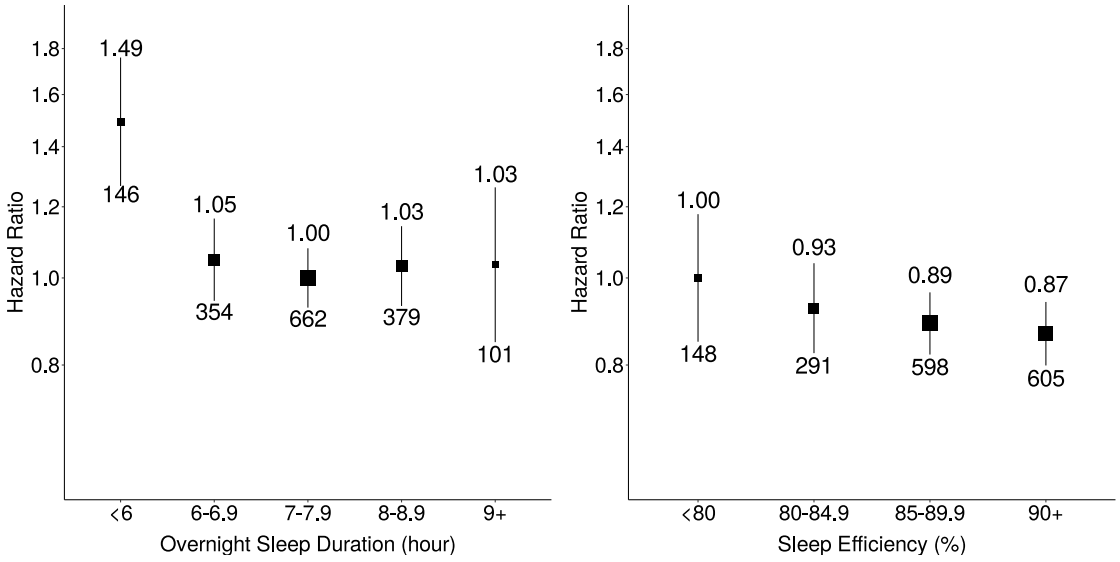
(c) All

(d) All

Figure 6.11: Associations of overnight sleep duration with all-cause mortality for groups with low and high sleep efficiency. The model used 1,642 events among 62,214 participants. Age was used as the timescale and adjusted for sex, ethnicity, Townsend Deprivation Index of baseline address (split by quarter in the study population), educational qualifications, smoking status, alcohol consumption (Never, <3 times/week, 3+ times/week), overall activity (measured in milli-gravity units). Areas of squares represent the inverse of the variance of the log risk. The I bars denote the 95% confidence interval for the floated risks.

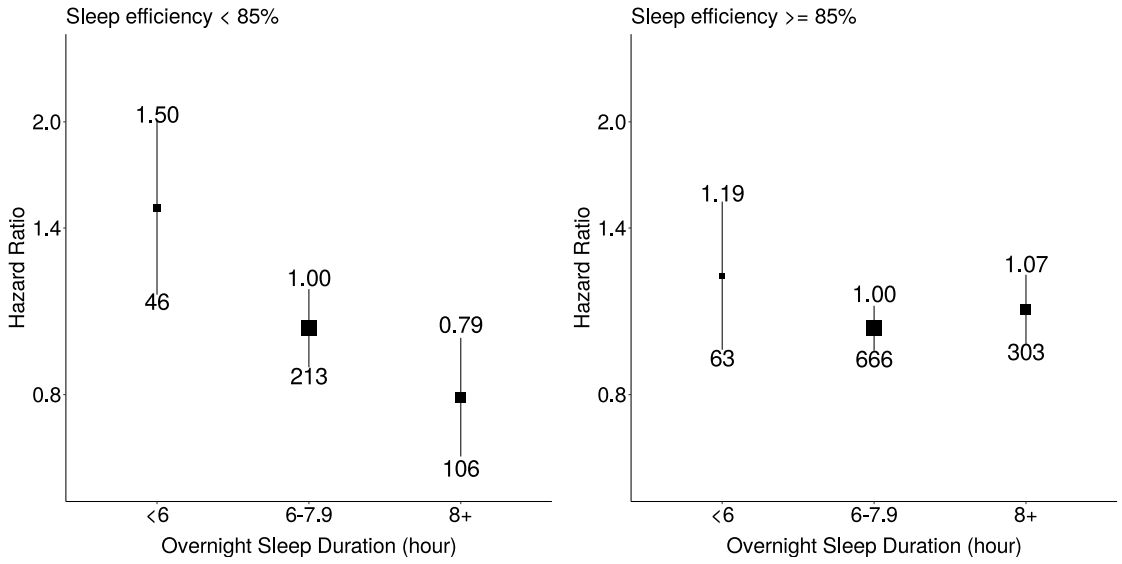


(a) Adjusted without body mass index **(b)** Adjusted without body mass index



(c) All **(d)** All

Figure 6.12: Associations of overnight sleep duration (a) and sleep efficiency (b) with all-cause mortality. The model used 1,642 events among 62,214 participants. Age was usedas the timescale and adjusted for sex, ethnicity, Townsend Deprivation Index of baseline address (split by quarter in the study population), educational qualifications, smoking status, alcohol consumption (Never, <3 times/week, 3+ times/week), overall activity (measured in milli-gravity units), and body mass index. Areas of squares represent the inverse of the variance of the log risk. The I bars denote the 95% confidence interval for the floated risks.



(a) Adjusted without body mass index **(b)** Adjusted without body mass index

Figure 6.13: Associations of overnight sleep duration with all-cause mortality for groups with low and high sleep efficiency after removing the first two years of follow-up. The model used 1,642 events among 62,214 participants. Age was used as the timescale and adjusted for sex, ethnicity, Townsend Deprivation Index of baseline address (split by quarter in the study population), educational qualifications, smoking status, alcohol consumption (Never, <3 times/week, 3+ times/week), overall activity (measured in milli-gravity units). Areas of squares represent the inverse of the variance of the log risk. The I bars denote the 95% confidence interval for the floated risks.

6.4 Discussion

6.4.1 Descriptive analysis in the UK Biobank

This Chapter deployed a deep learning sleep classification model on the UK Biobank accelerometer dataset to obtain sleep parameter estimates at the population level. The descriptive analysis demonstrated good face validity using accelerometer-based sleep classification for population inference, as exemplified by the summary statistics conformed with expected variations by sleep timing, population demographics, health status. Using the machine-learning derived sleep duration and sleep efficiency in an epidemiological analysis of 66,214 participants, the results indicated that shorter sleep duration was associated with an increased risk of all-cause mortality regardless of sleep continuity, indexed by sleep efficiency.

Self-report sleep duration from questionnaires had a low correlation with accelerometer-derived sleep parameters, including device-measured overnight sleep duration (correlation=0.25). The disparity seemed to be most pronounced in participants who reported short or long sleep duration. Participants with a low self-reported total sleep duration tended to under-estimate, and participants with a high self-reported total sleep duration tended to over-estimate their sleep duration (Figure 6.6). A low-level agreement in the UK Biobank could be due to the multi-year time gap between the baseline questionnaire and the accelerometer enrichment collection. However, Girschik and colleagues also found a similar difference pattern between the sleep questionnaire and accelerometer-derived sleep parameters in 56 participants who filled in a sleep log right before wearing an accelerometer for

seven days [172]. This disparity highlights the need for epidemiological analysis using device-measured sleep.

The descriptive analysis represented one of the largest studies for device-measured sleep stages on a population scale. The sleep stage distribution for REM, NREM, WASO remained relatively stable for the majority of the participants in the UK Biobank for both females and males, except for those over 80 years of age. Changes for participants aged over 80 might not generalise due to the low number of participants in this age group. A meta-analysis of polysomnography in healthy individuals from childhood to old age in healthy individuals also observed a stable sleep stage distribution between the age of 40 and 80 [18]. Participants with self-report insomnia symptoms had similar NREM sleep duration but a shorter REM sleep duration when compared with those without insomnia symptoms, which might support the theory that restless REM sleep plays a key role in the manifestation of insomnia [173].

6.4.2 Association analysis between device-measured sleep and mortality

Short overnight sleep duration has been linked with mortality outcomes in self-report and accelerometer-based studies [174, 175]. However, few studies have investigated the joint effect of sleep duration and efficiency. One recent study has suggested that participants with short and long total sleep time had an increased risk after accounting for sleep efficiency [176]. However, this analysis did not find that long overnight sleep duration was associated with increased risk, potentially because daytime naps were not included in the measurement of overnight sleep

duration. Daytime napping has been found to be associated with an increased risk of cardiovascular events and deaths in those with longer nighttime sleep [177]. Several other studies with limited sizes have suggested a U-shape association between device-measured sleep and mortality, which was not found in the results from this Chapter [174]. Instead, the data support adverse associations with short sleep duration only, which is concordant with pre-clinical human and animal studies [178]. Finally, when considering the sleep duration concerning the time spent in other activities throughout the day in a compositional manner, Walmsley and colleagues have reported that reallocating time from light or sedentary activities to sleep did not increase cardiovascular risks. Cardiovascular risks only increased when more time was spent in sleep than moderate-to-vigorous activities [61]. Future studies could investigate the joint effects of sleep and physical activity on disease outcomes.

6.4.3 Strengths and limitations

This study has several strengths, including the analysis of sleep architecture in a large, prospective Biobank with longitudinal follow-up. Compared with self-reported sleep questionnaires that only captured sleep duration to the nearest hour, accelerometer-based methods like SleepNet can provide more fine-grained sleep duration and efficiency estimates. The extensive multicentre evaluation of the sleep classification allowed for the characterisation of the measurement uncertainty and a less biased interpretation of the health association analysis. Sleep stage identification from accelerometers is highly challenging, especially for wake periods in bed that are not characterised by wrist movement. With the proposed SleepNet,

the face validity of the sleep parameters was assessed using the UK Biobank, opening the door to obtaining sleep stage estimates for population health inference in other datasets. While future work might improve sleep staging performance by incorporating additional physiological signals, such as electrocardiogram, to improve sleep staging performance, multi-modal sensor signals are not yet available for population-scale studies with longitudinal follow-up beyond a few years [160].

The UK Biobank mainly consists of healthy populations from a Caucasian ethnic background, which could bias the association analysis [179]. Thankfully, several large Biobanks aiming to recruit participants from different ethnic backgrounds are being collected, which could shed light on how ethnic differences could influence the cause-effect association for sleep [41, 180]. The health association showed signs of reverse causality, which could overestimate the effect of sleep. It would be important to assess the effect of reverse causality when more years of follow-up become available [181]. Finally, there might be a risk of residual confounding. For instance, the baseline employment status might not reflect the true employment during the device-wear, which occurred several years later. Future studies could consider replicating the analysis using the China Kadoorie Biobank, for which the employment and accelerometer data were collected roughly at the same time [41].

6.5 Summary

Results in this Chapter demonstrate a key opportunity to use device-measured sleep parameters to further enhance the impact of sleep on our health and overall well-being. Sleep is multi-dimensional and structurally complex. Cost-effective

wearable sensor datasets provide longitudinal and often orthogonal perspectives about sleep compared to commonly used sleep diaries and polysomnography. Both sleep and physical activity are two key lifestyle behaviours that have a significant impact on our health. Even though long-term changes in physical activity might be difficult to achieve [182], effective interventions for sleep have been identified, potentially leading to greater health benefits [183, 184]. The association analysis in this Chapter counteracts the social stigma that one does not need to sleep long if the sleep quality is preserved, calling for further research in leveraging device-measured sleep in large health datasets that might yield additional insights for public health intervention or behaviour changes in each individual.

7 Discussion

Contents

7.1 Summary of findings 138
7.2 Thesis limitations 141
7.3 Ethical implications 143
7.4 Future work 145
7.4.1 Multi-modal sleep classification 145
7.4.2 Towards consumer-grade devices 146
7.4.3 Beyond accelerometer-based sleep staging 147
7.5 Conclusion 147

This thesis aimed to identify performance of accelerometer-based sleep stage classification and investigate the association between device-measured sleep and mortality risks. To address the proposed aims, this thesis had three objectives:

- Develop a self-supervised feature extractor by leveraging large-scale unlabelled human sensing data and assess its generalisability across a range of human activity recognition tasks and sleep classification;
- Develop a state-of-the-art accelerometer-based sleep stage classifier validated against multi-centre polysomnography;

- Study the association between device-measured sleep and all-cause mortality using the developed sleep classifier on a large health dataset with longitudinal follow-up.

7.1 Summary of findings

This thesis started with a systematic review of existing literature on accelerometer-based sleep stage classification (**Chapter 2**). The seven studies included in the review indicated that accelerometer-based sleep staging methods have some ability to classify sleep stages against polysomnography. A large heterogeneity was noted in the reporting of the sleep stage classification performance. Even though the AASM guidelines divide sleep into five stages, arbitrary stage groupings were seen to perhaps reduce the classification difficulty. Nonetheless, adhering to the same reporting standard will aid comparability across studies, especially the reporting of classification performance in the five-stage setting. Furthermore, the review highlighted a reliance on hand-crafted features used by prior sleep stage classification algorithms and the validation using polysomnography data of limited size, prompting the need for additional method development in the remainder of this thesis.

Feature learning from large-scale unlabelled data The first aim of the thesis was to develop a feature extractor by learning directly from large-scale unlabelled data to remove the need for hand-crafted features and enhance the subsequent model development with a limited amount of labelled data. **Chapter 4** built a multi-task self-supervised model to learn features relevant to human motion dynamics by pre-training on 700,000 person-days of unlabelled human activity tracker data from

the UK Biobank. The self-supervised features were able to improve the performance of sleep staging and to a less extend sleep-wake classification. The features were also evaluated on human activity recognition, a common accelerometer application with more open-access benchmark datasets. The pre-trained self-supervised features showed generalisability across activity classes, devices, and populations. When compared against random forest models, the relative performance ranged between 2.5% to 100% for F1 on the seven benchmark datasets. The cluster analysis for the pre-trained features indicated that self-supervised approach learned information relevant to activity frequency and intensity. Compared to the previous work, this was the first pre-trained model in human activity recognition that did not require training again on the downstream dataset to be used as a feature extractor [65, 128–130]. All the pre-trained model weights have been made available freely online to allow others to build high-performance activity recognition models, especially when limited data label is available. Even though the downstream evaluation for the pre-trained features was limited to human activity recognition, the self-supervised pre-training was designed to learn generic features about human motion dynamics that could benefit other accelerometer-based applications such as sleep stage classification.

Sleep stage classification from accelerometers The second aim was to develop an accelerometer-based sleep stage classifier by training on multi-centre polysomnography. The self-supervised feature extractor was used to improve the classification performance, as lab-based polysomnography labels were expensive to acquire. The sleep stage classifier in **Chapter 5** consisted of the self-supervised

feature extractor and a deep recurrent neural network (SleepNet). The SleepNet with self-supervision obtained the best performance against current state-of-the-art with an F1 score of 0.75 for sleep and wake classification and 0.573 for wake/REM/NREM classification. On the external datasets, the mean bias was 34.7 minutes for total sleep duration, -2.6 minutes for REM duration, and 32.1 minutes for NREM duration. Overall, SleepNet had a fair to moderate agreement with polysomnography. The measurement error for sleep-wake classification was more reliable than for sleep stage classification. In particular, the Kappa score for sleep stage classification was 0.39. Without obtaining a moderate agreement, the accelerometer-based sleep staging may not be suitable for health association analysis.

Accelerometer-measured sleep and mortality risk The final aim of this thesis was to investigate the association between device-measured sleep and mortality risks. In **Chapter 6**, the SleepNet was deployed on the UK Biobank accelerometry dataset to investigate the association between overnight sleep duration and sleep efficiency with all-cause mortality. The summary sleep parameters showed good face validity with expected sleep parameter distribution across population sub-groups. Cox regression analysis suggested that short overnight sleep duration (<6 hours) was associated with higher mortality risk regardless of sleep quality. Concretely, short sleepers had a higher risk of mortality in groups of low sleep efficiency (Hazard ratios (HRs): 1.58; 95% confidence intervals (CIs): 1.19 to 2.11) and high sleep efficiency (HRs: 1.45; 95% CIs: 1.16 to 1.181) compared to participants with normal sleep duration (6 to 7.9 hours). Further adjustment for body mass index

only attenuated the effect slightly. The association analysis showcased how device-measured sleep could enrich the understanding of the cause-effect association of sleep and diseases that would not be possible otherwise.

7.2 Thesis limitations

There are several limitations regarding the results presented. The first limitation is the lack of population diversity for the data used in Chapter 4 to Chapter 6. The lack of population diversity could limit the model generalisability: both the self-supervised pre-trained model and SleepNet were developed using datasets that consisted of mostly Caucasians and those who tended to be healthier and received more years of education than the general population [117]. The developed models may fail to generalise to populations of different ethnic and social-demographic backgrounds. Future work should incorporate datasets from more diverse populations when they become available [185].

The second limitation is the potential risk for confounding and reverse causality in the association analysis. Because of the time gap between the baseline questionnaire and the accelerometer collection, social-demographic variables, including employment status and Body Mass Index, despite already factored in the analysis, might not be the same during the accelerometer collection. The imperfectly measured covariates could introduce regression dilution bias, leading to an underestimation of the effects [32]. Furthermore, sleep is known to suffer from reverse causality in observational studies as the sleep parameters likely have changed before death. Our sensitivity analysis indeed suggested evidence for reverse causality. Future

work should investigate differences in the associations using different follow-up time cut-offs when longer follow-up time has been accumulated [181]. Alternatively, one could use genetics techniques such as Mendelian Randomization on the sleep traits to investigate the causal effects, which are not influenced by reverse causality and residual confounding due to the random assortment of alleles at conception and the genetics remain largely stable from birth [186].

The third limitation is that SleepNet cannot identify nap time during the day but overnight sleep. Lab-based polysomnography is the gold standard for sleep stage scoring, requiring the study participants to lie in bed throughout. However, napping in older adults and shift workers tends to occur in different environments, including chairs and couches. In addition, napping could involve different physical locations and body gestures, rendering the validity of overnight lab-based polysomnography unclear. Ambulatory polysomnography could be used to capture nap time under different environments and body gestures to measure sleep stages to complement lab-based polysomnography.

The fourth limitation is that lab-based polysomnography could still be considered an indirect sleep measurement despite its wide acceptance among sleep clinicians and researchers. Sleep is characterised by physical immobility, reduced responses to external stimuli and varying patterns in brain activities. In a historical context, the definitions for sleep stages evolved from a one-state definition to the REM and NREM duality [187, 188], then gradually the more granular NREM sleep definitions were developed [10, 45]. The iterative development of different sleep stages depends on empirical observations and expert consensus. However, even

today, the inter-rater agreement between polysomnography scorers peaks around 0.8 in Kappa [157], leaving room for further improvement despite substantial agreement for most stages. To measure the true sleep stages, one might have to directly record the electrophysiological signals in the human brain rather than the voltage changes from the scalp. Nonetheless, polysomnography remains the best tool available for sleep stage measurement. As the understanding of sleep improves, future sleep scoring guidelines might improve agreement between different raters, providing better ground truth for model development.

7.3 Ethical implications

This thesis sits in the intersection between ubiquitous computing, sleep, population health and artificial intelligence. Thus, the underlying ethical implications need to be addressed in the context of the broader impact of this work to individuals and to the society as a whole.

The methodological advancement in my thesis aimed to improve the measurement of sleep and physical activity that can then be used in population health studies. However, the open-sourced self-supervised models could expose the public to the misuse of the technology if adversaries build models to capture hand-writing to eavesdrop on sensitive content [189] and other activities that the general public might not feel comfortable with sharing. On the one hand, making all the developed models open-access will enable others to build high-performance human sensing models. The decision to share the model weights without restrictions depends on

an implicit assumption that the net benefits of open-access models will outweigh the potential damage from the misuse of the technology.

Today, the annual wearable device shipments are around 520 million devices globally, according to the International Data Corporation [190]. The prevalence of wearable devices is projected to grow in the coming decade. Nonetheless, those who could benefit most from having a wearable device use it the least, as found in a US representative survey [191]. Wearable device users tend to be healthier, wealthier and receive more years of education. The uptake of wearable devices has been slowest in more deprived populations, which could create health equity issues and hinder social mobility as existing consumer-grade wearables range between 50 and 500 pounds per device and often require some technical sophistication for usage. Early evidence might support the need for public health programs to offer free or discounted smartwatches at the population level to help end-users enjoy the benefits of an active lifestyle [192, 193].

Lastly, privacy is an important ethical issue for the use of wearable devices in health research. Wearables are one of the few technologies that enable passive continuous monitoring with high-frequency data over a long period. A recent systematic review of 64 studies found that when using common wearable devices, including accelerometers, electroencephalography (EEG) and electrocardiogram (ECG), the reidentification rates ranged between 86% and 100% [194]. Reidentification is defined as the deliberate act of determining an individual‘s identity from a purposely anonymised data source. Even though prior works have only attempted the reidentification of individuals from much smaller datasets, and

the risk of reidentification remains small, the breach of confidentially can have a dire impact on public health research programs and on the individuals being reidentified [195]. Even though the risk of not sharing the data is greater due to the potential for selection bias and algorithmic bias in health research, as a community, it is important to think of appropriate storage and sharing of wearable sensor data and even use privacy-persevering methods in future work, for instance, federated learning [196] and differential privacy [197].

7.4 Future work

This work has shown great promise in using device-measured sleep in epidemiological studies. The following section outlines possible directions for further exploration.

7.4.1 Multi-modal sleep classification

Sleep, at its core, is a physiological behaviour coupled with immobility. Wrist-worn accelerometers can capture key parameters related to sleep-wake patterns with a high enough precision for clinical applications [76]. While accelerometer-based sleep staging was found to have a fair agreement with polysomnography, the lack of physiological signals limits the room for improvement from additional methodological development. Other modalities from wearables can be used to improve the classification of sleep stages in a daily living environment. A recent review on sleep stagings using wearable devices found the most useful modalities to be EEG and photoplethysmography (PPG) [83]. EEG-based sleep staging was the only modality to reliably classify all the sleep stages thus far, as EEG directly measures brain activities. Nonetheless, wearable EEG devices could still carry a

high user burden for large-scale epidemiological research. Using a combination of PPG and accelerometers for sleep stagings might strike a better balance between user burden and accuracy [64]. Inevitably, multi-modal sleep classification models will also benefit from using a pre-trained multi-modal model, which is currently not available in the wearable sensing space. The development of such a foundation model for multi-modal wearable sensing data offers immense opportunities for not only sleep classification but also other important downstream applications such as human activity recognition [198] and cardio-fitness prediction [199].

7.4.2 Towards consumer-grade devices

Existing large-scale cohort studies mostly used research-grade devices because research-grade devices can store raw accelerometry data without charging up to several weeks. Minimal participant burden and raw data availability are two key factors in the device selection for epidemiological studies. Thanks to the continuous improvement in consumer-grade devices, the collection of wearable sensing signals in cohort studies has an increasing appetite for consumer-grade devices [180]. The advantages of consumer-grade devices are greater availability, much longer-term device wear, multi-modal sensors, and potential ease of data collection by retrospective consenting. However, most sleep detection algorithms used in consumer-grade devices are proprietary, making it difficult to compare the sleep parameters from one device to another. Future work could be developed to understand the agreement and differences between research-grade and consumer-grade devices [200] and how to complement the high-fidelity raw data from research-grade devices with large-scale

longitudinal measurements from consumer-grade devices.

7.4.3 Beyond accelerometer-based sleep staging

Accurate accelerometer-based sleep staging is instrumental in answering scientific and clinical queries. However, when it comes to intervention at an individual and the population level, it is less clear what interventions will effectively improve the duration or the architecture of sleep stages [201]. Compared to the gold-standard laboratory-based polysomnography, wrist-worn accelerometers excel at capturing wake-sleep patterns and regularities over a long period, which are also important pillars of sleep health but that can be acted on via behaviour changes [202].

In addition to accelerometer-based sleep staging, one could also study how environmental factors can influence sleep and human diseases. Recently, others have used light and temperature sensors embedded in accelerometers from the UK Biobank to investigate how the light exposure and the diurnal rhythms of wrist temperature could elucidate one's risk of diseases [203, 204]. Even though the absolute value for the light and temperature sensor might not be satisfactory, the rhythmicity of light and temperature could yield orthogonal insights about circadian health, which is closely related to sleep health [205].

7.5 Conclusion

This thesis has assessed the performance of accelerometer-based sleep stage classification and investigated the association between device-measured sleep and mortality risks. The proposed self-supervised representation learning framework will allow others to build high-performance human activity recognition and sleep

classification models using wrist-worn accelerometers. Descriptive analyses for device-measured sleep stages were made possible for the first time by developing a recurrent neural network with a large-scale multi-centre polysomnography dataset. Device-measured short sleep duration was associated with an elevated mortality risk in those with high and low sleep efficiency. The overall associations remained strong in sensitivity analyses and after further adjustment for covariates. Future work could incorporate multi-modal signals from diverse populations to obtain more accurate device-measured sleep for both epidemiological research and clinical studies. The analyses presented form the foundation to investigate the association between device-measured sleep and diseases in large health datasets with longitudinal follow-up.

Appendices

A

Data for the systematic review of the performance of actigraphy in measuring sleep stages

A.1 Data extraction item

- Authors, Reference number (text, number)
- Sample size (number)
- Funding sources (text)
- PSG type (text): at home recording or in a lab?
- PSG scoring rubric (text): R&K or AASM standard+year
- PSG channel/montage (text)
- PSG level if applicable: Level I/II (attended/unattended)
- Scoring method (sing-choice): manual, semi-automated, automated.

- Population type (text): age, population type (e.g. have sleep disorders or not),
- Recruitment approach (text)
- Device type/brand/mode
- Device placement
- Actigrpahy sampling rate
- Best performing algorithm name (text)
- Data source: activity count or raw accelerometery
- Sensitivity, specificity, accuracy and F1 for N1, N2, N3, combined NREM, R, and W.
- Balanced accuracy, kappa and F1 scores for all four sleep stages and W.

A.2 Search terms

Table A.1: Search terms for Ovid.

Term index	Search term
1	exp Actigraphy/
2	wearable*.mp.
3	actigraphy.ti,ab.
4	actimetry.ti,ab.
5	Actiwatch.mp.
6	Somnowatch.mp.
7	Sleepwatch.mp.
8	“WatchPAT 100”.mp.
9	Action-W.mp.
10	“WP 100”.mp.
11	WatchPAT.mp.
12	Motionlogger.mp.
13	Actillume.mp.
14	wGT3X-BT.mp.
15	Gt3x.mp.
16	geneactiv.mp.
17	camntech.mp.
18	axivity.mp.
19	Fitbit.mp
20	“Apple watch”.mp
21	acceleromet*.mp.
22	accelometer*.mp.
23	actigraph.ti,ab.
24	exp Wrist/
25	wrist.ti,ab.
26	exp Monitoring, Physiologic/
27	exp "Sleep stage"/
28	“Sleep classification".mp.
29	26 or 27 or 28
30	24 or 25
31	29 and 30
32	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 31
33	exp Polysomnography/
34	exp Electroencephalography/
35	Polysomnography.ti,ab.
36	Electroencephalography.ti,ab.
37	33 or 34 or 35 or 36
38	32 and 37
39	limit 38 to (english language and (adult < 18 to 64 years> or aged < 65+ years>))

Embase and Medline The above strategy is adopted from the strategy used in [78]. The search terms were used for both Embase and Medline. When performing

Table A.2: Search terms for Web of Science.

Term index	Search term	Result
1	TS=Actigraphy	
2	TS=wearable*	
3	TS= actimetry	
4	TS=Actiwatch	
5	TS=Somnowatch	
6	TS=Sleepwatch	
7	TS="WatchPAT 100"	
8	TS="Action-W"	
9	TS="WP 100"	
10	TS=WatchPAT	
11	TS=Motionlogger	
12	TS=Actillume	
13	TS=wGT3X-B	
14	TS=Gt3x	
15	TS=geneactiv	
16	TS=camntech.	
17	TS=axivity	
18	TS=Fitbit	
19	Fitbit.mp	
20	TS="Apple watch"	
21	TS=acceleromet*	
22	TS=accelometer*	
23	TS=actigraph	
24	TS=Wrist	
25	TS=Monitoring Physiologic	
26	TS=sleep stages	
27	TS=sleep classification	
28	#24 OR #25 OR #26	
29	#28 AND #24	
30	#1 or #2 or #3 or #4 or #5 or #6 or #7 or #8 or #9 or #10 or #11 or #12 or #13 or #14 or #15 or #16 or #17 or #18 or #19 or #20 or #21 or #22 or #22 or #23 or #29	
31	TS=Polysomnography	
32	TS=Electroencephalography	
33	#31 OR #32	
36	#30 and #33	

the the search, the time limit for Embase was 1974 to present and the time limit for Medline was 1946 to present. The search terms and results are shown in Supplement Table A.1

Web of Science The search syntax was adopted from the Embase syntax. The search terms and results are shown in Supplement Table A.2

Google Scholar Stop at page 20 because no results are relevant on page 21. 10 links per page. (Actigraphy OR wearable OR actigraphy OR Actimetry OR Actiwatch OR Somnowatch OR Sleepwatch OR "WatchPAT 100" OR "Action-W" OR 'WP 100" OR WatchPAT OR Motionlogger OR Actillume OR "wGT3X-BT" OR Gt3x OR geneactiv OR camntech OR axivity OR Fitbit OR Apple watch OR accelerometer OR accelerometry OR actigraph OR (Wrist AND ((Monitoring OR Physiologic) OR Sleep stage OR Sleep classification)) AND (Polysomnography OR Electroencephalography)

We put all the references into EndNote to search for abstracts before uploading them to Covidence for further processing.

Excluded studies

Table A.3: Exclusion reasons for studies that might be eligible. AASM: American Academy of Sleep Medicine.

Reference	Reason for exclusion
Haghayegh et al. [206]	Comparison to single channel EEG only
Darbyshire et al. [207]	Comparison to single channel EEG only
Toedebusch et al. [208]	Comparison to single channel EEG only
Menghini et al. [209]	Conference abstract only
Kumar and Dash [210]	Conference abstract only
Pollet et al. [211]	Conference abstract only
Kim, Shin, and Byun [212]	Conference abstract only
Kawhawage et al. [213]	Conference abstract only
Kahawage et al. [214]	Conference abstract only
Olsen et al. [215]	Conference abstract only
Samman et al. [216]	Conference abstract only
Collen and Lesage [217]	Conference abstract only
Chung et al. [218]	Conference abstract only
Alakuijala, Jokela, and Toppila [219]	Conference abstract only
Dietz-Terjung, Martin, and Schobel [220]	Conference abstract only
Baker et al. [221]	Conference abstract only
Allen et al. [222]	Conference abstract only
Devine et al. [223]	Duplicate
Domingues, Paiva, and Sanches [98]	No comparison to actigraphy alone
Alcantara et al. [224]	No comparison to actigraphy alone
Haghayegh et al. [99]	No comparison to actigraphy alone
Herscovici et al. [97]	No comparison to actigraphy alone
Willemen et al. [225]	No comparison to actigraphy alone
Hedner et al. [226]	No comparison to actigraphy alone
Kalkbrenner et al. [227]	Wrong measurements: no wrist actigraphy
Zaffaroni et al. [228]	Wrong measurements: no wrist actigraphy
Concheiro-Moscoso et al. [229]	Wrong measurements: no wrist actigraphy
Turetken, Van Zaen, and Delgado-Gonzalo [230]	Wrong measurements: use only hear rate
Haghayegh et al. [231]	Wrong model input: classification model input is not acceleration.

Moreno-Pino et al. [232]	Wrong model input: classification model input is not acceleration.
Liang and Chapa-Martell [233]	Wrong model input: classification model input is not acceleration.
Wulterkens et al. [234]	Wrong model input: classification model input is not acceleration.
Miller et al. [235]	Wrong model input: classification model input is not acceleration.
Miller et al. [236]	Wrong model input: classification model input is not acceleration.
Renevey et al. [237]	Wrong model input: classification model input is not acceleration.
Kuula and Pesonen [238]	Wrong model input: classification model input is not acceleration.
Muzet et al. [239]	Wrong model input: classification model input is not acceleration.
Zhang et al. [240]	Wrong model input: classification model input is not acceleration.
Kahawage et al. [241]	Wrong model input: classification model input is not acceleration.
Ribeiro et al. [242]	Wrong model input: classification model input is not acceleration.
Fujimoto, Ding, and Takahashi [243]	Wrong model input: classification model input is not acceleration.
Fonseca et al. [244]	Wrong model input: classification model input is not acceleration.
Fonseca et al. [245]	Wrong model input: classification model input is not acceleration.
Zambotti et al. [246]	Wrong model input: classification model input is not acceleration.
Zhang et al. [247]	Wrong model input: classification model input is not acceleration.
Schneider, Zhang, and Schwartz [248]	Wrong model input: classification model input is not acceleration.
Cook et al. [249]	Wrong model input: classification model input is not acceleration.
Bresler et al. [250]	Wrong model input: classification model input is not acceleration.
Choi et al. [251]	Wrong model input: classification model input is not acceleration.
Chinoy et al. [252]	Wrong model input: classification model input is not acceleration.
Willemen et al. [253]	Wrong model input: classification model input is not acceleration.
Stucky et al. [254]	Wrong model input: classification model input is not acceleration.

Onder et al. [255]	Wrong model input: classification model input is not acceleration.
Kanady et al. [256]	Wrong sleep parameters: no AASM sleep stage results were reported.
Kim et al. [257]	Wrong sleep parameters: no AASM sleep stage results were reported.
Barouni et al. [258]	Wrong sleep parameters: no AASM sleep stage results were reported.
Barouni et al. [259]	Wrong sleep parameters: no AASM sleep stage results were reported.
Beecroft et al. [260]	Wrong sleep parameters: no AASM sleep stage results were reported.
Choi et al. [261]	Wrong sleep parameters: no AASM sleep stage results were reported.
Cook, Prairie, and Plante [262]	Wrong sleep parameters: no AASM sleep stage results were reported.
Edinger et al. [263]	Wrong sleep parameters: no AASM sleep stage results were reported.
Danzig et al. [264]	Wrong sleep parameters: no AASM sleep stage results were reported.
Jean-Louis et al. [265]	Wrong sleep parameters: no AASM sleep stage results were reported.
Honma et al. [266]	Wrong sleep parameters: no AASM sleep stage results were reported.
Cook, Prairie, and Plante [267]	Wrong sleep parameters: no AASM sleep stage results were reported.
Mikulec et al. [268]	Wrong sleep parameters: no AASM sleep stage results were reported.
Aktaruzzaman et al. [269]	Wrong sleep parameters: no AASM sleep stage results were reported.
Shen et al. [270]	Wrong study design: no parallel actigraphy and polysomnography.
Levendowski et al. [271]	Wrong study design: no parallel actigraphy and polysomnography.
Bell et al. [272]	Wrong study design: no parallel actigraphy and polysomnography.
Barash et al. [273]	Wrong study design: no parallel actigraphy and polysomnography.
Alcantara et al. [274]	Wrong study design: no parallel actigraphy and polysomnography.

PRISMA 2020 Main Checklist

TITLE			
Title	1	Identify the report as a systematic review.	page 1
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist	
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	page 2
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	page 2
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	page 3; subsection: eligible studies
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	page 3; subsection: search strategies
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	page 3; subsection: search strategies
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	page 4; subsection: study selection and data extraction

(continued)

Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	page 4; subsection: study selection and data extraction
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	page 4; subsection: study selection and data extraction
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	page 4; subsection: study selection and data extraction
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	page 4; subsection: data evaluation
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	not applicable: no quantitative synthesis done
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item 5)).	not applicable: no quantitative synthesis done

(continued)

	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	not applicable: no quantitative synthesis done
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	not applicable: no quantitative synthesis done
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	not applicable: no quantitative synthesis done
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	not applicable: no quantitative synthesis done
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	not applicable: no quantitative synthesis done
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	not applicable: no quantitative synthesis done
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	not applicable: no quantitative synthesis done
RESULTS			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	page 4; last paragraph
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	page 4; last paragraph
Study characteristics	17	Cite each included study and present its characteristics.	figure 1

(continued)

Risk of bias in studies	18	Present assessments of risk of bias for each included study.	table 3
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	table 2
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	page 6; subsection: risk of bias of included studies
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	not applicable: no statistic syntheses done
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	not applicable: no statistic syntheses done
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	not applicable: no statistic syntheses done
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	page 12; subsection: limitations
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	not applicable for diagnostic test review but for treatment
DISCUSSION			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	page 11; discussion 1st paragraph
	23b	Discuss any limitations of the evidence included in the review.	page 12; subsection: limitations
	23c	Discuss any limitations of the review processes used.	page 12; subsection: limitations

(continued)

	23d	Discuss implications of the results for practice, policy, and future research.	page 12; 3rd paragraph
OTHER INFORMATION			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	page 3
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	page 3
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	page 12; section: limitations
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	page 13; section: acknowledgements
Competing interests	26	Declare any competing interests of review authors.	page 13; section: conflicts of interests
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	page 13; sections: data and materials sharing

PRIMSA Abstract Checklist

TITLE			
Title	1	Identify the report as a systematic review.	Yes
BACKGROUND			
Objectives	2	Provide an explicit statement of the main objective(s) or question(s) the review addresses.	Yes
METHODS			
Eligibility criteria	3	Specify the inclusion and exclusion criteria for the review.	Yes
Information sources	4	Specify the information sources (e.g. databases, registers) used to identify studies and the date when each was last searched.	Yes
Risk of bias	5	Specify the methods used to assess risk of bias in the included studies.	No
Synthesis of results	6	Specify the methods used to present and synthesize results.	Yes
RESULTS			
Included studies	7	Give the total number of included studies and participants and summarise relevant characteristics of studies.	Yes
Synthesis of results	8	Present results for main outcomes, preferably indicating the number of included studies and participants for each. If meta-analysis was done, report the summary estimate and confidence/credible interval. If comparing groups, indicate the direction of the effect (i.e. which group is favoured).	Yes
DISCUSSION			
Limitations of evidence	9	Provide a brief summary of the limitations of the evidence included in the review (e.g. study risk of bias, inconsistency and imprecision).	No
Interpretation	10	Provide a general interpretation of the results and important implications.	Yes
OTHER			
Funding	11	Specify the primary source of funding for the review.	No
Registration	12	Provide the register name and registration number.	Yes

From: Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. MetaArXiv. 2020, September 14. DOI: 10.31222/osf.io/v7gm2. For more information, visit: www.prisma-statement.org

B

Data for self-supervised learning for human activity recognition

B.1 Datasets

Table B.1: The license, consent and device used for each dataset

Dataset	Personal Info	Licence	Consent
UK Biobank	✗	Non-exclusive licence (but not ownership rights) for permitted purpose only	Informed consent for research
Capture-24	✗	CC Attribution	Informed consent for research
Rowlands	✗	Non-specified	Informed consent for research
WISDM	✗	Non-specified	Non-specified
REALWORLD	✗	Non-specified	Non-specified
Opportunity	✗	Non-specified	Non-specified
PAMAP2	✗	Non-specified	Informed consent for research
ADL	✗	Free for research	Non-specified

B.2 Results

Table B.2: Test accuracies on different tasks 1000 subjects.

AoT	Permutation	Time warped
0.93	.90	.97

C

Data for sleep classification using wrist-worn accelerometers

C.1 Additional data for Agreement between derived sleep parameters and polysomnography

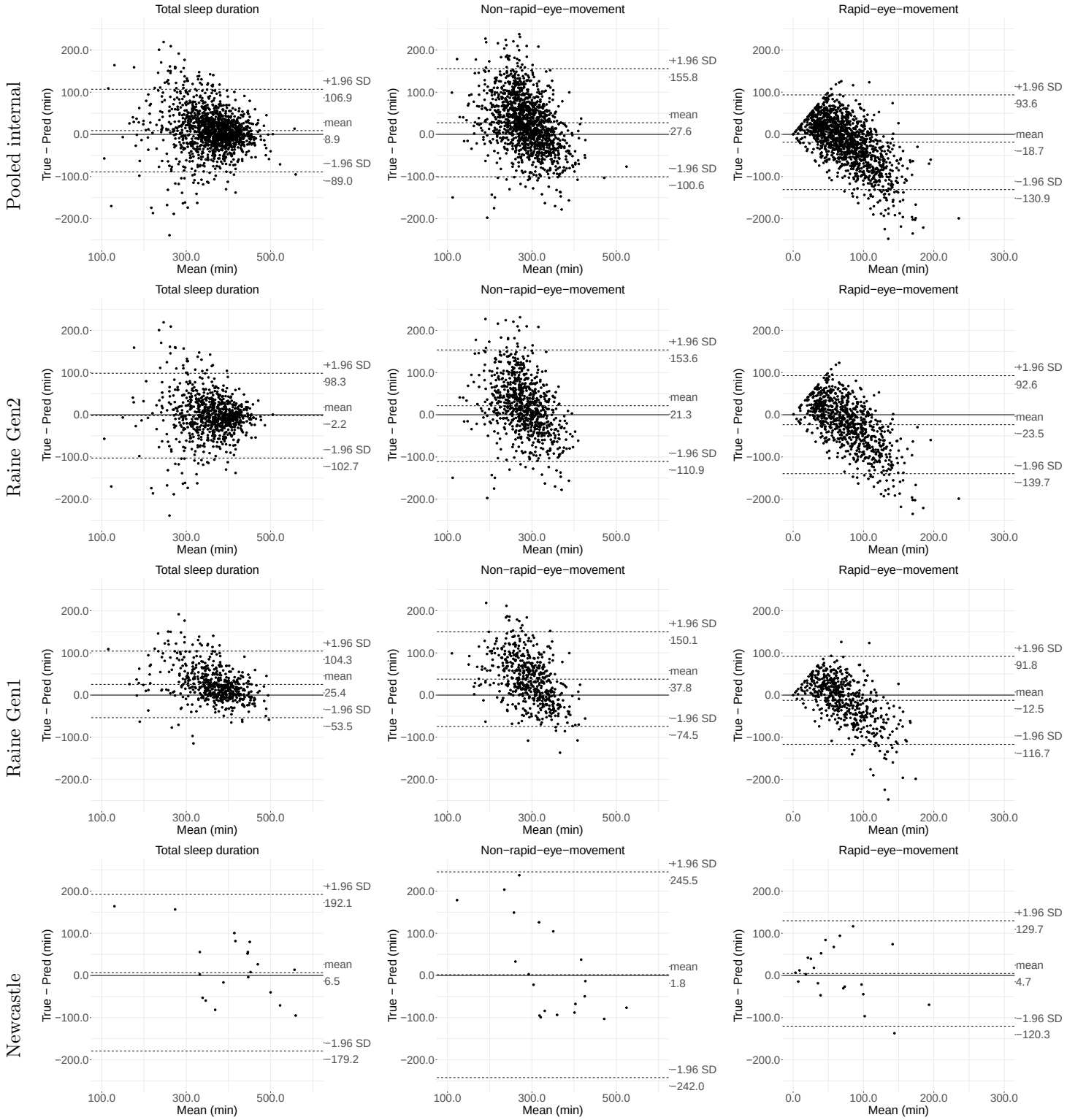


Figure C.1: Agreement assessment via Bland-Altman plots for internal validation: total sleep duration (TSD), non-rapid-eye-movement sleep (NREM), and rapid-eye-movement sleep (REM).

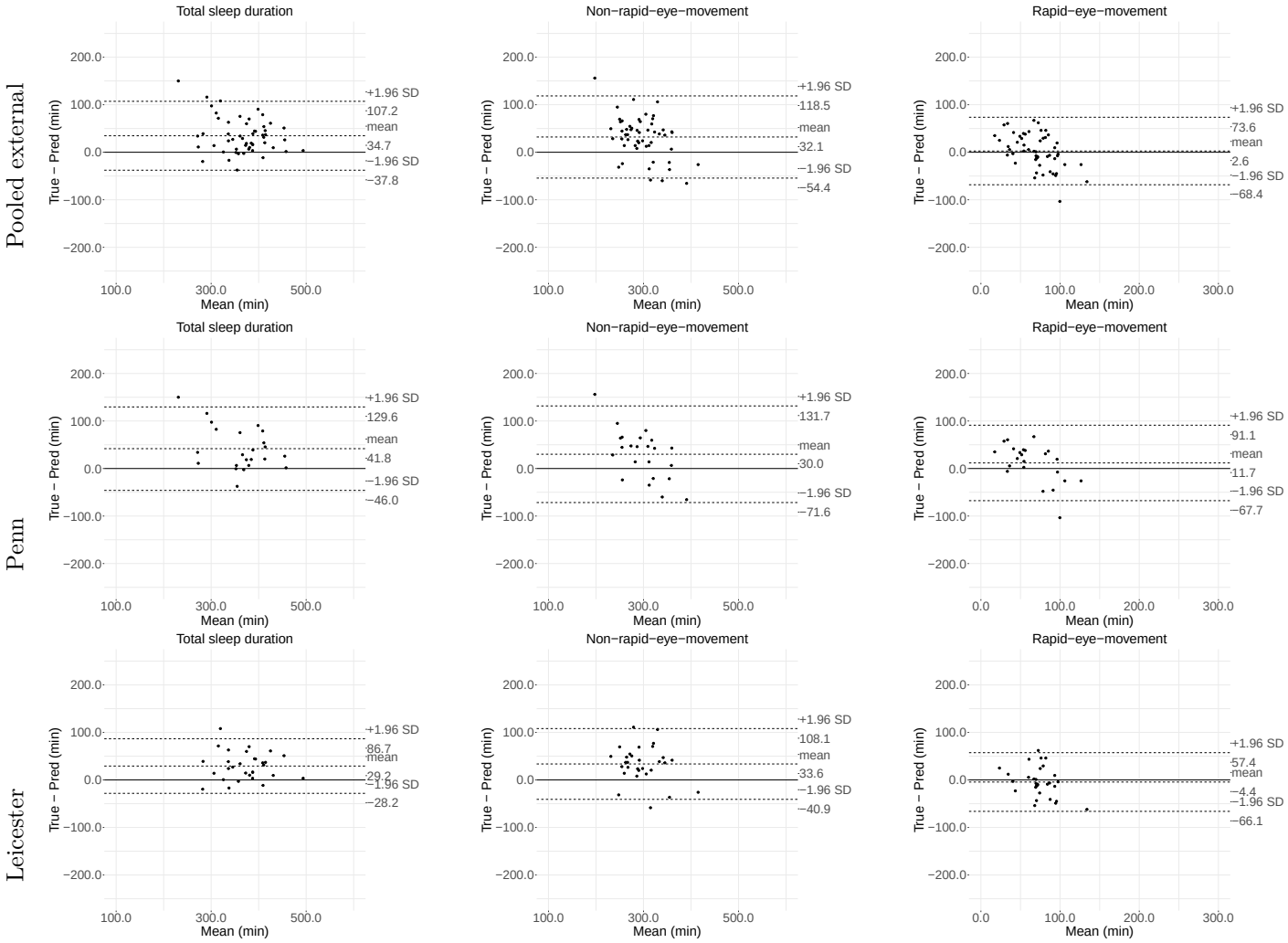


Figure C.2: Agreement assessment via Bland-Altman plots for external validation: total sleep duration, wake after sleep onset (WASO), non-rapid-eye-movement sleep (NREM), and rapid-eye-movement sleep (REM).

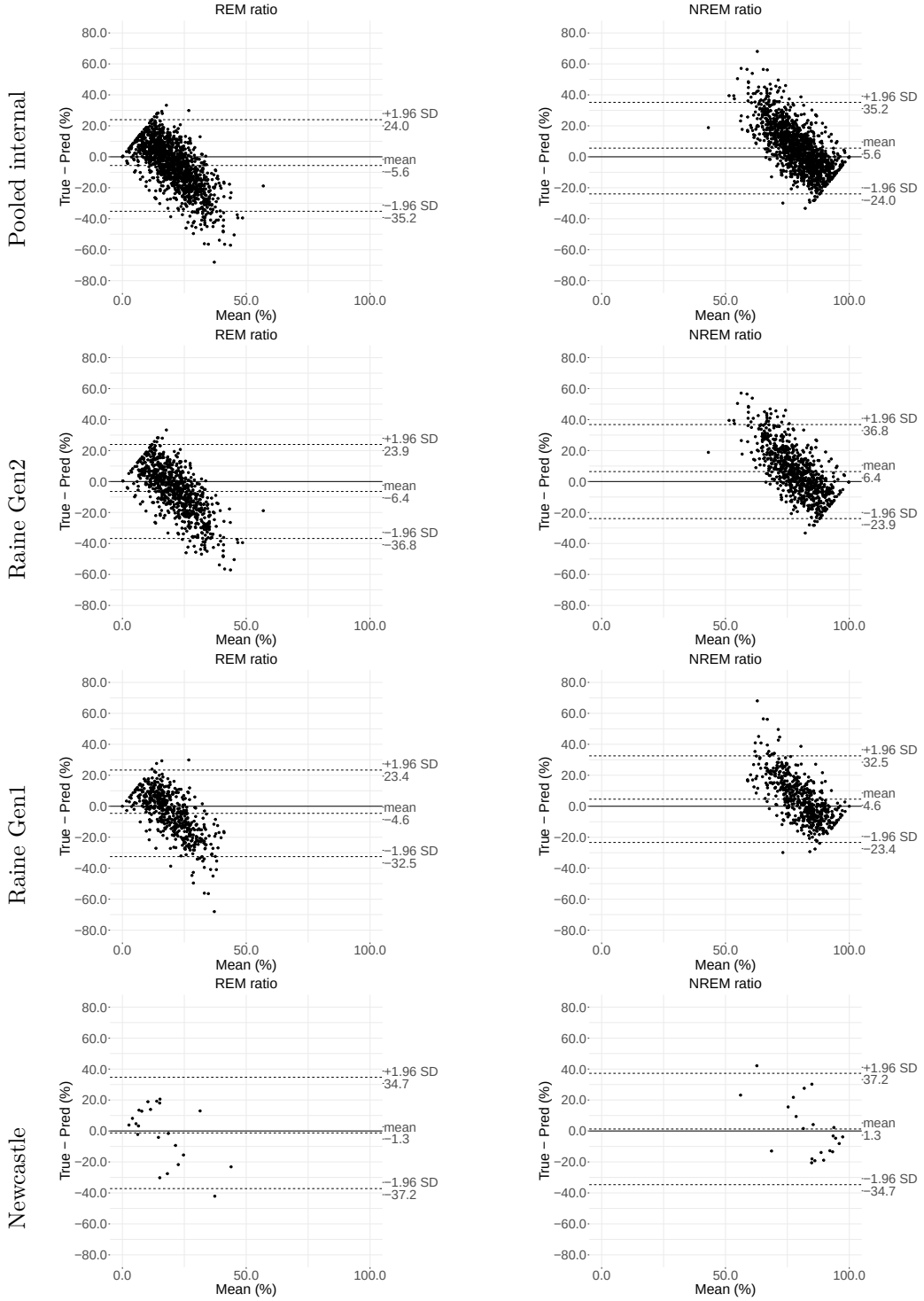


Figure C.3: Agreement assessment via Bland-Altman plots for internal validation: non-rapid-eye-movement sleep (NREM) ratio, and rapid-eye-movement sleep (REM) ratio.

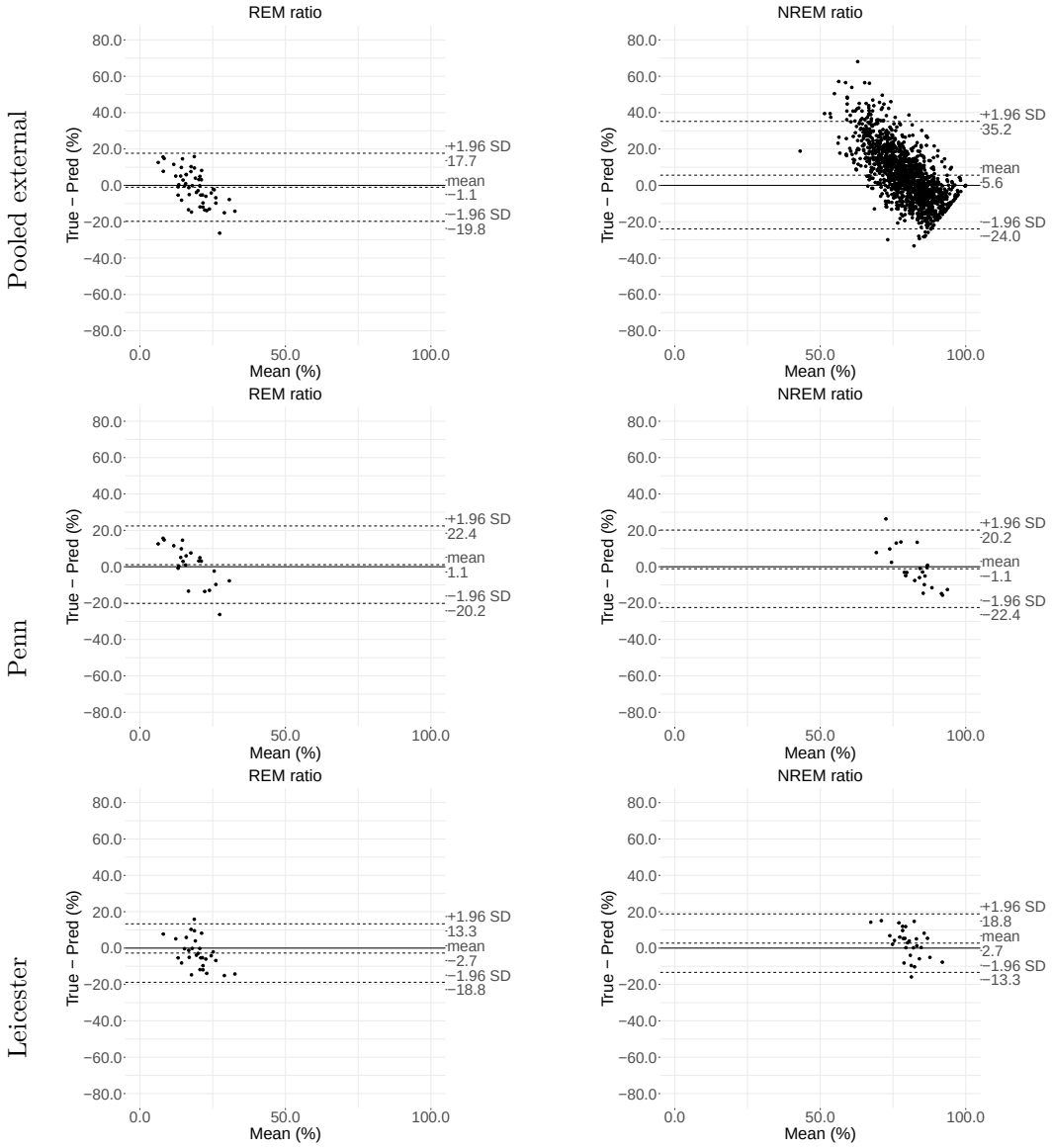


Figure C.4: Agreement assessment via Bland-Altman plots for external validation: non-rapid-eye-movement sleep (NREM) ratio, and rapid-eye-movement sleep (REM) ratio.

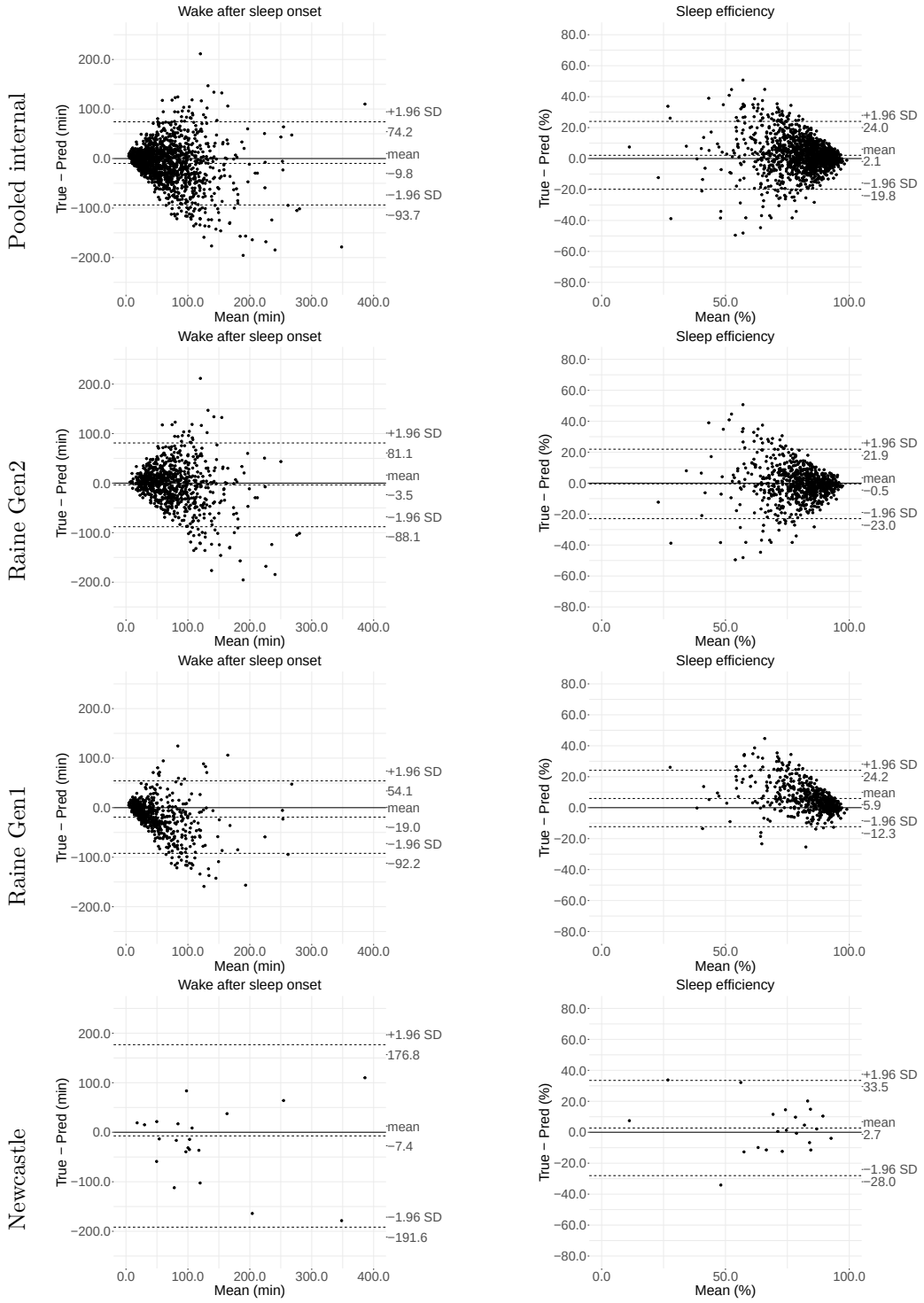


Figure C.5: Agreement assessment via Bland-Altman plots for internal validation: wake after sleep onset (WASO), and sleep efficiency (SE).

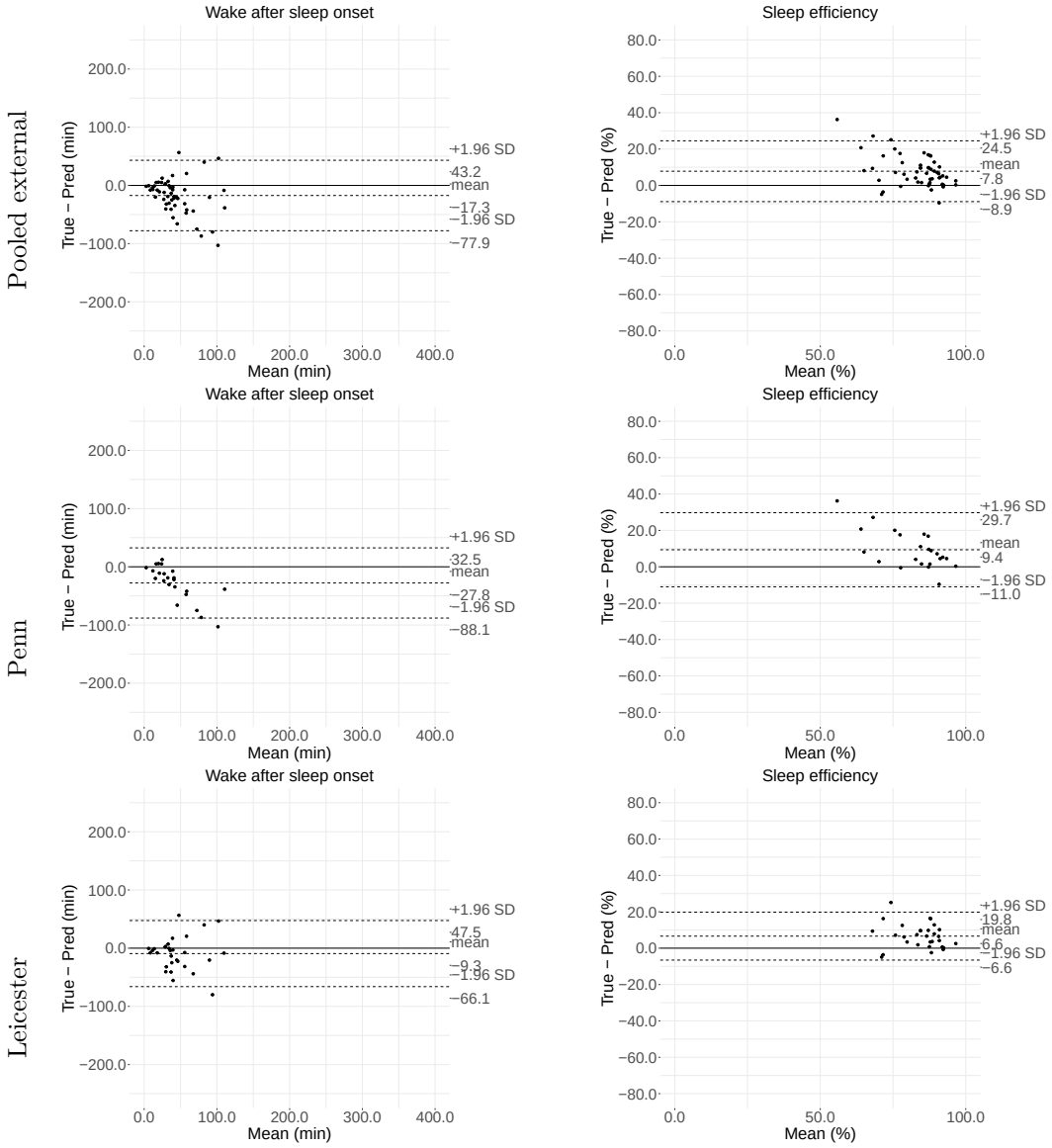


Figure C.6: Agreement assessment via Bland-Altman plots for internal validation: wake after sleep onset (WASO), and sleep efficiency (SE).

Bibliography

[1] **Hang Yuan***, Shing Chan*, Andrew P Creagh, Catherine Tong, Aidan Acquah, David A Clifton, and Aiden Doherty. "Self-supervised Learning for Human Activity Recognition Using 700,000 Person-days of Wearable Data". In: *npj Digital Medicine* (2024).

[2] **Hang Yuan**, Tatiana Plekhanova, Rosemary Walmsley, Amy C. Reynolds, Kathleen J. Maddison, Maja Bucan, Philip Gehrman, Alex Rowlands, David W. Ray, Derrick Bennett, Joanne McVeigh, Leon Straker, Peter Eastwood, Simon D. Kyle, and Aiden Doherty. "Self-supervised learning of accelerometer data provides new insights for sleep and its association with mortality". In: *npj Digital Medicine* (2024).

[3] **Hang Yuan**, Elizabeth A Hill, Simon D Kyle, and Aiden Doherty. "A systematic review of the performance of actigraphy in measuring sleep stages". In: *Journal of Sleep Research* (2023).

[4] Organizers of Queer in AI, Anaelia Ovalle, Arjun Subramonian, Ashwin Singh, Claas Voelcker, Danica J. Sutherland, Davide Locatelli, Eva Breznik, Filip Klubicka, **Hang Yuan**, J Hetvi, Huan Zhang, Jaidev Shriram, Kruno Lehman, Luca Soldaini, Maarten Sap, Marc Peter Deisenroth, Maria Leonor Pacheco, Maria Ryskina, Martin Mundt, Milind Agarwal, Nyx McLean, Pan Xu, A Pranav, Raj Korpan, Ruchira Ray, Sarah Mathew, Sarthak Arora, St John, Tanvi Anand, Vishakha Agrawal, William Agnew, Yanan Long, Zijie J. Wang, Zeerak Talat, Avijit Ghosh, Nathaniel Dennler, Michael Noseworthy, Sharvani Jha, Emily Baylor, Aditya Joshi, Natalia Y. Bilenko, Andrew McNamara, Raphael Gontijo-Lopes, Alex Markham, Evyn Dǒng, Jackie Kay, Manu Saraswat, Nikhil Vytla, and Luke Stark. "Queer In AI: A Case Study in Community-Led Participatory AI". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (2023).

[5] Andrew P Creagh, Valentin Hamy, Hang Yuan, Gert Mertes, Ryan Tomlinson, Wen-Hung Chen, Rachel Williams, Christopher Llop, Christopher Yee, Mei Sheng Duh, et al. "Digital health technologies and machine learning augment patient reported outcomes to remotely characterise rheumatoid arthritis". In: *npj Digital Medicine* 7.1 (2024), p. 33.

[6] **Hang Yuan***, Claudia Vanea*, Federica Lucivero, and Nina Hallowell. "Training Ethically Responsible AI Researchers: a Case Study". In: *Navigating the Broader Impacts of AI Research Workshop at NeurIPS* (2020).

[7] Federica Lucivero, Gabrielle Samuel, Gordon Blair, Sarah J Darby, Tina Fawcett, Mike Hazas, Carolyn Ten Holter, Marina Jirotka, Michael Parker, Helena Webb, and **Hang Yuan**. "Data-driven unsustainability? An interdisciplinary perspective on governing the environmental impacts of a data-driven society". In: *An Interdisciplinary Perspective on Governing the Environmental Impacts of a Data-Driven Society (June 19, 2020)* (2020).

[8] Mary A Carskadon, William C Dement, et al. "Normal human sleep: an overview". In: *Principles and practice of sleep medicine* 4.1 (2005), pp. 13–23.

[9] William C Dement. "History of sleep medicine". In: *Neurologic clinics* 23.4 (2005), pp. 945–965.

[10] Richard B Berry, Rita Brooks, Charlene Gamaldo, Susan M Harding, Robin M Lloyd, Stuart F Quan, Matthew T Troester, and Bradley V Vaughn. *AASM scoring manual updates for 2017 (version 2.4)*. 2017.

[11] I Feinberg and TC Floyd. "Systematic trends across the night in human sleep cycles". In: *Psychophysiology* 16.3 (1979), pp. 283–291.

[12] Sudhansu Chokroverty. "Overview of sleep & sleep disorders". In: *Indian Journal of Medical Research* 131.2 (2010), pp. 126–140.

[13] Giulio Tononi and Chiara Cirelli. "Sleep and synaptic homeostasis: a hypothesis". In: *Brain research bulletin* 62.2 (2003), pp. 143–150.

[14] Samson Z Assefa, Montserrat Diaz-Abad, Emerson M Wickwire, and Steven M Scharf. "The Functions of Sleep." In: *AIMS Neuroscience* 2.3 (2015).

[15] Aakash K Patel, Vamsi Reddy, Karlie R Shumway, and John F Araujo. "Physiology, sleep stages". In: *StatPearls [Internet]*. StatPearls Publishing, 2022.

[16] Masaki Nishida and Matthew P Walker. "Daytime naps, motor memory consolidation and regionally specific sleep spindles". In: *PloS one* 2.4 (2007), e341.

[17] Eve Van Cauter and Laurence Plat. "Physiology of growth hormone secretion during sleep". In: *The Journal of pediatrics* 128.5 (1996), S32–S37.

[18] Maurice M Ohayon, Mary A Carskadon, Christian Guilleminault, and Michael V Vitiello. "Meta-analysis of quantitative sleep parameters from childhood to old age in healthy individuals: developing normative sleep values across the human lifespan". In: *Sleep* 27.7 (2004), pp. 1255–1273.

[19] Chiara Baglioni, Svetoslava Nanovska, Wolfram Regen, Kai Spiegelhalder, Bernd Feige, Christoph Nissen, Charles F Reynolds III, and Dieter Riemann. "Sleep and mental disorders: A meta-analysis of polysomnographic research." In: *Psychological bulletin* 142.9 (2016), p. 969.

[20] Ye Zhang, Rong Ren, Linghui Yang, Haipeng Zhang, Yuan Shi, Larry D Sanford, and Xiangdong Tang. "Polysomnographic nighttime features of narcolepsy: a systematic review and meta-analysis". In: *Sleep Medicine Reviews* 58 (2021), p. 101488.

[21] Michael A Grandner and Sean PA Drummond. "Who are the long sleepers? Towards an understanding of the mortality relationship". In: *Sleep medicine reviews* 11.5 (2007), pp. 341–360.

[22] Michael A Grandner, Lauren Hale, Melisa Moore, and Nirav P Patel. "Mortality associated with short sleep duration: the evidence, the possible mechanisms, and the future". In: *Sleep medicine reviews* 14.3 (2010), pp. 191–203.

[23] Baixin Chen, Alexandros N Vgontzas, and Yun Li. "Good sleep and health". In: *The Lancet Healthy Longevity* 4.7 (2023), e295–e296.

[24] Faith S Luyster, Patrick J Strollo Jr, Phyllis C Zee, and James K Walsh. "Sleep: a health imperative". In: *Sleep* 35.6 (2012), pp. 727–734.

[25] Uthman Albakri, Elizabeth Drotos, and Ree Meertens. "Sleep health promotion interventions and their effectiveness: an umbrella review". In: *International Journal of Environmental Research and Public Health* 18.11 (2021), p. 5533.

[26] URL: `https://www.healthypeople.gov/2020/topics-objectives`.

[27] *Waking up to the Health Benefits of Sleep*. 2020. URL: `https://www.rsph.org.uk/%20resourceLibrary/waking-up-to-the-health-benefits-of-sleep.html`.

[28] Andressa Alves da Silva, Renato Gorga Bandeira de Mello, Camila Wohlgemuth Schaan, Flávio D Fuchs, Susan Redline, and Sandra C Fuchs. "Sleep duration and mortality in the elderly: a systematic review with meta-analysis". In: *BMJ open* 6.2 (2016), e008119.

[29] Tong-Zu Liu, Chang Xu, Matteo Rota, Hui Cai, Chao Zhang, Ming-Jun Shi, Rui-Xia Yuan, Hong Weng, Xiang-Yu Meng, Joey SW Kwong, et al. "Sleep duration and risk of all-cause mortality: a flexible, non-linear, meta-regression of 40 prospective cohort studies". In: *Sleep medicine reviews* 32 (2017), pp. 28–36.

[30] Marieke P Hoevenaar-Blom, Annemieke MW Spijkerman, Daan Kromhout, Julia F van den Berg, and WM Verschuren. "Sleep duration and sleep quality in relation to 12-year cardiovascular disease incidence: the MORGEN study". In: *Sleep* 34.11 (2011), pp. 1487–1492.

[31] Shahram Nikbakhtian, Angus B Reed, Bernard Dillon Obika, Davide Morelli, Adam C Cunningham, Mert Aral, and David Plans. "Accelerometer-derived sleep onset timing and cardiovascular disease incidence: a UK Biobank cohort study". In: *European Heart Journal-Digital Health* 2.4 (2021), pp. 658–666.

[32] Jennifer A Hutcheon, Arnaud Chiolero, and James A Hanley. "Random measurement error and regression dilution bias". In: *Bmj* 340 (2010).

[33] 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". In: *NPJ digital medicine* 3.1 (2020), p. 42.

[34] Jane E Ferrie, Meena Kumari, Paula Salo, Archana Singh-Manoux, and Mika Kivimaki. *Sleep epidemiology—a rapidly growing field*. 2011.

[35] Daniel J Buysse, Charles F Reynolds III, Timothy H Monk, Susan R Berman, and David J Kupfer. "The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research". In: *Psychiatry research* 28.2 (1989), pp. 193–213.

[36] Murray W Johns. "A new method for measuring daytime sleepiness: the Epworth sleepiness scale". In: *sleep* 14.6 (1991), pp. 540–545.

[37] Timothy H Monk, CHARLES F REYNOLDS III, David J Kupfer, Daniel J Buysse, Patricia A Coble, Amy J Hayes, Mary Ann Machen, Susan R Petrie, and Angela M Ritenour. "The Pittsburgh sleep diary". In: *Journal of sleep research* 3.2 (1994), pp. 111–120.

[38] Colleen E Carney, Daniel J Buysse, Sonia Ancoli-Israel, Jack D Edinger, Andrew D Krystal, Kenneth L Lichstein, and Charles M Morin. "The consensus sleep diary: standardizing prospective sleep self-monitoring". In: *Sleep* 35.2 (2012), pp. 287–302.

[39] Vanessa Ibáñez, Josep Silva, and Omar Cauli. "A survey on sleep questionnaires and diaries". In: *Sleep medicine* 42 (2018), pp. 90–96.

[40] 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". In: *PLoS Medicine* 12.3 (2015), e1001779.

[41] Zhengming Chen, Junshi Chen, Rory Collins, Yu Guo, Richard Peto, Fan Wu, and Liming Li. "China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up". In: *International journal of epidemiology* 40.6 (2011), pp. 1652–1666.

[42] H Matthew Lehrer, Zhigang Yao, Robert T Krafty, Marissa A Evans, Daniel J Buysse, Howard M Kravitz, Karen A Matthews, Ellen B Gold, Sioban D Harlow, Laura B Samuelsson, et al. "Comparing polysomnography, actigraphy, and sleep diary in the home environment: The Study of Women's Health Across the Nation (SWAN) Sleep Study". In: *Sleep Advances* 3.1 (2022), zpac001.

[43] Fee Benz, Dieter Riemann, Katharina Domschke, Kai Spiegelhalder, Anna F Johann, Nathaniel S Marshall, and Bernd Feige. "How many hours do you sleep? A comparison of subjective and objective sleep duration measures in a sample of insomnia patients and good sleepers". In: *Journal of Sleep Research* 32.2 (2023), e13802.

[44] Diego de Alcantara Borba, Raquel Sousa Reis, Pedro Henrique Tadeu de Melo Lima, Lucas Alves Facundo, Fernanda Veruska Narciso, Andressa Silva, and Marco Túlio de Mello. "How many days are needed for a reliable assessment by the Sleep Diary?" In: *Sleep Science* 13.1 (2020), p. 49.

[45] Allan Rechtschaffen. "A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects". In: *Public Health Service* (1968), pp. 1–55.

[46] Johan Newell, Olivier Mairesse, Paul Verbanck, and Daniel Neu. "Is a one-night stay in the lab really enough to conclude? First-night effect and night-to-night variability in polysomnographic recordings among different clinical population samples". In: *Psychiatry Research* 200.2-3 (2012), pp. 795–801.

[47] Clete A Kushida, Arthur Chang, Chirag Gadkary, Christian Guilleminault, Oscar Carrillo, and William C Dement. "Comparison of actigraphic, polysomnographic, and subjective assessment of sleep parameters in sleep-disordered patients". In: *Sleep Medicine* 2.5 (2001), pp. 389–396.

[48] Kevin So, PAT Buckley, T Michael Adamson, and Rosemary SC Horne. "Actigraphy correctly predicts sleep behavior in infants who are younger than six months, when compared with polysomnography". In: *Pediatric Research* 58.4 (2005), pp. 761–765.

[49] Catherine McCall and W Vaughn McCall. "Comparison of actigraphy with polysomnography and sleep logs in depressed insomniacs". In: *Journal of Sleep Research* 21.1 (2012), pp. 122–127.

[50] Elicia Toon, Margot J Davey, Samantha L Hollis, Gillian M Nixon, Rosemary SC Horne, and Sarah N Biggs. "Comparison of commercial wrist-based and smartphone accelerometers, actigraphy, and PSG in a clinical cohort of children and adolescents". In: *Journal of Clinical Sleep Medicine* 12.3 (2016), pp. 343–350.

[51] John B Webster, Daniel F Kripke, Sam Messin, Daniel J Mullaney, and Grant Wyborney. "An activity-based sleep monitor system for ambulatory use". In: *Sleep* 5.4 (1982), pp. 389–399.

[52] Roger J Cole, Daniel F Kripke, William Gruen, Daniel J Mullaney, and J Christian Gillin. "Automatic sleep/wake identification from wrist activity". In: *Sleep* 15.5 (1992), pp. 461–469.

[53] Nigel R Oakley. "Validation with polysomnography of the Sleepwatch sleep/wake scoring algorithm used by the Actiwatch activity monitoring system. Mini Mitter Co". In: *Sleep* 2 (1997), pp. 0–140.

[54] Alex V Rowlands, Evgeny M Mirkes, TOM Yates, Stacey Clemes, Melanie Davies, Kamlesh Khunti, and Charlotte L Edwardson. "Accelerometer-assessed physical activity in epidemiology: are monitors equivalent?" In: (2017).

[55] Vincent T van Hees, Kathrin Thaler-Kall, Klaus-Hendrik Wolf, Jan C Brønd, Alberto Bonomi, Mareike Schulze, Matthäus Vigl, Bente Morseth, Laila Arnesdatter Hopstock, Lukas Gorzelniak, et al. "Challenges and opportunities for harmonizing research methodology: Raw accelerometry". In: *Methods of information in medicine* 55.06 (2016), pp. 525–532.

[56] Ali Neishabouri, Joe Nguyen, John Samuelsson, Tyler Guthrie, Matt Biggs, Jeremy Wyatt, Doug Cross, Marta Karas, Jairo H Migueles, Sheraz Khan, et al. "Quantification of acceleration as activity counts in ActiGraph wearable". In: *Scientific Reports* 12.1 (2022), p. 11958.

[57] 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". In: *PloS One* 12.2 (2017), e0169649.

[58] Andrew Leroux, Junrui Di, Ekaterina Smirnova, Elizabeth J Mcguffey, Quy Cao, Elham Bayatmokhtari, Lucia Tabacu, Vadim Zipunnikov, Jacek K Urbanek, and Ciprian Crainiceanu. "Organizing and analyzing the activity data in NHANES". In: *Statistics in biosciences* 11 (2019), pp. 262–287.

[59] Samuel E Jones, Jacqueline M Lane, Andrew R Wood, Vincent T van Hees, Jessica Tyrrell, Robin N Beaumont, Aaron R Jeffries, Hassan S Dashti, Melvyn Hillsdon, Katherine S Ruth, et al. "Genome-wide association analyses of chronotype in 697,828 individuals provides insights into circadian rhythms". In: *Nature communications* 10.1 (2019), p. 343.

[60] Aiden Doherty, Karl Smith-Byrne, Teresa Ferreira, Michael V Holmes, Chris Holmes, Sara L Pulit, and Cecilia M Lindgren. "GWAS identifies 14 loci for device-measured physical activity and sleep duration". In: *Nature Communications* 9.1 (2018), pp. 1–8.

[61] Rosemary Walmsley, Shing Chan, Karl Smith-Byrne, Rema Ramakrishnan, Mark Woodward, Kazem Rahimi, Terence Dwyer, Derrick Bennett, and Aiden Doherty. "Reallocation of time between device-measured movement behaviours and risk of incident cardiovascular disease". In: *British Journal of Sports Medicine* 56.18 (2022), pp. 1008–1017.

[62] Sebastien Chastin, Duncan McGregor, Javier Palarea-Albaladejo, Keith M Diaz, Maria Hagströmer, Pedro Curi Hallal, Vincent T Van Hees, Steven Hooker, Virginia J Howard, I-Min Lee, et al. "Joint association between accelerometry-measured daily combination of time spent in physical activity, sedentary behaviour and sleep and all-cause mortality: a pooled analysis of six prospective cohorts using compositional analysis". In: *British journal of sports medicine* 55.22 (2021), pp. 1277–1285.

[63] Mustafa Radha, Pedro Fonseca, Arnaud Moreau, Marco Ross, Andreas Cerny, Peter Anderer, Xi Long, and Ronald M Aarts. "A deep transfer learning approach for wearable sleep stage classification with photoplethysmography". In: *NPJ digital medicine* 4.1 (2021), p. 135.

[64] Olivia Walch, Yitong Huang, Daniel Forger, and Cathy Goldstein. "Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device". In: *Sleep* 42.12 (2019), zsz180.

[65] Aaqib Saeed, Tanir Ozcelebi, and Johan Lukkien. "Multi-task self-supervised learning for human activity detection". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3.2 (2019), pp. 1–30.

[66] Yash Jain, Chi Ian Tang, Chulhong Min, Fahim Kawsar, and Akhil Mathur. "Collossl: Collaborative self-supervised learning for human activity recognition". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6.1 (2022), pp. 1–28.

[67] Harish Haresamudram, Irfan Essa, and Thomas Plötz. "Assessing the state of self-supervised human activity recognition using wearables". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6.3 (2022), pp. 1–47.

[68] Vincent T Van Hees, Séverine Sabia, Kirstie N Anderson, Sarah J Denton, James Oliver, Michael Catt, Jessica G Abell, Mika Kivimäki, Michael I Trenell, and Archana Singh-Manoux. "A novel, open access method to assess sleep duration using a wrist-worn accelerometer". In: *PloS one* 10.11 (2015), e0142533.

[69] Vincent Theodoor van Hees, Severine Sabia, Samuel E Jones, Andrew R Wood, Kirstie N Anderson, M Kivimäki, Timothy M Frayling, Allan I Pack, Maja Bucan, MI Trenell, et al. "Estimating sleep parameters using an accelerometer without sleep diary". In: *Scientific Reports* 8.1 (2018), pp. 1–11.

[70] Matthew R Patterson, Adonay AS Nunes, Dawid Gerstel, Rakesh Pilkar, Tyler Guthrie, Ali Neishabouri, and Christine C Guo. "40 years of actigraphy in sleep medicine and current state of the art algorithms". In: *NPJ Digital Medicine* 6.1 (2023), p. 51.

[71] Michelle L Trevenen, Berwin A Turlach, Peter R Eastwood, Leon M Straker, and Kevin Murray. "Using hidden Markov models with raw, triaxial wrist accelerometry data to determine sleep stages". In: *Australian & New Zealand Journal of Statistics* 61.3 (2019), pp. 273–298.

[72] Kalaivani Sundararajan, Sonja Georgievska, Bart H. W. te Lindert, Philip R. Gehrman, Jennifer Ramautar, Diego R. Mazzotti, Séverine Sabia, Michael N. Weedon, Eus J. W. van Someren, Lars Ridder, Jian Wang, and Vincent T. van Hees. "Sleep classification from wrist-worn accelerometer data using random forests". In: *Scientific Reports* 11.1 (Dec. 2021), p. 24. URL: `http://www.nature.com/articles/s41598-020-79217-x`.

[73] Ulysses J Magalang, Ning-Hung Chen, Peter A Cistulli, Annette C Fedson, Thorarinn Gıslason, David Hillman, Thomas Penzel, Renaud Tamisier, Sergio Tufik, Gary Phillips, et al. "Agreement in the scoring of respiratory events and sleep among international sleep centers". In: *Sleep* 36.4 (2013), pp. 591–596.

[74] J Richard Landis and Gary G Koch. "The measurement of observer agreement for categorical data". In: *biometrics* (1977), pp. 159–174.

[75] Avi Sadeh. "The role and validity of actigraphy in sleep medicine: an update". In: *Sleep medicine reviews* 15.4 (2011), pp. 259–267.

[76] Michael T Smith, Christina S McCrae, Joseph Cheung, Jennifer L Martin, Christopher G Harrod, Jonathan L Heald, and Kelly A Carden. "Use of actigraphy for the evaluation of sleep disorders and circadian rhythm sleep-wake disorders: an American Academy of Sleep Medicine systematic review, meta-analysis, and GRADE assessment". In: *Journal of Clinical Sleep Medicine* 14.7 (2018), pp. 1209–1230.

[77] Kristin E Schwab, Bonnie Ronish, Dale M Needham, An Q To, Jennifer L Martin, and Biren B Kamdar. "Actigraphy to evaluate sleep in the intensive care unit. A systematic review". In: *Annals of the American Thoracic Society* 15.9 (2018), pp. 1075–1082.

[78] Samantha Conley, Andrea Knies, Janene Batten, Garrett Ash, Brienne Miner, Youri Hwang, Sangchoon Jeon, and Nancy S Redeker. "Agreement between actigraphic and polysomnographic measures of sleep in adults with and without chronic conditions: a systematic review and meta-analysis". In: *Sleep Medicine Reviews* 46 (2019), pp. 151–160.

[79] Alexander TM Van de Water, Alison Holmes, and Deirdre A Hurley. "Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography–a systematic review". In: *Journal of sleep research* 20.1pt2 (2011), pp. 183–200.

[80] Ping-Ru T Ko, Julie A Kientz, Eun Kyoung Choe, Matthew Kay, Carol A Landis, and Nathaniel F Watson. "Consumer sleep technologies: a review of the landscape". In: *Journal of clinical sleep medicine* 11.12 (2015), pp. 1455–1461.

[81] Evan D Chinoy, Joseph A Cuellar, Kirbie E Huwa, Jason T Jameson, Catherine H Watson, Sara C Bessman, Dale A Hirsch, Adam D Cooper, Sean PA Drummond, and Rachel R Markwald. "Performance of seven consumer sleep-tracking devices compared with polysomnography". In: *Sleep* 44.5 (2021), zsaa291.

[82] Richard B Berry, Stuart F Quan, Alexandre R Abreu, Marietta L Bibbs, Lourdes DelRosso, Susan M Harding, Molly-Min Mao, David T Plante, Mark R Pressman, Matthew M Troester, et al. *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*. American Academy of Sleep Medicine, 2020.

[83] Syed Anas Imtiaz. "A systematic review of sensing technologies for wearable sleep staging". In: *Sensors* 21.5 (2021), p. 1562.

[84] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews". In: *Systematic reviews* 10.1 (2021), pp. 1–11.

[85] Heidi Danker-Hopfe, Dieter Kunz, Georg Gruber, Gerhard Klösch, José L Lorenzo, Sari-Leena Himanen, Bob Kemp, Thomas Penzel, Joachim Röschke, Hans Dorn, et al. "Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders". In: *Journal of sleep research* 13.1 (2004), pp. 63–69.

[86] Penny F Whiting, Anne WS Rutjes, Marie E Westwood, Susan Mallett, Jonathan J Deeks, Johannes B Reitsma, Mariska MG Leeflang, Jonathan AC Sterne, and Patrick MM Bossuyt. "QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies". In: *Annals of internal medicine* 155.8 (2011), pp. 529–536.

[87] Quyen NT Nguyen, Toan Le, Quyen BT Huynh, Arveity Setty, Toi V Vo, and Trung Q Le. "Validation Framework for Sleep Stage Scoring in Wearable Sleep Trackers and Monitors with Polysomnography Ground Truth". In: *Clocks & sleep* 3.2 (2021), pp. 274–288.

[88] Janna Mantua, Nickolas Gravel, and Rebecca Spencer. "Reliability of sleep measures from four personal health monitoring devices compared to research-based actigraphy and polysomnography". In: *Sensors* 16.5 (2016), p. 646.

[89] Jaime K Devine, Evan D Chinoy, Rachel R Markwald, Lindsay P Schwartz, and Steven R Hursh. "Validation of Zulu watch against polysomnography and actigraphy for on-wrist sleep-wake determination and sleep-depth estimation". In: *Sensors* 21.1 (2020), p. 76.

[90] Alexander J Boe, Lori L McGee Koch, Megan K O'Brien, Nicholas Shawen, John A Rogers, Richard L Lieber, Kathryn J Reid, Phyllis C Zee, and Arun Jayaraman. "Automating sleep stage classification using wireless, wearable sensors". In: *NPJ Digital Medicine* 2.1 (2019), pp. 1–9.

[91] Kalaivani Sundararajan, Sonja Georgievska, Bart HW Te Lindert, Philip R Gehrman, Jennifer Ramautar, Diego R Mazzotti, Séverine Sabia, Michael N Weedon, Eus JW van Someren, Lars Ridder, et al. "Sleep classification from wrist-worn accelerometer data using random forests". In: *Scientific Reports* 11.1 (2021), pp. 1–10.

[92] Richard B Berry, Rita Brooks, Charlene E Gamaldo, Susan M Harding, C Marcus, Bradley V Vaughn, et al. "The AASM manual for the scoring of sleep and associated events". In: *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine* 176 (2012), p. 2012.

[93] et al. Berry RB. *the American Academy of Sleep Medicine.The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications. Version 2.3.0 ed.* Vol. 1. American academy of sleep medicine Westchester, IL, 2016.

[94] Conrad Iber, Sonia Ancoli-Israel, Andrew L Chesson, Stuart F Quan, et al. *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications.* Vol. 1. American academy of sleep medicine Westchester, IL, 2007.

[95] Massimiliano de Zambotti, Luca Menghini, Michael A Grandner, Susan Redline, Ying Zhang, Meredith L Wallace, and Orfeu M Buxton. "Rigorous performance evaluation (previously,"validation") for informed use of new technologies for sleep health measurement". In: *Sleep health* 8.3 (2022), pp. 263–269.

[96] Avi Sadeh, M Sharkey, and Mary A Carskadon. "Activity-based sleep-wake identification: an empirical test of methodological issues". In: *Sleep* 17.3 (1994), pp. 201–207.

[97] Sarah Herscovici, Avivit Pe'er, Surik Papyan, and Peretz Lavie. "Detecting REM sleep from the finger: an automatic REM sleep algorithm based on peripheral arterial tone (PAT) and actigraphy". In: *Physiological measurement* 28.2 PG - 129-40 (2007), pp. 129–140.

[98] Alexandre Domingues, Teresa Paiva, and J Miguel Sanches. "Hypnogram and sleep parameter computation from activity and cardiovascular data". In: *IEEE transactions on bio-medical engineering* 61.6 PG - 1711-9 (2014), pp. 1711–1719.

[99] Shahab Haghayegh, Sepideh Khoshnevis, Michael H Smolensky, Kenneth R Diller, and Richard J Castriotta. "Deep Neural Network Sleep Scoring Using Combined Motion and Heart Rate Variability Data". In: *Sensors (Basel, Switzerland)* 21.1 PG - (2020).

[100] JP Newnham and SE Reid. "A randomised controlled trial of doppler waveform analysis in obstetrics". In: *National Health and Medical Research Council grant* 880441 (1987).

[101] Leon Straker, Jenny Mountain, Angela Jacques, Scott White, Anne Smith, Louis Landau, Fiona Stanley, John Newnham, Craig Pennell, and Peter Eastwood. "Cohort profile: the Western Australian pregnancy cohort (Raine) study–generation 2". In: *International journal of epidemiology* 46.5 (2017), 1384–1385j.

[102] Manon L Dontje, Peter Eastwood, and Leon Straker. "Western Australian pregnancy cohort (Raine) study: generation 1". In: *BMJ open* 9.5 (2019), e026276.

[103] Richard B Berry, Rohit Budhiraja, Daniel J Gottlieb, David Gozal, Conrad Iber, Vishesh K Kapur, Carole L Marcus, Reena Mehra, Sairam Parthasarathy, Stuart F Quan, et al. "Rules for scoring respiratory events in sleep: update of the 2007 AASM manual for the scoring of sleep and associated events: deliberations of the sleep apnea definitions task force of the American Academy of Sleep Medicine". In: *Journal of Clinical Sleep Medicine* 8.5 (2012), pp. 597–619.

[104] ActiGraph. *ActiGraph wGT3X-BT*. December, 5, 2023. 2023. URL: `https://theactigraph.com/actigraph-wgt3x-bt`.

[105] Vincent van Hees, Sarah Charman, and Kirstie Anderson. *Newcastle polysomnography and accelerometer data*. Version 1.0. Zenodo, Jan. 2018. URL: `https://doi.org/10.5281/zenodo.1160410`.

[106] Tatiana Plekhanova, Alex V Rowlands, Melanie J Davies, Andrew P Hall, Tom Yates, and Charlotte L Edwardson. "Validation of an automated sleep detection algorithm using data from multiple accelerometer brands". In: *Journal of Sleep Research* (2022).

[107] Enda M Byrne, Philip R Gehrman, Maciej Trzaskowski, Henning Tiemeier, and Allan I Pack. "Genetic correlation analysis suggests association between increased self-reported sleep duration in adults and schizophrenia and type 2 diabetes". In: *Sleep* 39.10 (2016), pp. 1853–1857.

[108] Machiko Katori, Shoi Shi, Koji L Ode, Yasuhiro Tomita, and Hiroki R Ueda. "The 103,200-arm acceleration dataset in the UK Biobank revealed a landscape of human sleep phenotypes". In: *Proceedings of the National Academy of Sciences* 119.12 (2022), e2116729119.

[109] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. "The UK Biobank resource with deep phenotyping and genomic data". In: *Nature* 562.7726 (2018), pp. 203–209.

[110] Naomi Allen, Cathie Sudlow, Paul Downey, Tim Peakman, John Danesh, Paul Elliott, John Gallacher, Jane Green, Paul Matthews, Jill Pell, Tim Sprosen, and Rory Collins. "UK Biobank: Current status and what it means for epidemiology". In: *Health Policy and Technology* 1.3 (2012), pp. 123–126. URL: `http://dx.doi.org/10.1016/j.hlpt.2012.07.003`.

[111] Karla L Miller, Fidel Alfaro-Almagro, Neal K Bangerter, David L Thomas, Essa Yacoub, Junqian Xu, Andreas J Bartsch, Saad Jbabdi, Stamatios N Sotiropoulos, Jesper LR Andersson, et al. "Multimodal population brain imaging in the UK Biobank prospective epidemiological study". In: *Nature neuroscience* 19.11 (2016), pp. 1523–1536.

[112] Kathryn E Bradbury, Heather J Young, Wenji Guo, and Timothy J Key. "Dietary assessment in UK Biobank: an evaluation of the performance of the touchscreen dietary questionnaire". In: *Journal of nutritional science* 7 (2018), e6.

[113] Katrina AS Davis, Jonathan RI Coleman, Mark Adams, Naomi Allen, Gerome Breen, Breda Cullen, Chris Dickens, Elaine Fox, Nick Graham, Jo Holliday, et al. "Mental health in UK Biobank–development, implementation and results from an online questionnaire completed by 157 366 participants: a reanalysis". In: *BJPsych open* 6.2 (2020), e18.

[114] UK Biobank. "UK Biobank ethics and governance framework". In: *UK BIOBANK* 3 (2007).

[115] Axivity. *Axivity AX3 Technical User Guide*. December, 5, 2023. 2023. URL: `https://axivity.com/userguides/ax3/technical/`.

[116] Vincent T Van Hees, Zhou Fang, Joss Langford, Felix Assah, Anwar Mohammad, Inacio CM Da Silva, Michael I Trenell, Tom White, Nicholas J Wareham, and Søren Brage. "Autocalibration of accelerometer data for free-living physical activity assessment using local gravity and temperature: an evaluation on four continents". In: *Journal of applied physiology* 117.7 (2014), pp. 738–744.

[117] Anna Fry, Thomas J Littlejohns, Cathie Sudlow, Nicola Doherty, Ligia Adamska, Tim Sprosen, Rory Collins, and Naomi E Allen. "Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population". In: *American journal of epidemiology* 186.9 (2017), pp. 1026–1034.

[118] Xiao-Li Meng. "Statistical paradises and paradoxes in big data (i) law of large populations, big data paradox, and the 2016 us presidential election". In: *The Annals of Applied Statistics* 12.2 (2018), pp. 685–726.

[119] Valerie C Bradley, Shiro Kuriwaki, Michael Isakov, Dino Sejdinovic, Xiao-Li Meng, and Seth Flaxman. "Unrepresentative big surveys significantly overestimated US vaccine uptake". In: *Nature* 600.7890 (2021), pp. 695–700.

[120] Valerie Bradley and Thomas E Nichols. "Addressing selection bias in the UK Biobank neurological imaging cohort". In: *MedRxiv* (2022), pp. 2022–01.

[121] Andreas Bulling, Ulf Blanke, and Bernt Schiele. "A tutorial on human activity recognition using body-worn inertial sensors". In: *ACM Computing Surveys (CSUR)* 46.3 (2014), pp. 1–33.

[122] Nils Y Hammerla, Reuben Kirkham, Peter Andras, and Thomas Ploetz. "On preserving statistical characteristics of accelerometry data using their empirical cumulative distribution". In: *Proceedings of the 2013 International Symposium on Wearable Computers*. 2013, pp. 65–68.

[123] Jianbo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiaoli Li, and Shonali Krishnaswamy. "Deep convolutional neural networks on multichannel time series for human activity recognition." In: *IJCAI*. Vol. 15. Buenos Aires, Argentina. 2015, pp. 3995–4001.

[124] Nils Y Hammerla, Shane Halloran, and Thomas Plötz. "Deep, convolutional, and recurrent models for human activity recognition using wearables". In: *arXiv preprint arXiv:1604.08880* (2016).

[125] Charissa Ann Ronao and Sung-Bae Cho. "Human activity recognition with smartphone sensors using deep learning neural networks". In: *Expert Systems with Applications* 59 (2016), pp. 235–244.

[126] Matthew Willetts, Sven Hollowell, Louis Aslett, Chris Holmes, and Aiden Doherty. "Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 UK Biobank participants". In: *Scientific Reports* 8.1 (2018), pp. 1–10.

[127] Hyeokhyen Kwon, Catherine Tong, Harish Haresamudram, Yan Gao, Gregory D Abowd, Nicholas D Lane, and Thomas Ploetz. "IMUTube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4.3 (2020), pp. 1–29.

[128] Setareh Rahimi Taghanaki and Ali Etemad. "Self-supervised Wearable-based Activity Recognition by Learning to Forecast Motion". In: *arXiv preprint arXiv:2010.13713* (2020).

[129] Harish Haresamudram, Apoorva Beedu, Varun Agrawal, Patrick L Grady, Irfan Essa, Judy Hoffman, and Thomas Plötz. "Masked reconstruction based self-supervision for human activity recognition". In: *Proceedings of the 2020 International Symposium on Wearable Computers*. 2020, pp. 45–49.

[130] Harish Haresamudram, Irfan Essa, and Thomas Ploetz. "Contrastive Predictive Coding for Human Activity Recognition". In: *arXiv preprint arXiv:2012.05333* (2020).

[131] Chi Ian Tang, Ignacio Perez-Pozuelo, Dimitris Spathis, and Cecilia Mascolo. "Exploring Contrastive Learning in Human Activity Recognition for Healthcare". In: *arXiv preprint arXiv:2011.11542* (2020).

[132] Chi Ian Tang, Ignacio Perez-Pozuelo, Dimitris Spathis, Soren Brage, Nick Wareham, and Cecilia Mascolo. "SelfHAR: Improving Human Activity Recognition through Self-training with Unlabeled Data". In: *arXiv preprint arXiv:2102.06073* (2021).

[133] Andrew Creagh et al. "Digital health technologies and machine learning augment patient reported outcomes to remotely characterise rheumatoid arthritis". In: *arXiv preprint arXiv* (2022).

[134] Terry T. Um, Franz M. J. Pfister, Daniel Pichler, Satoshi Endo, Muriel Lang, Sandra Hirche, Urban Fietzek, and Dana Kulić. "Data Augmentation of Wearable Sensor Data for Parkinson's Disease Monitoring Using Convolutional Neural Networks". In: *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ICMI 2017. Glasgow, UK: ACM, 2017, pp. 216–220.

[135] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.

[136] DW Eslinger, Ann V Rowlands, Tina L Hurst, Michael Catt, Peter Murray, and Roger G Eston. "Validation of the GENEA Accelerometer". In: *Med Sci Sports Exerc* (2011).

[137] Gary M Weiss, Kenichi Yoneda, and Thaier Hayajneh. "Smartphone and smartwatch-based biometrics using activities of daily living". In: *IEEE Access* 7 (2019), pp. 133190–133202.

[138] Timo Sztyler and Heiner Stuckenschmidt. "On-body localization of wearable devices: An investigation of position-aware activity recognition". In: *2016 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE. 2016, pp. 1–9.

[139] Daniel Roggen, Alberto Calatroni, Mirco Rossi, Thomas Holleczek, Kilian Förster, Gerhard Tröster, Paul Lukowicz, David Bannach, Gerald Pirkl, Alois Ferscha, et al. "Collecting complex activity datasets in highly rich networked sensor environments". In: *2010 Seventh International Conference on Networked Sensing Systems (INSS)*. IEEE. 2010, pp. 233–240.

[140] Attila Reiss and Didier Stricker. "Introducing a new benchmarked dataset for activity monitoring". In: *2012 16th International Symposium on Wearable Computers*. IEEE. 2012, pp. 108–109.

[141] Barbara Bruno, Fulvio Mastrogiovanni, Antonio Sgorbissa, Tullio Vernazza, and Renato Zaccaria. "Analysis of human behavior recognition algorithms based on acceleration data". In: *2013 IEEE International Conference on Robotics and Automation*. IEEE. 2013, pp. 1602–1607.

[142] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Identity mappings in deep residual networks". In: *European Conference on Computer Vision*. Springer. 2016, pp. 630–645.

[143] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[144] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. "Accurate, large minibatch sgd: Training imagenet in 1 hour". In: *arXiv preprint arXiv:1706.02677* (2017).

[145] Shaoyan Zhang, Alex V Rowlands, Peter Murray, Tina L Hurst, et al. "Physical activity classification using the GENEA wrist-worn accelerometer". In: *Medicine and Science in Sports and Exercise* (2012).

[146] Andrea Mannini, Stephen S Intille, Mary Rosenberger, Angelo M Sabatini, and William Haskell. "Activity recognition using a single accelerometer placed at the wrist or ankle". In: *Medicine and Science in Sports and Exercise* 45.11 (2013), p. 2193.

[147] Katherine Ellis, Jacqueline Kerr, Suneeta Godbole, John Staudenmayer, and Gert Lanckriet. "Hip and wrist accelerometer algorithms for free-living behavior classification". In: *Medicine and Science in Sports and Exercise* 48.5 (2016), p. 933.

[148] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. "Realistic evaluation of deep semi-supervised learning algorithms". In: *NeurIPS* 31 (2018).

[149] Leland McInnes, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426* (2018).

[150] Harish Haresamudram, David V Anderson, and Thomas Plötz. "On the role of features in human activity recognition". In: *Proceedings of the 23rd International Symposium on Wearable Computers*. 2019, pp. 78–88.

[151] Luca A Lotta, Laura BL Wittemans, Verena Zuber, Isobel D Stewart, Stephen J Sharp, Jian'an Luan, Felix R Day, Chen Li, Nicholas Bowker, Lina Cai, et al. "Association of genetic variants related to gluteofemoral vs abdominal fat distribution with type 2 diabetes, coronary disease, and cardiovascular risk factors". In: *Jama* 320.24 (2018), pp. 2553–2563.

[152] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. "A simple framework for contrastive learning of visual representations". In: *ICML*. PMLR. 2020, pp. 1597–1607.

[153] Nenad Tomasev, Ioana Bica, Brian McWilliams, Lars Buesing, Razvan Pascanu, Charles Blundell, and Jovana Mitrovic. "Pushing the limits of self-supervised ResNets: Can we outperform supervised learning without labels on ImageNet?" In: *arXiv preprint arXiv:2201.05119* (2022).

[154] Johanna Wilde-Frenz and Hartmut Schulz. "Rate and distribution of body movements during sleep in humans". In: *Perceptual and motor skills* 56.1 (1983), pp. 275–283.

[155] Jairo H Migueles, Pablo Molina-Garcia, Lucia V Torres-Lopez, Cristina Cadenas-Sanchez, Alex V Rowlands, Ulrich W Ebner-Priemer, Elena D Koch, Andreas Reif, and Francisco B Ortega. "Equivalency of four research-grade movement sensors to assess movement behaviors and its implications for population surveillance". In: *Scientific Reports* 12.1 (2022), pp. 1–9.

[156] Christopher B Miller, Simon D Kyle, Kerri L Melehan, and Delwyn J Bartlett. "Methodology for the assessment of sleep". In: *Sleep and Affect*. Elsevier, 2015, pp. 65–90.

[157] Yun Ji Lee, Jae Yong Lee, Jae Hoon Cho, and Ji Ho Choi. "Interrater reliability of sleep stage scoring: a meta-analysis". In: *Journal of Clinical Sleep Medicine* 18.1 (2022), pp. 193–202.

[158] Zhiheng Huang, Wei Xu, and Kai Yu. "Bidirectional LSTM-CRF models for sequence tagging". In: *arXiv preprint arXiv:1508.01991* (2015).

[159] Mary L McHugh. "Interrater reliability: the kappa statistic". In: *Biochemia medica* 22.3 (2012), pp. 276–282.

[160] Jessica R Golbus, Nicole A Pescatore, Brahmajee K Nallamothu, Nirav Shah, and Sachin Kheterpal. "Wearable device signals and home blood pressure data across age, sex, race, ethnicity, and clinical phenotypes in the Michigan Predictive Activity & Clinical Trajectories in Health (MIPACT) study: a prospective, community-based observational study". In: *The Lancet Digital Health* 3.11 (2021), e707–e715.

[161] Jane E Ferrie, Meena Kumari, Paula Salo, Archana Singh-Manoux, and Mika Kivimäki. "Sleep epidemiology—a rapidly growing field". In: *International Journal of Epidemiology* 40.6 (Dec. 2011), pp. 1431–1437. eprint: https://academic.oup.com/ije/article-pdf/40/6/1431/2407775/dyr203.pdf. URL: https://doi.org/10.1093/ije/dyr203.

[162] Michelle A Short, Michael Gradisar, Leon C Lack, Helen Wright, and Mary A Carskadon. "The discrepancy between actigraphic and sleep diary measures of sleep in adolescents". In: *Sleep Medicine* 13.4 (2012), pp. 378–384.

[163] Michael Wainberg, Samuel E Jones, Lindsay Melhuish Beaupre, Sean L Hill, Daniel Felsky, Manuel A Rivas, Andrew SP Lim, Hanna M Ollila, and Shreejoy J Tripathy. "Association of accelerometer-derived sleep measures with lifetime psychiatric diagnoses: A cross-sectional study of 89,205 participants from the UK Biobank". In: *PLoS Medicine* 18.10 (2021), e1003782.

[164] Samuel E Jones, Vincent T van Hees, Diego R Mazzotti, Pedro Marques-Vidal, Séverine Sabia, Ashley van der Spek, Hassan S Dashti, Jorgen Engmann, Desana Kocevska, Jessica Tyrrell, et al. "Genetic studies of accelerometer-based sleep measures yield new insights into human sleep behaviour". In: *Nature Communications* 10.1 (2019), pp. 1–12.

[165] Terry K Koo and Mae Y Li. "A guideline of selecting and reporting intraclass correlation coefficients for reliability research". In: *Journal of Chiropractic Medicine* 15.2 (2016), pp. 155–163.

[166] Max Hirshkowitz, Kaitlyn Whiton, Steven M Albert, Cathy Alessi, Oliviero Bruni, Lydia DonCarlos, Nancy Hazen, John Herman, Paula J Adams Hillard, Eliot S Katz, et al. "National Sleep Foundation's updated sleep duration recommendations". In: *Sleep health* 1.4 (2015), pp. 233–243.

[167] Bin Yan, Jian Yang, Binbin Zhao, Yajuan Fan, Wei Wang, and Xiancang Ma. "Objective sleep efficiency predicts cardiovascular disease in a community population: the sleep heart health study". In: *Journal of the American Heart Association* 10.7 (2021), e016201.

[168] Gail M Sullivan and Richard Feinn. "Using effect size—or why the P value is not enough". In: *Journal of graduate medical education* 4.3 (2012), pp. 279–282.

[169] Douglas F Easton, Julian Peto, and Abdel GAG Babiker. "Floating absolute risk: an alternative to relative risk in survival and case-control analysis avoiding an arbitrary reference group". In: *Statistics in Medicine* 10.7 (1991), pp. 1025–1035.

[170] Martyn Plummer and Bendix Carstensen. "Lexis: An R Class for Epidemiological Studies with Long-Term Follow-Up". In: *Journal of Statistical Software* 38.5 (2011), pp. 1–12. URL: https://www.jstatsoft.org/v38/i05/.

[171] Martyn Plummer. "Improved estimates of floating absolute risk". In: *Statistics in Medicine* 23.1 (2004), pp. 93–104.

[172] Jennifer Girschik, Lin Fritschi, Jane Heyworth, and Flavie Waters. "Validation of self-reported sleep against actigraphy". In: *Journal of Epidemiology* 22.5 (2012), pp. 462–468.

[173] Eus JW Van Someren. "Brain mechanisms of insomnia: new perspectives on causes and consequences". In: *Physiological reviews* 101.3 (2021), pp. 995–1046.

[174] Jiawei Yin, Xiaoling Jin, Zhilei Shan, Shuzhen Li, Hao Huang, Peiyun Li, Xiaobo Peng, Zhao Peng, Kaifeng Yu, Wei Bao, et al. "Relationship of sleep duration with all-cause mortality and cardiovascular events: a systematic review and dose-response meta-analysis of prospective cohort studies". In: *Journal of the American Heart Association* 6.9 (2017), e005947.

[175] Osamu Itani, Maki Jike, Norio Watanabe, and Yoshitaka Kaneita. "Short sleep duration and health outcomes: a systematic review, meta-analysis, and meta-regression". In: *Sleep Medicine* 32 (2017), pp. 246–256.

[176] Yannis Yan Liang, Sizhi Ai, Huachen Xue, Yilin Chen, Jiajin Zhou, Xinyue Shu, Foqian Weng, Mingqing Zhou, Huan Ma, Jihui Zhang, et al. "Joint Associations of Device-measured Sleep Duration and Efficiency with All-cause and Cause-specific Mortality: A Prospective Cohort Study of 90 398 UK Biobank Participants". In: *The Journals of Gerontology: Series A* (2023), glad108.

[177] Chuangshi Wang, Shrikant I Bangdiwala, Sumathy Rangarajan, Scott A Lear, Khalid F AlHabib, Viswanathan Mohan, Koon Teo, Paul Poirier, Lap Ah Tse, Zhiguang Liu, et al. "Association of estimated sleep duration and naps with mortality and cardiovascular events: a study of 116 632 people from 21 countries". In: *European Heart Journal* 40.20 (2019), pp. 1620–1629.

[178] Shahrad Taheri. "Sleep and cardiometabolic health—not so strange bedfellows". In: *The Lancet Diabetes & Endocrinology* (2023).

[179] Jay S Kaufman, Richard S Cooper, and Daniel L McGee. "Socioeconomic status and health in blacks and whites: the problem of residual confounding and the resiliency of race". In: *Epidemiology* (1997), pp. 621–628.

[180] Hiral Master, Jeffrey Annis, Shi Huang, Joshua A Beckman, Francis Ratsimbazafy, Kayla Marginean, Robert Carroll, Karthik Natarajan, Frank E Harrell, Dan M Roden, et al. "Association of step counts over time with the risk of chronic disease in the All of Us Research Program". In: *Nature medicine* 28.11 (2022), pp. 2301–2308.

[181] Tessa Strain, Katrien Wijndaele, Stephen J Sharp, Paddy C Dempsey, Nick Wareham, and Søren Brage. "Impact of follow-up time and analytical approaches to account for reverse causality on the association between physical activity and health outcomes in UK Biobank". In: *International journal of epidemiology* 49.1 (2020), pp. 162–172.

[182] Andrea L Dunn, Ross E Andersen, and John M Jakicic. "Lifestyle physical activity interventions: History, short-and long-term effects, and recommendations". In: *American journal of preventive medicine* 15.4 (1998), pp. 398–412.

[183] Matthew J Blake, Lisa B Sheeber, George J Youssef, Monika B Raniti, and Nicholas B Allen. "Systematic review and meta-analysis of adolescent cognitive–behavioral sleep interventions". In: *Clinical child and family psychology review* 20 (2017), pp. 227–249.

[184] Katie Whale, Jane Dennis, Vikki Wylde, Andrew Beswick, and Rachael Gooberman-Hill. "The effectiveness of non-pharmacological sleep interventions for people with chronic pain: a systematic review and meta-analysis". In: *BMC Musculoskeletal Disorders* 23.1 (2022), p. 440.

[185] Yuanyuan Chen, Shing Chan, Derrick Bennett, Xiaofang Chen, Xianping Wu, Yalei Ke, Jun Lv, Dianjianyi Sun, Lang Pan, Pei Pei, et al. "Device-measured movement behaviours in over 20,000 China Kadoorie Biobank participants". In: *International Journal of Behavioral Nutrition and Physical Activity* 20.1 (2023), p. 138.

[186] Rebecca C Richmond, Emma L Anderson, Hassan S Dashti, Samuel E Jones, Jacqueline M Lane, Linn Beate Strand, Ben Brumpton, Martin K Rutter, Andrew R Wood, Kurt Straif, et al. "Investigating causal relations between sleep traits and risk of breast cancer in women: Mendelian randomisation study". In: *bmj* 365 (2019).

[187] Eugene Aserinsky and Nathaniel Kleitman. "Regularly occurring periods of eye motility, and concomitant phenomena, during sleep". In: *Science* 118.3062 (1953), pp. 273–274.

[188] Eugene Aserinsky and Nathaniel Kleitman. "Two types of ocular motility occurring in sleep". In: *Journal of applied physiology* 8.1 (1955), pp. 1–10.

[189] Hao Jiang. "Motion eavesdropper: Smartwatch-based handwriting recognition using deep learning". In: *2019 International Conference on Multimodal Interaction*. 2019, pp. 145–153.

[190] *Wearables Bounce Back with 8.5% Growth in Q2 2023 and a Positive Forecast, According to IDC*. Sept. 2023. URL: `https://www.bloomberg.com/press-releases/2023-09-20/wearables-bounce-back-with-8-5-growth-in-q2-2023-and-a-positive-forecast-according-to-idc`.

[191] Lovedeep S Dhingra, Arya Aminorroaya, Evangelos Oikonomou, Arash Aghajani Nargesi, and Rohan Khera. "Use of Wearable Devices by Patients With and At-Risk for Cardiovascular Disease in the United States: A Nationally Representative Study". In: *Circulation* 146.Suppl_1 (2022), A15333–A15333.

[192] Inbar Raber, Cian P McCarthy, and Robert W Yeh. "Health insurance and mobile health devices: opportunities and concerns". In: *Jama* 321.18 (2019), pp. 1767–1768.

[193] Ministry of Health. *Healthier SG*. Accessed on 2023, 10, 10. 2023. URL: `https://www.healthiersg.gov.sg`.

[194] Lucy Chikwetu, Yu Miao, Melat K Woldetensae, Diarra Bell, Daniel M Goldenholz, and Jessilyn Dunn. "Does deidentification of data from wearable devices give us a false sense of security? A systematic review". In: *The Lancet Digital Health* (2023).

[195] Henry T Greely. "The uneasy ethical and legal underpinnings of large-scale genomic biobanks". In: *Annu. Rev. Genomics Hum. Genet.* 8 (2007), pp. 343–364.

[196] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. "The future of digital health with federated learning". In: *NPJ digital medicine* 3.1 (2020), p. 119.

[197] Zhihan Lv and Francesco Piccialli. "The security of medical data on internet based on differential privacy technology". In: *ACM Transactions on Internet Technology* 21.3 (2021), pp. 1–18.

[198] Zhongzheng Fu, Xinrun He, Enkai Wang, Jun Huo, Jian Huang, and Dongrui Wu. "Personalized human activity recognition based on integrated wearable sensor and transfer learning". In: *Sensors* 21.3 (2021), p. 885.

[199] Dimitris Spathis, Ignacio Perez-Pozuelo, Tomas I Gonzales, Yu Wu, Soren Brage, Nicholas Wareham, and Cecilia Mascolo. "Longitudinal cardio-respiratory fitness prediction through wearables in free-living environments". In: *NPJ Digital Medicine* 5.1 (2022), p. 176.

[200] Daniel M Roberts, Margeaux M Schade, Gina M Mathew, Daniel Gartenberg, and Orfeu M Buxton. "Detecting sleep using heart rate and motion data from multisensor consumer-grade wearables, relative to wrist actigraphy and polysomnography". In: *Sleep* 43.7 (2020), zsaa045.

[201] Youngsin Jung and Erik K St. Louis. "Treatment of REM sleep behavior disorder". In: *Current treatment options in neurology* 18 (2016), pp. 1–12.

[202] Daniel J Buysse. "Sleep health: can we define it? Does it matter?" In: *Sleep* 37.1 (2014), pp. 9–17.

[203] Angus C Burns, Daniel P Windred, Martin K Rutter, Patrick Olivier, Céline Vetter, Richa Saxena, Jacqueline M Lane, Andrew JK Phillips, and Sean W Cain. "Day and night light exposure are associated with psychiatric disorders: an objective light study in> 85,000 people". In: *Nature Mental Health* (2023), pp. 1–10.

[204] Thomas G Brooks, Nicholas F Lahens, Gregory R Grant, Yvette I Sheline, Garret A FitzGerald, and Carsten Skarke. "Diurnal rhythms of wrist temperature are associated with future disease risk in the UK Biobank". In: *Nature Communications* 14.1 (2023), p. 5172.

[205] Till Roenneberg and Martha Merrow. "The circadian clock and human health". In: *Current biology* 26.10 (2016), R432–R443.

[206] Shahab Haghayegh, Sepideh Khoshnevis, Michael H Smolensky, Kenneth R Diller, and Richard J Castriotta. "Performance assessment of new-generation Fitbit technology in deriving sleep parameters and stages". In: *Chronobiology international* 37.1 (2020), pp. 47–59.

[207] Julie L Darbyshire, Mark Borthwick, Peter Edmonds, Sarah Vollam, Lisa Hinton, and J Duncan Young. "Measuring sleep in the intensive care unit: Electroencephalogram, actigraphy, or questionnaire?" In: *Journal of the Intensive Care Society* 21.1 PG - 22-27 (2020), pp. 22–27.

[208] Cristina D Toedebusch, Jennifer S McLeland, Claire M Schaibley, Ian R Banks, Jill Boyd, John C Morris, David M Holtzman, and Brendan P Lucey. "Multi-Modal Home Sleep Monitoring in Older Adults". In: *Journal of visualized experiments : JoVE* 143 PG - (2019).

[209] L Menghini, V Alschuler, S Claudatos, A Goldstone, F Baker, N Cellini, I Colrain, and M De Zambotti. "Accuracy of a commercial wearable in detecting sleep stages compared to polysomnography in adults: Considering sleep classification methods and effects of evening alcohol consumption". In: *Sleep* 43.SUPPL 1 PG - A456-A457 (2020), A456–A457.

[210] S Kumar and S S Dash. "Nonparametric approach for sleep stage classification using cardiorespiratory and movement features". In: *Journal of sleep research* 27.Supplement 1 PG - 231 (2018), p. 231.

[211] E P Pollet, D P Pollet, B Long, and A A Qutub. "Activity trackers as a tool in sleep research: determining discrepancies in trackers vs. psg". In: *Sleep* 43.SUPPL 1 PG - A460-A461 (2020), A460–A461.

[212] D Kim, W Shin, and J Byun. "Utility of fitbit charge 2 for sleep monitoring in patients with obstructive sleep apnea". In: *Sleep* 43.SUPPL 1 PG - A458 (2020), A458.

[213] P Kawhawage, K Hamill, R Jumabhoy, M De Zambotti, and S P A Drummond. "Validation of consumer and research-grade activity trackers in Insomnia Disorder I: In-lab validation against polysomnography". In: *Journal of sleep research* 28.SUPPL 1 PG - (2019).

[214] P Kahawage, R Jumabhoy, K Hamill, T Nguyen, and S P A Drummond. "Validation of a consumer activity tracker against polysomnography and actigraphy in insomnia". In: *Sleep* 42.Supplement 1 PG - A128 (2019), A128.

[215] M Olsen, H Sorensen, P Jennum, and E Mignot. "Sleep stage prediction and sleep disordered breathing detection using raw actigraphy and photoplethysmography from wearable consumer device". In: *Sleep* 43.SUPPL 1 PG - A461-A462 (2020), A461–A462.

[216] H Samman, M Mansukhani, S Dhankikar, and B P Kolla. "Utility of actigraphy and polysomnography prior to conducting the adult multiple sleep latency test". In: *Sleep medicine* 64.Supplement 1 PG - S242 (2019), S242.

[217] J Collen and S Lesage. "The accuracy of actigraphy in the evaluation of patients with hypersomnia". In: *Sleep* 36.SUPPL. 1 PG - A251-A252 (2013), A251–A252.

[218] J Chung, M O Goodman, T Huang, M Wallace, S Bertisch, D Johnson, and S Redline. "Racial/ethnic differences in actigraphy, questionnaire, and polysomnographymeasured indicators of sleep health and sleep quality: the multi-ethnic study of atherosclerosis". In: *Sleep* 43.SUPPL 1 PG - A138-A139 (2020), A138–A139.

[219] A Alakuijala, T Jokela, and J Toppila. "Concomitant actigraphy and polysomnography at home in different sleep disorders". In: *Journal of sleep research* 27.Supplement 1 PG - 136 (2018), p. 136.

[220] S Dietz-Terjung, A Martin, and C Schobel. "A novel algorithm for the estimation of sleep states based on breathing and movement". In: *Sleep* 43.SUPPL 1 PG - A170 (2020), A170.

[221] F Baker, A Goldstone, S Claudatos, L Rosas, V Alschuler, Y Q Lim, M Gil, I Colrain, and M De Zambotti. "Validation of a multi-sensory commercially available wristband in measuring sleep composition against polysomnography". In: *Sleep medicine* 40.Supplement 1 PG - e21-e22 (2017), e21–e22.

[222] S Allen, G Elder, R Sharman, Z M Gotts, and J G Ellis. "Identification of objective and subjective markers of sleep health". In: *Sleep* 41.Supplement 1 PG - A125 (2018), A125.

[223] Jaime K Devine, Evan D Chinoy, Rachel R Markwald, Lindsay P Schwartz, and Steven R Hursh. "Validation of Zulu Watch against Polysomnography and Actigraphy for On-Wrist Sleep-Wake Determination and Sleep-Depth Estimation". In: *Sensors (Basel, Switzerland)* 21.1 PG - (2020).

[224] C Alcantara, M L Biggs, K W Davidson, J A Delaney, C L Jackson, P C Zee, S J Shea, and S Redline. "Sleep disturbances and depression in the multi-ethnic study of atherosclerosis". In: *Sleep* 38.SUPPL. 1 PG - A328 (2015), A328.

[225] Tim Willemen, Dorien Van Deun, Vincent Verhaert, Sandra Pirrera, Vasileios Exadaktylos, Johan Verbraecken, Bart Haex, and Jos Vander Sloten. "Automatic sleep stage classification based on easy to register signals as a validation tool for ergonomic steering in smart bedding systems". In: *WORK-A JOURNAL OF PREVENTION ASSESSMENT & REHABILITATION* 41.1 PG - 1985-1989 (2012), pp. 1985–1989.

[226] Jan Hedner, David P White, Atul Malhotra, Sarah Herscovici, Stephen D Pittman, Ding Zou, Ludger Grote, and Giora Pillar. "Sleep staging based on autonomic signals: a multi-center validation study". In: *Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine* 7.3 PG - 301-6 (2011), pp. 301–306.

[227] Christoph Kalkbrenner, Rainer Brucher, Tibor Kesztyus, Manuel Eichenlaub, Wolfgang Rottbauer, and Dominik Scharnbeck. "Automated sleep stage classification based on tracheal body sound and actigraphy". In: *German medical science : GMS e-journal* 17.PG - Doc02 (2019), Doc02.

[228] A Zaffaroni, E P Doheny, L Gahan, Y Ivanov, H Kilroy, N O'Mahony, D O'Rourke, H Eskola, O Vaisanen, J Viik, and J Hyttinen. "Non-Contact Estimation of Sleep Staging". In: 65.PG - 77-80 (2018), pp. 77–80.

[229] Patricia Concheiro-Moscoso, Francisco Jose Martinez-Martinez, Maria del Carmen Miranda-Duro, Thais Pousada, Laura Nieto-Riveiro, Betania Groba, Francisco Javier Mejuto-Muino, and Javier Pereira. "Study Protocol on the Validation of the Quality of Sleep Data from Xiaomi Domestic Wristbands". In: *INTERNATIONAL JOURNAL OF ENVIRONMENTAL RESEARCH AND PUBLIC HEALTH* 18.3 PG - (2021).

[230] Engin Turetken, Jerome Van Zaen, and Ricard Delgado-Gonzalo. "Embedded Deep Learning for Sleep Staging". In: PG - 95-96 (2019), pp. 95–96.

[231] Shahab Haghayegh, Sepideh Khoshnevis, Michael H Smolensky, Kenneth R Diller, and Richard J Castriotta. "Deep Neural Network Sleep Scoring Using Combined Motion and Heart Rate Variability Data". In: *Sensors* 21.1 PG - (2021).

[232] Fernando Moreno-Pino, Alejandro Porras-Segovia, Pilar Lopez-Esteban, Antonio Artes, and Enrique Baca-Garcia. "Validation of Fitbit Charge 2 and Fitbit Alta HR Against Polysomnography for Assessing Sleep in Adults With Obstructive Sleep Apnea". In: *Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine* 15.11 PG - 1645-1653 (2019), pp. 1645–1653.

[233] Zilu Liang and Mario Alberto Chapa-Martell. "Accuracy of Fitbit Wristbands in Measuring Sleep Stage Transitions and the Effect of User-Specific Factors". In: *JMIR mHealth and uHealth* 7.6 PG - e13384 (2019), e13384.

[234] Bernice M Wulterkens, Pedro Fonseca, Lieke W A Hermans, Marco Ross, Andreas Cerny, Peter Anderer, Xi Long, Johannes P van Dijk, Nele Vandenbussche, and Sigrid Pillen. "It is all in the wrist: wearable sleep staging in a clinical population versus reference polysomnography". In: *Nature and Science of Sleep* 13.PG - 885 (2021), p. 885.

[235] Dean J Miller, Michele Lastella, Aaron T Scanlan, Clint Bellenger, Shona L Halson, Gregory D Roach, and Charli Sargent. "A validation study of the WHOOP strap against polysomnography to assess sleep". In: *JOURNAL OF SPORTS SCIENCES* 38.22 PG - 2631-2636 (2020), pp. 2631–2636.

[236] Dean J Miller, Gregory D Roach, Michele Lastella, Aaron T Scanlan, Clint R Bellenger, Shona L Halson, and Charli Sargent. "A Validation Study of a Commercial Wearable Device to Automatically Detect and Estimate Sleep". In: *Biosensors* 11.6 PG - 185 (2021), p. 185.

[237] Ph Renevey, R Delgado-Gonzalo, A Lemkaddem, M Proenca, M Lemay, J Sola, A Tarniceriu, M Bertschi, H Eskola, O Vaisanen, J Viik, and J Hyttinen. "Optical wrist-worn device for sleep monitoring". In: 65.PG - 615-618 (2018), pp. 615–618.

[238] L Kuula and A K Pesonen. "Heart Rate Variability and Firstbeat Method for Detecting Sleep Stages in Healthy Young Adults: Feasibility Study". In: *JMIR mHealth and uHealth* 9.2 PG - e24704 (2021), e24704.

[239] A Muzet, S Werner, G Fuchs, T Roth, J B Saoud, A U Viola, J Y Schaffhauser, and R Luthringer. "Assessing sleep architecture and continuity measures through the analysis of heart rate and wrist movement recordings in healthy subjects: Comparison with results based on polysomnography". In: *Sleep medicine* 21.PG - 47-56 (2016), pp. 47–56.

[240] Yuezhou Zhang, Zhicheng Yang, Ke Lan, Xiaoli Liu, Zhengbo Zhang, Peiyao Li, Desen Cao, Jiewen Zheng, and Jianli Pan. "Sleep Stage Classification Using Bidirectional LSTM in Wearable Multi-sensor Systems". In: PG - 443-448 (2019), pp. 443–448.

[241] Piyumi Kahawage, Ria Jumabhoy, Kellie Hamill, Massimiliano de Zambotti, and Sean P A Drummond. "Validity, potential clinical utility, and comparison of consumer and research-grade activity trackers in Insomnia Disorder I: In-lab validation against polysomnography". In: *Journal of sleep research* 29.1 PG - e12931 (2020), e12931.

[242] A Ribeiro, R Gabriel, A William, M Alejandro, G Bernardo, C Landeen, D Pfaff, and N B Kavey. "Evaluation of the Watch-PAT apparatus as a sleep-staging tool". In: *Sleep* 40.Supplement 1 PG - A120-A121 (2017), A120–A121.

[243] K Fujimoto, Y Ding, and E Takahashi. "Sleep stage detection using a wristwatch-type physiological sensing device". In: *SLEEP AND BIOLOGICAL RHYTHMS* 16.4 PG - 449-456 (2018), pp. 449–456.

[244] P Fonseca, T Weysen, M S Goelema, E I S Most, M Radha, C L Scheurleer, L Van Den Heuvel, and R M Aarts. "Validation of photoplethysmography-based sleep staging compared with polysomnography in healthy middle-aged adults". In: *Sleep* 40.7 PG - 097 (2017), p. 97.

[245] Pedro Fonseca, Merel M van Gilst, Mustafa Radha, Marco Ross, Arnaud Moreau, Andreas Cerny, Peter Anderer, Xi Long, Johannes P van Dijk, and Sebastiaan Overeem. "Automatic sleep staging using heart rate variability, body movements, and recurrent neural networks in a sleep disordered population". In: *Sleep* 43.9 PG - (2020).

[246] Massimiliano de Zambotti, Aimee Goldstone, Stephanie Claudatos, Ian M Colrain, and Fiona C Baker. "A validation study of Fitbit Charge 2 TM compared with polysomnography in adults". In: *Chronobiology international* 35.4 PG - 465-476 (2018), pp. 465–476.

[247] Xin Zhang, Weixuan Kou, Eric I Chao Chang, He Gao, Yubo Fan, and Yan Xu. "Sleep stage classification based on multi-level feature learning and recurrent neural networks via wearable device". In: *Computers in biology and medicine* 103.PG - 71-81 (2018), pp. 71–81.

[248] H Schneider, Z Zhang, and A Schwartz. "A comparison of automated and manual sleep staging and respiratory event recognition in a portable sleep diagnostic device with in-lab sleep study". In: *EUROPEAN RESPIRATORY JOURNAL* 56.Supplement 64 PG - (2020).

[249] Jesse D Cook, Sahand C Eftekari, Erika Dallmann, Megan Sippy, and David T Plante. "Ability of the Fitbit Alta HR to quantify and classify sleep in patients with suspected central disorders of hypersomnolence: A comparison against polysomnography". In: *Journal of sleep research* 28.4 PG - e12789 (2019), e12789.

[250] Ma'ayan Bresler, Koby Sheffy, Giora Pillar, Meir Preiszler, and Sarah Herscovici. "Differentiating between light and deep sleep stages using an ambulatory device based on peripheral arterial tonometry". In: *Physiological measurement* 29.5 PG - 571-84 (2008), pp. 571–584.

[251] Ji Ho Choi, Eun Joong Kim, Yang Soo Kim, June Choi, Tae Hoon Kim, Soon Young Kwon, Heung Man Lee, Sang Hag Lee, Chol Shin, and Seung Hoon Lee. "Validation study of portable device for the diagnosis of obstructive sleep apnea according to the new AASM scoring criteria: Watch-PAT 100". In: *Acta oto-laryngologica* 130.7 PG - 838-43 (2010), pp. 838–843.

[252] Evan D Chinoy, Joseph A Cuellar, Kirbie E Huwa, Jason T Jameson, Catherine H Watson, Sara C Bessman, Dale A Hirsch, Adam D Cooper, Sean P A Drummond, and Rachel R Markwald. "Performance of Seven Consumer Sleep-Tracking Devices Compared with Polysomnography". In: *Sleep* PG - (2020).

[253] T Willemen, D Van Deun, V Verhaert, M Vandekerckhove, V Exadaktylos, J Verbraecken, S Van Huffel, B Haex, and J Vander Sloten. "An Evaluation of Cardiorespiratory and Movement Features With Respect to Sleep-Stage Classification". In: *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS* 18.2 PG - 661-669 (2014), pp. 661–669.

[254] Benjamin Stucky, Ian Clark, Yasmine Azza, Walter Karlen, Peter Achermann, Birgit Kleim, and Hans-Peter Landolt. "Validation of Fitbit Charge 2 Sleep and Heart Rate Estimates Against Polysomnographic Measures in Shift Workers: Naturalistic Study". In: *J Med Internet Res* 23.10 PG - e26476 (2021), e26476.

[255] Nilgun Surmen Onder, Meltem Esen Akpinar, Ozgur Yigit, and Ayse Pelin Gor. "Watch peripheral arterial tonometry in the diagnosis of obstructive sleep apnea: influence of aging". In: *The Laryngoscope* 122.6 PG - 1409-14 (2012), pp. 1409–1414.

[256] Jennifer C Kanady, Leslie Ruoff, Laura D Straus, Jonathan Varbel, Thomas Metzler, Anne Richards, Sabra S Inslicht, Aoife O'Donovan, Jennifer Hlavin, and Thomas C Neylan. "Validation of sleep measurement in a multisensor consumer grade wearable device in healthy young adults". In: *Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine* 16.6 PG - 917-924 (2020), pp. 917–924.

[257] M J Kim, G H Lee, C S Kim, W S Kim, Y S Chung, S Chung, and S A Lee. "Comparison of three actigraphic algorithms used to evaluate sleep in patients with obstructive sleep apnea". In: *Sleep and Breathing* 17.1 PG - 297-304 (2013), pp. 297–304.

[258] Amna Barouni, Jorg Ottenbacher, Johannes Schneider, Bernd Feige, Dieter Riemann, Anne Herlan, Driss El Hardouz, and Darren McLennan. "Ambulatory sleep scoring using accelerometers-distinguishing between nonwear and sleep/wake states". In: *PEERJ* 8.PG - e8284 (2020), e8284.

[259] A Barouni, J Ottenbacher, J Schneider, B Feige, D Riemann, A Herlan, D E Hardouz, and D McLennan. "Ambulatory sleep scoring using accelerometers-distinguishing between nonwear and sleep/wake states". In: *PEERJ* 2020.1 PG - 8284 (2020), p. 8284.

[260] Jaime M Beecroft, Michael Ward, Magdy Younes, Shelley Crombach, Orla Smith, and Patrick J Hanly. "Sleep monitoring in the intensive care unit: comparison of nurse assessment, actigraphy and polysomnography". In: *INTENSIVE CARE MEDICINE* 34.11 PG - 2076-83 (2008), pp. 2076–2083.

[261] Su Jung Choi, Miri Kang, Min Je Sung, and Eun Yeon Joo. "Discordant sleep parameters among actigraphy, polysomnography, and perceived sleep in patients with sleep-disordered breathing in comparison with patients with chronic insomnia disorder". In: *Sleep & breathing = Schlaf & Atmung* 21.4 PG - 837-843 (2017), pp. 837–843.

[262] Jesse D Cook, Michael L Prairie, and David T Plante. "Utility of the Fitbit Flex to evaluate sleep in major depressive disorder: A comparison against polysomnography and wrist-worn actigraphy". In: *Journal of affective disorders* 217.PG - 299-305 (2017), pp. 299–305.

[263] Jack D Edinger, Melanie K Means, Karen M Stechuchak, and Maren K Olsen. "A pilot study of inexpensive sleep-assessment devices". In: *BEHAVIORAL SLEEP MEDICINE* 2.1 PG - 41-9 (2004), pp. 41–49.

[264] Rachel Danzig, Mengxi Wang, Amit Shah, and Lynn Marie Trotti. "The wrist is not the brain: Estimation of sleep by clinical and consumer wearable actigraphy devices is impacted by multiple patient- and device-specific factors". In: *Journal of sleep research* 29.1 PG - e12926 (2020), e12926.

[265] G Jean-Louis, D F Kripke, W J Mason, J A Elliott, and S D Youngstedt. "Sleep estimation from wrist movement quantified by different actigraphic modalities". In: *JOURNAL OF NEUROSCIENCE METHODS* 105.2 PG - 185-91 (2001), pp. 185–191.

[266] H Honma, M Kohsaka, I Suzuki, N Fukuda, R Kobayashi, S Sakakibara, S Matubara, and T Koyama. "Motor activity rhythm in dementia with delirium". In: *Psychiatry and clinical neurosciences* 52.2 PG - 196-8 (1998), pp. 196–198.

[267] Jesse D Cook, Michael L Prairie, and David T Plante. "Ability of the Multisensory Jawbone UP3 to Quantify and Classify Sleep in Patients With Suspected Central Disorders of Hypersomnolence: A Comparison Against Polysomnography and Actigraphy". In: *Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine* 14.5 PG - 841-848 (2018), pp. 841–848.

[268] M Mikulec, J Mekyska, J Sigmund, Z Galaz, L Brabenec, I Moravkova, and I Rektorova. *Automatic Segmentation of Actigraphy Data Utilising Gradient Boosting Algorithm*. 2021.

[269] Md Aktaruzzaman, Massimo Walter Rivolta, Ruby Karmacharya, Nello Scarabottolo, Luigi Pugnetti, Massimo Garegnani, Gabriele Bovi, Giovanni Scalera, Maurizio Ferrarin, and Roberto Sassi. "Performance comparison between wrist and chest actigraphy in combination with heart rate variability for sleep classification". In: *Computers in biology and medicine* 89 (2017), pp. 212–221.

[270] Huaming Shen, Meihua Xu, Allon Guez, Ang Li, and Feng Ran. "An Accurate Sleep Stages Classification Method Based on State Space Model". In: *IEEE ACCESS* 7.PG - 125268-125279 (2019), pp. 125268–125279.

[271] Daniel J Levendowski, Sean Seagraves, Djordje Popovic, and Philip R Westbrook. "Assessment of a neck-based treatment and monitoring device for positional obstructive sleep apnea". In: *Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine* 10.8 PG - 863-71 (2014), pp. 863–871.

[272] I R Bell, A Howerter, N Jackson, M Aickin, C M Baldwin, and R R Bootzin. "Effects of homeopathic medicines on polysomnographic sleep of young adults with histories of coffee-related insomnia". In: *Sleep medicine* 12.5 PG - 505-511 (2011), pp. 505–511.

[273] I A Barash, C Beatty, F L Powell, G K Prisk, and J B West. "Nocturnal oxygen enrichment of room air at 3800 meter altitude improves sleep architecture". In: *HIGH ALTITUDE MEDICINE & BIOLOGY* 2.4 PG - 525-33 (2001), pp. 525–533.

[274] Carmela Alcantara, Mary L Biggs, Karina W Davidson, Joseph A Delaney, Chandra L Jackson, Phyllis C Zee, Steven J C Shea, and Susan Redline. "Sleep Disturbances and Depression in the Multi-Ethnic Study of Atherosclerosis". In: *Sleep* 39.4 PG - 915-25 (2016), pp. 915–925.