

1 Self-supervised learning of accelerometer data provides new
2 insights for sleep and its association with mortality

3 Hang Yuan^{a,b}, Tatiana Plekhanova^g, Rosemary Walmsley^{a,b}, Amy C. Reynolds^l,
4 Kathleen J. Maddison^{j,k}, Maja Bucan^f, Philip Gehrman^e, Alex Rowlands^{g,h}, David
5 W. Ray^{c,1}, Derrick Bennett^{a,m}, Joanne McVeighⁱ, Leon Strakerⁱ, Peter Eastwoodⁿ,
6 Simon D. Kyle^o, Aiden Doherty^{a,b}

^a*Nuffield Department of Population Health, University of Oxford, UK*

^b*Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, UK*

^c*NIHR Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford, UK*

^d*Oxford Centre for Diabetes, Endocrinology and Metabolism, Oxford Kavli Centre for Nanoscience Discovery, University of Oxford, UK*

^e*Department of Psychiatry, University of Pennsylvania, USA*

^f*Department of Genetics, University of Pennsylvania, USA*

^g*Diabetes Research Centre, University of Leicester, UK*

^h*NIHR Leicester Biomedical Research Centre, University of Leicester, UK*

ⁱ*Curtin School of Allied Health, Curtin University, Australia*

^j*Centre of Sleep Science, School of Human Sciences, University of Western Australia, Australia*

^k*West Australian Sleep Disorders Research Institute, Department of Pulmonary Physiology, Sir Charles Gairdner Hospital, Australia*

^l*College of Medicine and Public Health, Flinders University, Australia*

^m*Medical Research Council Population Health Research Unit, University of Oxford, UK*

ⁿ*Health Futures Institute, Murdoch University, Australia*

^o*Sir Jules Thorn Sleep & Circadian Neuroscience Institute, Nuffield Department of Clinical Neurosciences, University of Oxford, UK*

July 7, 2023

7 Summary

8 **Background.** Sleep is essential to life. Accurate measurement and classification
9 of sleep/wake and sleep stages is important in clinical studies for sleep disorder
10 diagnoses and in the interpretation of data from consumer devices for monitoring
11 physical and mental well-being. Existing non-polysomnography sleep classification
12 techniques mainly rely on heuristic methods developed in relatively small cohorts.
13 Thus, we aimed to establish the accuracy of wrist-worn accelerometers for sleep stage
14 classification and subsequently describe the association between sleep duration and
15 efficiency (proportion of total time asleep when in bed) with mortality outcomes.

16 **Methods.** We developed and validated a self-supervised deep neural network for
17 sleep stage classification using concurrent laboratory-based polysomnography and
18 accelerometry data from three countries (Australia, the UK, and the USA). The
19 model was validated within-cohort using subject-wise five-fold cross-validation for
20 sleep-wake classification and in a three-class setting for sleep stage classification
21 wake, rapid-eye-movement sleep (REM), non-rapid-eye-movement sleep (NREM) and
22 by external validation. We assessed the face validity of our model for population
23 inference by applying the model to the UK Biobank with 100,000 participants, each
24 of whom wore a wristband for up to seven days. The derived sleep parameters were
25 used in a Cox regression model to study the association of sleep duration and sleep
26 efficiency with all-cause mortality.

27 **Findings.** After exclusion, 1,448 participant nights of data were used to train the
28 sleep classifier. The difference between polysomnography and the model classifica-
29 tions on the external validation was 34.7 minutes (95% limits of agreement (LoA):
30 -37.8 to 107.2 minutes) for total sleep duration, 2.6 minutes for REM duration (95%
31 LoA: -68.4 to 73.4 minutes) and 32.1 minutes (95% LoA: -54.4 to 118.5 minutes) for
32 NREM duration. The derived sleep architecture estimate in the UK Biobank sample
33 showed good face validity. Among 66,214 UK Biobank participants, 1,642 mortal-
34 ity events were observed. Short sleepers (<6 hours) had a higher risk of mortality
35 compared to participants with normal sleep duration (6 to 7.9 hours), regardless of
36 whether they had low sleep efficiency (Hazard ratios (HRs): 1.69; 95% confidence

37 intervals (CIs): 1.28 to 2.24) or high sleep efficiency (HRs: 1.42; 95% CIs: 1.14 to
38 1.77).

39 **Interpretation.** Deep-learning-based sleep classification using accelerometers has a
40 fair to moderate agreement with polysomnography. Our findings suggest that having
41 short overnight sleep confers mortality risk irrespective of sleep continuity.

42 **Funding.** This research has been conducted using the UK Biobank Resource under
43 Application Number 59070. The UK Biobank received ethical approval from the
44 National Health Service National Research Service (Ref 21/NW/0157). We would
45 like to acknowledge the Raine Study participants and their families for their on-
46 going participation in the study and the Raine Study team for study coordination
47 and data collection. We also thank the NHMRC for their long-term contribution to
48 funding the study over the last 30 years. The core management of the Raine Study
49 is funded by The University of Western Australia, Curtin University, Telethon Kids
50 Institute, Women and Infants Research Foundation, Edith Cowan University, Mur-
51 doch University, The University of Notre Dame Australia and the Raine Medical Re-
52 search Foundation. The 22-year Gen2 Raine Study follow-up was funded by NHMRC
53 project grants 1027449 & 1044840. The data collection for the Pennsylvania dataset
54 is funded, in part, by US National Institute of Health (NIMH) grant R21 MH103963
55 (MB).

56 HY, DB, and AD are supported by Novo Nordisk. RW and AD are supported by
57 Health Data Research UK, an initiative funded by UK Research and Innovation, De-
58 partment of Health and Social Care (England) and the devolved administrations, and
59 leading medical research charities. AD is additionally supported by Swiss Re, Well-
60 come Trust [223100/Z/21/Z], and the British Heart Foundation Centre of Research
61 Excellence (grant number RE/18/3/34214). DWR is supported by MRC programme
62 grant MR/P023576/1; Wellcome Trust (107849/Z/15/Z). TP and AR are supported
63 by the National Institute for Health Research (NIHR) Leicester Biomedical Research
64 Centre and NIHR Applied Research Collaboration East Midlands (ARC EM). SDK
65 is supported by the NIHR Oxford Health Biomedical Research Centre, Health Tech-
66 nology Assessment Programme, Efficacy and Mechanisms Evaluation Programme,
67 Programme Grants for Applied Research, and the Wellcome Trust. The views ex-

68 pressed are those of the authors and not necessarily those of the NHS, the NIHR or
69 the Department of Health.

70 Computational aspects of this research were funded from the National Institute
71 for Health Research (NIHR) Oxford Biomedical Research Centre (BRC) with addi-
72 tional support from Health Data Research (HDR) UK and the Wellcome Trust Core
73 Award [grant number 203141/Z/16/Z]. The views expressed are those of the authors
74 and not necessarily those of the NHS, the NIHR or the Department of Health.

75 For the purpose of open access, the author has applied a CC-BY public copyright
76 licence to any author accepted manuscript version arising from this submission.

Research in context

Evidence before this study

Sleep plays a crucial role in our mental and physical health. Nonetheless, much of our understanding of sleep relies on self-report sleep questionnaires, which are subject to recall bias. We searched on Web of Science, Medline, and Google Scholar from the database inception to June 23, 2023, using terms that included “wearable”, “actigraphy” or “accelerometer” in combination with “sleep stage” or “sleep classification”, and “polysomnography”. Existing studies have attempted to use machine learning to predict both sleep and sleep stages using accelerometry. However, prior methods were validated in populations of small sample sizes ($n < 100$), making the prediction validity unclear. To date, no study has examined variations of accelerometer-derived sleep stage estimates in large population datasets with longitudinal disease outcomes.

Added value of this study

We showed that our deep-learning-based method improves sleep staging for wrist-worn accelerometers against the current state-of-the-art. We quantified the model uncertainty in a large multicentre dataset with 1,448 nights of concurrent raw accelerometry and polysomnography recordings. We further demonstrated that our sleep staging method could capture population differences concerning age, season, and other sociodemographic characteristics using a large health database. Shorter overnight sleep duration was associated with an increased risk of all-cause mortality after seven years of follow-up in groups with both low and high sleep efficiencies.

Implications of all the available evidence

This study helps clinicians to interpret sleep measurements from wearable sensors in routine care. Researchers can use derived sleep parameters in large-scale accelerometer datasets to advance our understanding of the association between sleep and population subgroups with different clinical characteristics. Our findings further suggest that having a short overnight sleep is a risky behaviour regardless of the sleep quality, which requires immediate public attention to fight the social stigma that having a short sleep is acceptable as long as one sleeps well.

78 1. Introduction

79 Sleep is essential to life and is structurally complex. Humans spend approximately
80 one third of their lives asleep, yet sleep is hard to assess in free-living environ-
81 ments [1]. Our understanding of how sleep is associated with health and morbidity
82 primarily draws on studies that use self-report sleep diaries, which capture the sub-
83 jective experience [2]. However, sleep diaries have a low correlation with objective
84 device-measured sleep parameters [3, 4]. The accepted standard for sleep measure-
85 ment is laboratory-based polysomnography, which monitors sleep using a range of
86 physical and physiological signals. However, polysomnography is not feasible for use
87 at scale due to its high cost and technical complexity. Instead, wrist-worn accelerom-
88 eters are more viable to deploy in large-scale epidemiological studies because of their
89 portability and low user burden.

90 Despite the popularity of sleep monitoring in consumer and research-grade wrist-
91 worn devices, sleep assessment algorithms are frequently proprietary and validated in
92 small populations, making their measurement validity unclear [5, 6, 7, 8]. Methods
93 for Sleep classification (i.e. defining periods of wake, NREM and REM sleep) pri-
94 marily rely on hand-crafted spatiotemporal features such as device angle, which may
95 not make full use of all the information in the signals. Hence, data-driven methods
96 like deep learning could be advantageous. Furthermore, existing actigraphy-based
97 sleep studies on large health datasets have only focused on the differentiation be-
98 tween sleep and wakefulness [9, 4, 10, 11] without evaluating variations in the stages
99 of sleep.

100 We therefore set out to: (1) develop and internally validate an open-source novel
101 deep learning method to infer sleep stages from wrist-worn accelerometers, (2) ex-
102 ternally validate our proposed algorithm together with existing sleep staging bench-
103 marks, and (3) investigate the association between device-measured overnight sleep
104 duration and efficiency with all-cause mortality.

105 **2. Methods**

106 *2.1. Study design and participants*

107 In our multicentre cohort study, we developed and tested a sleep staging model for
108 accelerometers (SleepNet) using a self-supervised deep recurrent neural network. We
109 designed the model to classify each 30-second window of accelerometry data into
110 one of the three sleep stages, wake, rapid-eye-movement sleep (REM), and non-
111 rapid-eye movement sleep (NREM). Figure 1 illustrates the three main steps in our
112 study: (1) feature extraction from unlabelled free-living data, (2) sleep staging model
113 development, and (3) face validity assessment and health association analysis using
114 the machine learning-estimated sleep parameters.

115 We used the UK Biobank accelerometry dataset [12] for two purposes: learning
116 health-relevant accelerometer features to support the training of the sleep staging
117 model and conducting the downstream health association analyses using the devel-
118 oped sleep staging model.

119 For sleep staging model development, internal validation consisted of two gener-
120 ations of participants from the Raine Study [13, 14] and a sleep patient population
121 from the Newcastle cohort [15]. The Raine Study has followed up roughly 2900 chil-
122 dren since 1989 in Australia. A subset of children (Raine Generation 2, Gen2) at
123 the age of 22 and their parents (Raine Generation 1, Gen1) were invited to undergo
124 one night of laboratory-based polysomnography at Western Australia’s Center for
125 Sleep Science. The external validation consisted of two general populations from Le-
126 icester [16] and Pennsylvania [17]. Detailed population characteristics and inclusion
127 criteria are listed in Supplementary Section 5.

128 *2.2. Accelerometer devices and data preprocessing*

129 Three different devices were used to collect the accelerometry for the included
130 datasets, ActiGraph GT3X, Axivity AX3 and GENEActive Original accelerome-
131 ters. The devices used have been shown to have a high inter-instrument agreement
132 ($> 80\%$) in derived sedentary and sleep-related time estimates in free-living envi-
133 ronments [18]. As for device placement, we selected data from the dominant wrist
134 where possible to be consistent with the UK Biobank protocol.

135 We used the Biobank Accelerometer Analysis Tool [19, 20] to preprocess all the
136 data. The raw tri-axial accelerometry was first resampled into 30 Hz and clipped
137 to $\pm 3 g$. The accelerometry sequence was then divided into consecutive 30-second
138 windows. We considered stationary periods ($x/y/z$ sd < 13 mg) with a duration
139 greater than 60 minutes as non-wear [12]. We further excluded the data that could
140 not be parsed, had unrealistic high values (> 200 mg), or were poorly calibrated.

141 *2.3. Ascertainment of sleep stages via polysomnography*

142 The gold-standard, laboratory-based polysomnography sleep label was aligned
143 with its concurrent accelerometer data as the model ground truth. The polysomnog-
144 raphy labels were scored according to the American Academy of Sleep Medicine
145 (AASM) protocol [21], which divided sleep into five categories: wake, REM, and
146 NREM I, II, and III. In total, 1,157,913 ($\sim 10,000$ hours) sleep windows were used to
147 train the network. The sleep stage distributions were similar across all the datasets
148 except for the Newcastle cohort, which had a greater proportion of wakefulness than
149 the others (Supplementary Figure 1).

150 *2.4. Deep learning analysis of sleep stages from wrist-worn accelerometers*

151 A deep recurrent neural network (SleepNet) was trained to classify the sleep
152 stages for every 30-second window of tri-axial accelerometry data. The SleepNet has
153 three components: a ResNet-17 V2 [22] with 1D convolution for feature extraction, a
154 bi-directional Long-Short-Term-Memory (LSTM) network for temporal dependencies
155 learning [23], and two fully-connected (FC) layers for sleep stage prediction. During
156 training, we provided the SleepNet with five-stage polysomnography labels (wake,
157 REM, and NREM I, II, III). When evaluating the model, we collapsed all the NREM
158 stages into one class for classification (wake/REM/NREM). Similarly, we collapsed
159 all the REM and NREM stages together to classify wake vs sleep.

160 The SleepNet was pre-trained using multi-task self-supervision on the UK Biobank
161 to learn features of human motion dynamics [24]. Multi-task self-supervision auto-
162 matically extracts the features relevant to motion by learning to discriminate dif-
163 ferent spatiotemporal transformations applied to the unlabelled 700,000 person-days

164 of data. Self-supervised pre-training has been shown to help classify human activ-
165 ity recognition not just in healthy but clinical populations [25]. See Supplementary
166 Section 6 for further details of the model development.

167 For internal validation, we used subject-wise five-fold cross-validation on the
168 Raine Gen2, Raine Gen1, and Newcastle cohorts. For external validation, we trained
169 the SleepNet on all the internal datasets and then evaluated its performance on the
170 Leicester and Pennsylvania cohorts. We compared the SleepNet performance with
171 a random forest model that used the hand-crafted spatiotemporal features [20, 26].
172 The random forest feature definitions are listed in Supplementary Table 2.

173 We reported the staging performance in both subject-wise and epoch-to-epoch
174 fashion. Three-class and five-class confusion matrices were plotted for both internal
175 and external validation. Since Cohen Kappa, F1 scores, and balanced accuracies
176 (Supplementary Table 3) are less influenced by class imbalance, they were used to
177 evaluate the overall model. To assess the relationship between the model perfor-
178 mance and population characteristics, we stratified the subject-wise sleep staging
179 performance by age, sex, employment status, income level, body mass index (BMI),
180 presence and severity of sleep apnea using the apnea-hypopnea index (AHI), existing
181 sleep disorders, and neurological disorders where available.

182 Finally, we evaluated the agreement between summary sleep parameters per
183 each night derived from our deep learning method and polysomnography via Bland-
184 Altman plots for the following sleep parameters: total sleep duration, sleep efficiency
185 (proportion of total time asleep when in bed), time awake after sleep onset (WASO),
186 REM duration, NREM duration, REM ratio, NREM ratio. Supplementary Table 4
187 entails the sleep parameter definitions and their calculations.

188 *2.5. Measurements of sleep in 100,000 UK Biobank participants*

189 We obtained the sleep architecture estimates on the UK Biobank by applying
190 SleepNet on the longest overnight sleep windows. Since no concurrent sleep diaries
191 were collected in the UK Biobank, we used a random forest model trained on sleep
192 diaries with Hidden Markov Models smoothing to first obtain time in bed [19, 20].
193 The random forest model achieved 90%+ precision and recall for detecting sleep
194 windows in 152 free-living participants with sleep diaries that asked two questions:

195 “what time did you first fall asleep last night?” and “what time did you wake up
196 (eyes open, ready to get up)?” [20]. We used the sleep window output from the
197 random forest model as a proxy for the time in bed. We then merged any time in
198 bed windows within 60 minutes of one another [27]. Finally, we applied the SleepNet
199 on the longest window over each noon-to-noon interval to estimate the overnight sleep
200 duration. The difference between overnight and total sleep duration is that total sleep
201 duration is a sleep parameter used to assess the agreement between our SleepNet
202 output and polysomnography for model validation. Overnight sleep duration refers
203 to the estimate for the amount of sleep one obtains for a noon-to-noon interval in a
204 free-living environment using a random forest model for sleep window detection and
205 the SleepNet for sleep stage identification.

206 We simulated the effects of random missing data on the participants that had
207 no missing data across seven-days to determine the minimum wear time required for
208 stable weekly sleep parameter estimates (Supplementary Section 7.2). We found that
209 a minimum of 22 hours of wear time per day for at least three days were required to
210 ensure the intra-class correlation was greater than 0.75 between the weekly average
211 sleep duration from incomplete and perfect wear data. Moreover, we tried to mitigate
212 the weekend effect by only including the participants who had at least one weekday
213 and one weekend day during the device wear. Shift workers and participants whose
214 data had daylight saving cross-overs were also excluded, as circadian disruption is
215 not the focus of our paper.

216 Descriptive analyses were performed on the device-measured sleep parameters in
217 the UK Biobank to quantify variations by age, sex, device-measured physical activity
218 level, self-reported chronotype and insomnia symptoms. Estimated marginal means,
219 adjusted for age and sex, were also calculated for different self-rated health groups
220 and self-reported insomnia symptoms.

221 *2.6. Health association analysis*

222 The associations of overnight sleep duration and sleep efficiency with incident
223 mortality were assessed using Cox proportional hazards regression. All-cause mor-
224 tality was determined using death registry data (obtained by UK Biobank from NHS
225 Digital for participants in England and Wales and from the NHS Central Register,

226 National Records of Scotland, for participants in Scotland). Participants were cen-
227 sored at the earliest of UK Biobank’s record censoring date for mortality data (2021-
228 09-30 for participants in England and Wales and 2021-10-31 for participants in Scot-
229 land, with country assigned based on baseline assessment centre). Cox models used
230 age as the timescale, and the main analysis was adjusted for sex, ethnicity, Townsend
231 Deprivation Index, educational qualifications, smoking status, alcohol consumption,
232 and overall activity. See Supplementary Section 7.1 for the full specification of the
233 analysis.

234 *2.7. Role of the funding source*

235 The funders of the study had no role in study design, data collection, data anal-
236 ysis, data interpretation, or writing of the report.

237 **3. Results**

238 *3.1. Comparison to polysomnography*

239 After preprocessing, 1,395 participants were included in the internal validation, and
240 53 participants were included in the external validation. Our proposed deep recurrent
241 neural network (SleepNet) pre-trained with self-supervision achieved the best perfor-
242 mance when compared with other baseline models that used hand-crafted features
243 (Supplementary Table 6).

244 On the internal validation, SleepNet had a mean bias of 8.9 minutes (95% limits of
245 agreement (LoA): -89.0 to 106.9 minutes) for total sleep duration, -18.7 minutes (95%
246 LoA: -130.9 to 93.6 minutes) for REM duration, and 27.6 minutes (95% LoA: -100.6
247 to 155.8 minutes) for NREM duration (Figure 2). In comparison, on the external
248 validation, the mean bias was 34.7 minutes (95% LoA: -37.8 to 107.2 minutes) for to-
249 tal sleep duration, -2.6 minutes (95% LoA: -68.4 to 73.6 minutes) for REM duration,
250 and 32.1 minutes (95% LoA: -54.4 to 118.5 minutes) for NREM duration. Overall,
251 our model tends to underestimate REM and short sleep and overestimate NREM
252 and long sleep. Supplementary Figures 5 to 10 depict the agreement assessments for
253 other sleep parameters on the individual cohorts.

254 The subject-wise performance for both the internal and external validation us-
255 ing the pre-trained SleepNet is shown in Supplementary Table 7. On the pooled

256 internal validation, our model obtained an F1 of 0.75 ± 0.1 in the two-class setting
257 (sleep/wake) and an F1 of 0.57 ± 0.11 in the three-class setting (wake/REM/NREM).
258 The agreement decreased slightly on the external validation with an F1 of $0.67 \pm$
259 0.11 in the two-class setting (sleep/wake) and an F1 of 0.52 ± 0.10 in the three-
260 class setting (wake/REM/NREM). In the Newcastle cohort, for the sleep/wake clas-
261 sification, sensitivity decreased and specificity increased in participants with sleep
262 disorders. No obvious difference was observed in both Raine Gen1 and Gen2 co-
263 horts when the participants were stratified by sex, BMI, AHI, and sleep disorder
264 conditions.(Supplementary Table 8-10).

265 To classify any given window in an epoch-by-epoch fashion, the SleepNet achieved
266 a Kappa score of 0.39 on the internal validation set and a Kappa score of 0.32 on
267 the external validation set in the three-class setting (Supplementary Figure 11).
268 Cohort-specific confusion matrices can be found in Supplementary Figures 12-15.
269 Supplementary Figure 16 visualizes a one-night sample actigram, its ground-truth
270 polysomnography labels, and SleepNet predictions. We used SleepNet to generate
271 all the sleep parameters for the rest of the paper.

272 *3.2. Face validity in the UK Biobank*

273 Before deploying the SleepNet on the UK Biobank, we excluded participants
274 with unusable accelerometer data and participants with missing covariates in the
275 descriptive analysis. We further excluded participants with any prior hospitalisa-
276 tion for cardiovascular disease or cancer in the association analysis (Supplementary
277 Figure 17). In sum, 66,214 participants were included in the final analysis.

278 Table 1 describes the variations in overnight sleep duration, REM and NREM
279 durations, and sleep efficiency across population subgroups in the UK Biobank. Older
280 participants generally slept longer with higher sleep efficiency. Females had a longer
281 overnight sleep duration and NREM but a shorter REM than males. Participants
282 with better self-rated health had longer sleep duration and higher sleep efficiency
283 than those with poor self-rated health. Sleep efficiency was relatively stable across
284 different seasons and days of the week. The correlation coefficients between device-
285 measured sleep parameters during accelerometer wear and self-reported total sleep
286 duration at baseline assessment were all below 0.25 (Supplementary Figure 18). The

287 distributions of device-measured overnight sleep duration tend to have a greater
288 variability for participants who self-reported to have less than 5 or greater than
289 10 hours of total sleep duration (Supplementary Figure 19). Overall, sleep stage
290 distribution was similar for males and females aged between 45 and 75, with NREM
291 sleep fluctuating around 5 hours and REM sleep fluctuating around 2.5 hours per
292 night (Supplementary Figure 20). No major differences were seen between females
293 and males.

294 We found expected sleep-wake patterns in population subgroups. For exam-
295 ple, timing of the sleep opportunity for participants with a self-reported “morning”
296 chronotype was about one hour earlier when compared with those that had a self-
297 reported “evening” chronotype (Figure 3a). We saw similar but shorter phase ad-
298 vance (~30 mins) in participants who were most physically active compared to the
299 participants that were least physically active (Figure 3b). When comparing groups
300 that had a history of self-reported insomnia symptoms versus those who did not,
301 we found that participants with a history of insomnia symptoms were less likely
302 to be in REM sleep on average during the overnight sleep window (Figure 3d and
303 Figure 3c). Participants with a history of self-reported insomnia symptoms tended
304 to have a longer overnight sleep duration but with a lower sleep efficiency (Supple-
305 mentary Figure 21). The sleep architecture for different population subgroups were
306 similar between weekdays and weekends, with a slight phase delay over the weekend
307 (Supplementary Figure 22).

308 *3.3. Association with all-cause mortality*

309 Over 452,327 years of the follow-up, 1,642 mortality events among 66,214 par-
310 ticipants were observed. Short sleepers (<6 hours) had a higher risk of mortality in
311 groups of low sleep efficiency (Hazard ratios (HRs): 1.69; 95% confidence intervals
312 (CIs): 1.28 to 2.24) and high sleep efficiency (HRs: 1.42; 95% CIs: 1.14 to 1.77)
313 compared to participants with normal sleep duration (6 to 7.9 hours, Figure 4). The
314 risk of all-cause mortality appeared to decrease linearly as sleep efficiency increased.
315 However, a non-linear association was observed in the association for overnight sleep
316 duration (Supplementary Figure 23). When further adjusted for BMI, associations
317 of overnight sleep duration and sleep efficiency with all-cause mortality were slightly

318 attenuated (Supplementary Figure 24- 25). Longer overnight sleep duration was not
319 founded to have a higher risk than the reference group in both the main (Supple-
320 mentary Figure 23) and sensitivity analysis (Supplementary Figure 26).

321 **4. Discussion**

322 We have developed, and internally and externally validated a deep-learning method
323 to characterise sleep architecture from a wrist-worn accelerometer with competitive
324 performance against 1,448 nights of laboratory-based polysomnography recordings.
325 When applying our developed method in the UK Biobank in an epidemiological
326 analysis of 66,214 participants, we found that shorter sleep time was associated with
327 an increased risk of all-cause mortality individually regardless of sleep continuity,
328 indexed by sleep efficiency. Our open-source algorithm and the inferred sleep pa-
329 rameters will open the door to future studies on sleep and sleep architecture using
330 large-scale accelerometer databases.

331 Our novel self-supervised deep learning sleep staging method outperformed ex-
332 isting baseline methods that rely on hand-crafted features. The inferred sleep archi-
333 tecture estimates had a fair agreement ($\kappa = 0.39$) with the polysomnography ground
334 truth on the internal validation [28]. Unlike previous work in sleep classification
335 methods that depended on hand-crafted features [26, 29], our proposed method au-
336 tomatically extracted the features using self-supervision, hence removing the need for
337 manual engineering. Even for sleep/wake classification, SleepNet achieved compa-
338 rable results to a systematic evaluation of eight state-of-the-art sleep algorithms [8]
339 in the Newcastle dataset. However, our work offers a more robust evaluation and
340 identifies the upper limit of using accelerometry for sleep classification by developing
341 a model with one of the largest multicentre datasets with polysomnography ground
342 truth, at least ten times the size of existing studies.

343 In the subsequent epidemiological analysis, we found a clear association between
344 short overnight sleep duration with increased risk of all-cause mortality in both good
345 and poor sleepers defined by sleep efficiency. Short overnight sleep duration has been
346 linked with mortality outcomes in self-report and actigraphy-based studies [30, 31].
347 However, few studies have investigated the joint effect of sleep duration and efficiency.
348 One recent study has suggested that participants with short and long total sleep time

349 had an increased risk after accounting for sleep efficiency [32]. However, our analysis
350 did not find that long overnight sleep duration was associated with increased risk,
351 potentially because we did not include daytime naps in our measurement of overnight
352 sleep duration. Daytime napping has been found to be associated with an increased
353 risk of cardiovascular events and deaths in those with longer nighttime sleep [33]. We
354 did not find a U-shape association between device-measured sleep and mortality that
355 has been suggested by other smaller studies [30]. Instead, our data are supportive
356 of adverse associations with short sleep duration only, which is concordant with pre-
357 clinical human and animal studies [34].

358 This study has several strengths, including the analysis of sleep architecture
359 in a large, prospective Biobank with longitudinal follow-up. Compared with self-
360 reported sleep questionnaires that only captured sleep duration to the nearest hour,
361 actigraphy-based methods like ours can provide more fine-grained sleep duration
362 and efficiency estimates. The extensive multicentre evaluation of the sleep classifi-
363 cation allowed for the characterisation of the measurement uncertainty and a less
364 biased interpretation of the health association analysis. Sleep stage identification
365 from actigraphy is highly challenging, especially for wake periods in bed that are not
366 characterised by wrist movement. With the proposed SleepNet, we could obtain sleep
367 architecture estimates for population health inference after evaluating the face va-
368 lidity of the sleep parameters in the UK Biobank. While future work might improve
369 sleep staging performance by incorporating additional physiological signals, such as
370 electrocardiogram, to improve sleep staging performance, multi-modal sensor signals
371 are not yet available for population-scale studies with longitudinal follow-up beyond
372 a few years [35]. Despite our best efforts to include diverse validation cohorts from
373 different centres, the included datasets mainly consist of healthy populations from a
374 Caucasian ethnic background. Validation in populations with chronic diseases and
375 different ethnic backgrounds would aid in quantifying the measurement uncertainty.

376 In this work, we have developed and validated an open-source sleep staging
377 method that substantially improves the ability to measure sleep characteristics with
378 wrist-worn accelerometers in large biomedical datasets. Using the sleep parameters
379 generated by our model, we demonstrated that shorter overnight sleep was associ-
380 ated with a higher risk of all-cause mortality in both good and poor sleepers. Our

381 proposed method provides the community with a rich set of new measurements to
382 study how sleep parameters are longitudinally associated with clinical outcomes.

383 *Data sharing*

384 The data for the Newcastle cohort is available from direct download via <https://zenodo.org/record/1160410#.Y-065i-11qs>. The data for other cohorts can
385 be requested by contacting the corresponding host institute. All the sleep staging
386 models and analysis scripts are freely available for academic use on GitHub: <https://github.com/OxWearables/asleep>.
387
388

389 *Contributions*

390 HY, KM, JM, LS, PE, SD, and AD conceptualised and designed the study. TP,
391 MB, PG, AR, JM, LS, and PS did the data curation of the accelerometers and
392 polysomnography data. HY, TP, and RW did the formal analysis and validation.
393 DB, SK and AD provided supervision to HY and RW. HY wrote the manuscript,
394 and all the authors contributed to the review & editing process. HY and RW had
395 direct access to the summary statistics and verified the findings.

396 *Acknowledgments*

397 We would like to thank Andrew Creagh, Angel Wong, Scott Small, and Alaina
398 Shreves for their input on the revision of this manuscript. We would also like to
399 thank Andrew Creagh for his feedback in creating the graphic illustrations.

400 **Main text tables and figures**

Table 1: Overall sleep parameters by participant characteristics in the UK Biobank (mean \pm SD) for overnight sleep duration, non-rapid-eye-movement sleep (NREM), rapid-eye-movement sleep (REM), and sleep efficiency.

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

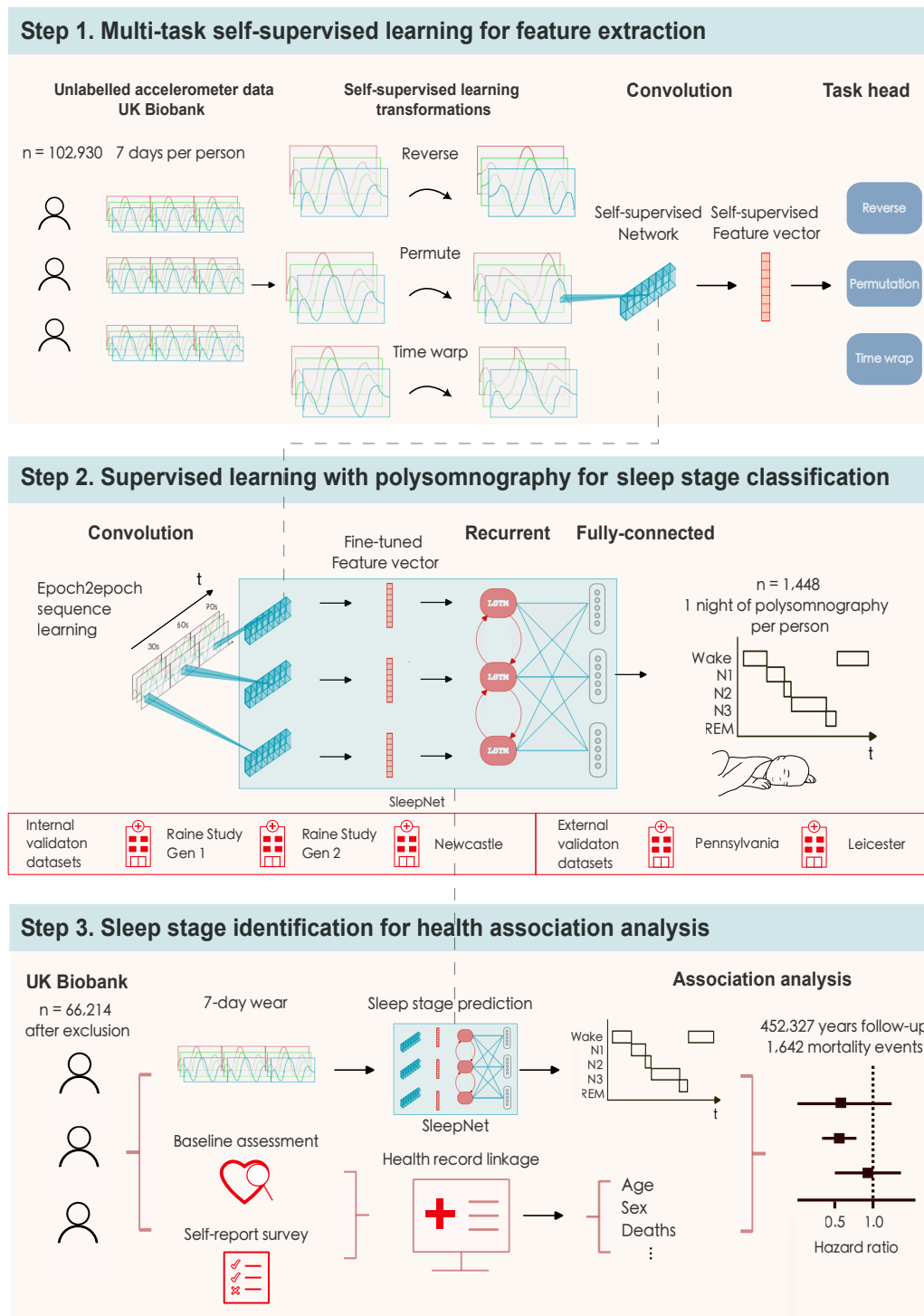


Figure 1: **The SleepNet development pipeline**¹⁴: 1. We use multi-task self-supervised learning to obtain a feature extractor by learning from 700,000 person-days of tri-axial accelerometry data in the UK Biobank. 2. The pre-trained feature extractor was then fine-tuned with a deep recurrent network to train a sleep-stage classifier using polysomnography as the ground truth. 3. We deploy the sleep prediction model on the UK Biobank and investigate the association between device-measured sleep and mortality outcomes.

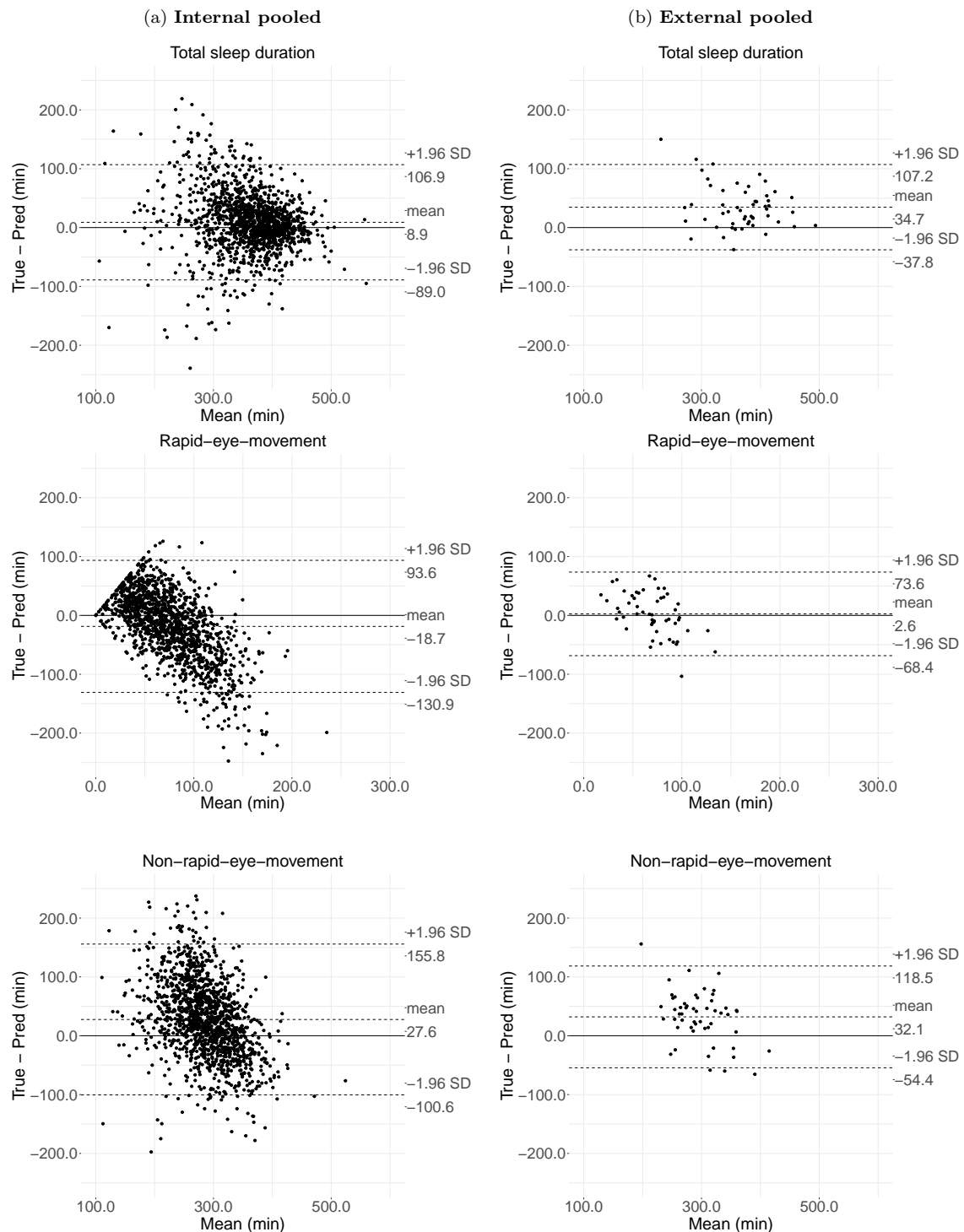


Figure 2: Agreement assessment via Bland-Atman plot for total sleep duration, rapid eye movement sleep (REM) duration, and non-rapid eye movement sleep (NREM) duration on internal and external validation. The internal validation consists of 1,373 polysomnography nights from the Raine Study and the Newcastle cohort, whereas the external validation consists of 53 polysomnography nights from the Leicester and Pennsylvania cohorts.

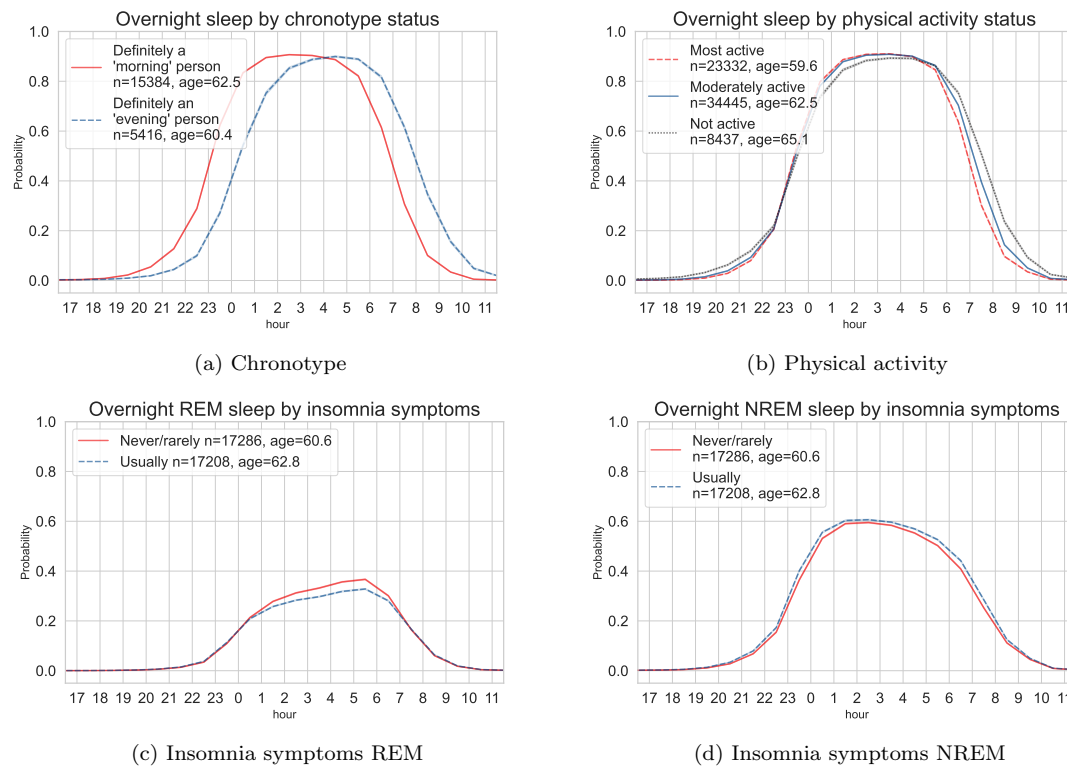


Figure 3: Device-measured sleep probability trajectories throughout the day for the UK Biobank participants. Top: variations of the average overnight sleep probability for the participants with self-reported “morning” and “evening” chronotype (a) and the overnight sleep distributions across thirds of 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.

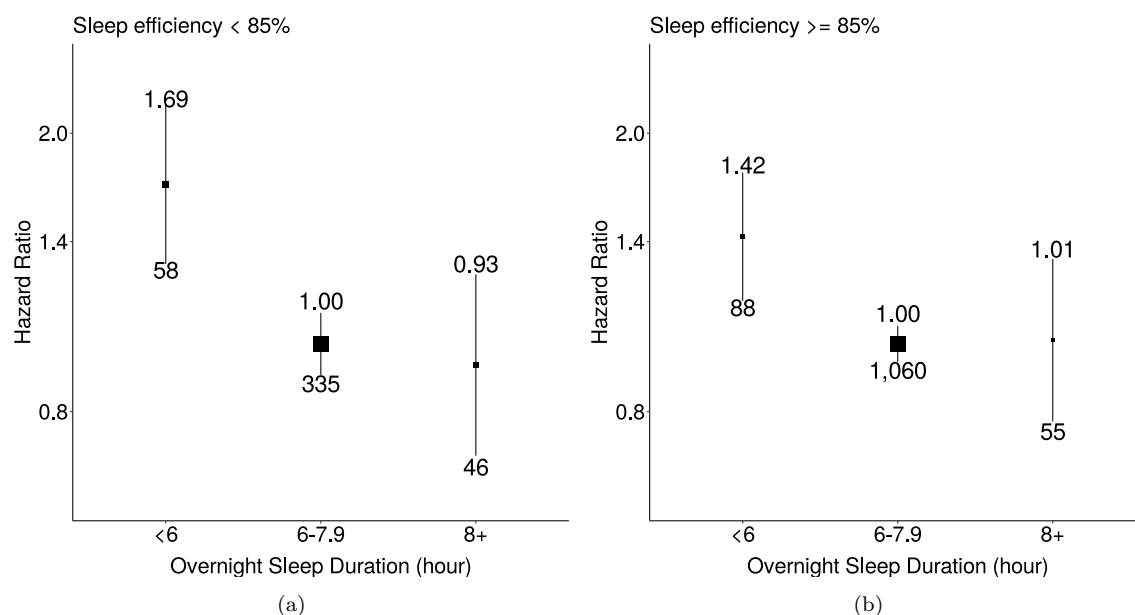


Figure 4: **Associations of overnight sleep duration with all-cause mortality for groups with low and high sleep efficiency.** The model used 1,642 events among 62,214 participants. We used age as the timescale and adjusted for sex, ethnicity, Townsend Deprivation Index of baseline address (split by 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.

401 References

- 402 [1] Nicholas Meyer et al. “Circadian rhythms and disorders of the timing of sleep”.
403 In: *The Lancet* (2022).
- 404 [2] Jane E Ferrie et al. “Sleep epidemiology—a rapidly growing field”. In: *International Journal of Epidemiology* 40.6 (Dec. 2011), pp. 1431–1437. ISSN: 0300-5771. DOI: 10.1093/ije/dyr203. eprint: <https://academic.oup.com/ije/article-pdf/40/6/1431/2407775/dyr203.pdf>. URL: <https://doi.org/10.1093/ije/dyr203>.
405
406
407
408
- 409 [3] Michelle A Short et al. “The discrepancy between actigraphic and sleep diary
410 measures of sleep in adolescents”. In: *Sleep Medicine* 13.4 (2012), pp. 378–384.
- 411 [4] Michael Wainberg et al. “Association of accelerometer-derived sleep measures
412 with lifetime psychiatric diagnoses: A cross-sectional study of 89,205 partici-
413 pants from the UK Biobank”. In: *PLoS Medicine* 18.10 (2021), e1003782.
- 414 [5] Janna Mantua, Nickolas Gravel, and Rebecca Spencer. “Reliability of sleep
415 measures from four personal health monitoring devices compared to research-
416 based actigraphy and polysomnography”. In: *Sensors* 16.5 (2016), p. 646.
- 417 [6] Alexander J Boe et al. “Automating sleep stage classification using wireless,
418 wearable sensors”. In: *NPJ Digital Medicine* 2.1 (2019), pp. 1–9.
- 419 [7] Jaime K Devine et al. “Validation of Zulu watch against polysomnography and
420 actigraphy for on-wrist sleep-wake determination and sleep-depth estimation”.
421 In: *Sensors* 21.1 (2020), p. 76.
- 422 [8] Matthew R Patterson et al. “40 years of actigraphy in sleep medicine and
423 current state of the art algorithms”. In: *NPJ Digital Medicine* 6.1 (2023), p. 51.
- 424 [9] Aiden Doherty et al. “GWAS identifies 14 loci for device-measured physical
425 activity and sleep duration”. In: *Nature Communications* 9.1 (2018), pp. 1–8.
- 426 [10] Samuel E Jones et al. “Genetic studies of accelerometer-based sleep measures
427 yield new insights into human sleep behaviour”. In: *Nature Communications*
428 10.1 (2019), pp. 1–12.

- 429 [11] Machiko Katori et al. “The 103,200-arm acceleration dataset in the UK Biobank
430 revealed a landscape of human sleep phenotypes”. In: *Proceedings of the Na-*
431 *tional Academy of Sciences* 119.12 (2022), e2116729119.
- 432 [12] Aiden Doherty et al. “Large scale population assessment of physical activity
433 using wrist worn accelerometers: the UK biobank study”. In: *PloS One* 12.2
434 (2017), e0169649.
- 435 [13] Leon Straker et al. “Cohort profile: the Western Australian pregnancy cohort
436 (Raine) study—Generation 2”. In: *International Journal of Epidemiology* 46.5
437 (2017), 1384–1385j.
- 438 [14] Manon L Dontje, Peter Eastwood, and Leon Straker. “Western Australian preg-
439 nancy cohort (Raine) study: generation 1”. In: *BMJ open* 9.5 (2019), e026276.
- 440 [15] Vincent van Hees, Sarah Charman, and Kirstie Anderson. *Newcastle polysomnog-*
441 *raphy and accelerometer data*. Version 1.0. Zenodo, Jan. 2018. DOI: 10.5281/
442 [zenodo.1160410](https://doi.org/10.5281/zenodo.1160410). URL: <https://doi.org/10.5281/zenodo.1160410>.
- 443 [16] Tatiana Plekhanova et al. “Validation of an automated sleep detection algo-
444 rithm using data from multiple accelerometer brands”. In: *Journal of Sleep*
445 *Research* (2022).
- 446 [17] Enda M Byrne et al. “Genetic correlation analysis suggests association between
447 increased self-reported sleep duration in adults and schizophrenia and type 2
448 diabetes”. In: *Sleep* 39.10 (2016), pp. 1853–1857.
- 449 [18] Jairo H Migueles et al. “Equivalency of four research-grade movement sensors
450 to assess movement behaviors and its implications for population surveillance”.
451 In: *Scientific Reports* 12.1 (2022), pp. 1–9.
- 452 [19] Matthew Willetts et al. “Statistical machine learning of sleep and physical
453 activity phenotypes from sensor data in 96,220 UK Biobank participants”. In:
454 *Scientific Reports* 8.1 (2018), pp. 1–10.
- 455 [20] Rosemary Walmsley et al. “Reallocation of time between device-measured
456 movement behaviours and risk of incident cardiovascular disease”. In: *British*
457 *Journal of Sports Medicine* 56.18 (2022), pp. 1008–1017.

- 458 [21] Richard B Berry et al. “Rules for scoring respiratory events in sleep: update
459 of the 2007 AASM manual for the scoring of sleep and associated events: de-
460 liberations of the sleep apnea definitions task force of the American Academy
461 of Sleep Medicine”. In: *Journal of Clinical Sleep Medicine* 8.5 (2012), pp. 597–
462 619.
- 463 [22] Kaiming He et al. “Identity mappings in deep residual networks”. In: *European
464 Conference on Computer Vision*. Springer. 2016, pp. 630–645.
- 465 [23] Zhiheng Huang, Wei Xu, and Kai Yu. “Bidirectional LSTM-CRF models for
466 sequence tagging”. In: *arXiv preprint arXiv:1508.01991* (2015).
- 467 [24] Hang Yuan et al. “Self-supervised Learning for Human Activity Recognition
468 Using 700,000 Person-days of Wearable Data”. In: *arXiv preprint arXiv:2206.02909*
469 (2022).
- 470 [25] Andrew P Creagh et al. “Digital health technologies and machine learning aug-
471 ment patient reported outcomes to remotely characterise rheumatoid arthritis”.
472 In: *medRxiv* (2022).
- 473 [26] Kalaivani Sundararajan et al. “Sleep classification from wrist-worn accelerom-
474 eter data using random forests”. In: *Scientific Reports* 11.1 (2021), pp. 1–10.
- 475 [27] Vincent Theodoor van Hees et al. “Estimating sleep parameters using an ac-
476 celerometer without sleep diary”. In: *Scientific reports* 8.1 (2018), p. 12975.
- 477 [28] Mary L McHugh. “Interrater reliability: the kappa statistic”. In: *Biochemia
478 medica* 22.3 (2012), pp. 276–282.
- 479 [29] Michelle L Trevenen et al. “Using hidden Markov models with raw, triaxial
480 wrist accelerometry data to determine sleep stages”. In: *Australian & New
481 Zealand Journal of Statistics* 61.3 (2019), pp. 273–298.
- 482 [30] Jiawei Yin et al. “Relationship of sleep duration with all-cause mortality and
483 cardiovascular events: a systematic review and dose-response meta-analysis of
484 prospective cohort studies”. In: *Journal of the American Heart Association* 6.9
485 (2017), e005947.

- 486 [31] Osamu Itani et al. “Short sleep duration and health outcomes: a systematic
487 review, meta-analysis, and meta-regression”. In: *Sleep Medicine* 32 (2017),
488 pp. 246–256.
- 489 [32] Yannis Yan Liang et al. “Joint Associations of Device-measured Sleep Dura-
490 tion and Efficiency with All-cause and Cause-specific Mortality: A Prospective
491 Cohort Study of 90 398 UK Biobank Participants”. In: *The Journals of Geron-
492 tology: Series A* (2023), glad108.
- 493 [33] Chuangshi Wang et al. “Association of estimated sleep duration and naps with
494 mortality and cardiovascular events: a study of 116 632 people from 21 coun-
495 tries”. In: *European Heart Journal* 40.20 (2019), pp. 1620–1629.
- 496 [34] Shahrads Taheri. “Sleep and cardiometabolic health—not so strange bedfel-
497 lows”. In: *The Lancet Diabetes & Endocrinology* (2023).
- 498 [35] Jessica R Golbus et al. “Wearable device signals and home blood pressure
499 data across age, sex, race, ethnicity, and clinical phenotypes in the Michi-
500 gan Predictive Activity & Clinical Trajectories in Health (MIPACT) study:
501 a prospective, community-based observational study”. In: *The Lancet Digital
502 Health* 3.11 (2021), e707–e715.

503 **Supplements**

504 **List of Tables**

505 1 **Characteristics of the datasets used for internal validation, external validation and health association analyses** “Patient” indicates whether a cohort
506 consists of sleep patients in a clinic. 26
507
508 2 **Hand-crafted features** 29
509
510 3 **Model performance metric definitions (TP: true positive; TN: true negative; FP: false positive; FN: false negative)** 30
511
512 4 **Sleep parameter definitions: total sleep duration (TSD), rapid-eye-movement (REM), non-rapid-eye-movement (NREM), sleep onset latency (SOL),**
513 **wake after sleep onset (WASO), and sleep efficiency (SE).** 31
514
515 5 **Code table for UK Biobank variables used in the study.** 32
516
517 6 **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
518 is the deep recurrent network without pre-training. SleepNet-SSL is the network
519 pre-trained using self-supervision. Five-fold subject-wise performance metrics (mean
520 \pm SD) are reported using the internal validation data. REM: rapid-eye-movement
521 sleep, NREM: non-rapid-eye-movement sleep, Kappa score: κ 36
522
523 8 **Model characteristics on the internal validation datasets (wake versus sleep):** subject-wise performance metrics (mean \pm SD) are reported using the internal
524 validation data. Sen: sensitivity, Spe: specificity. Wake is the negative class
525 and the sleep is the positive class when calculating model performance. 38
526
527 9 **Model characteristics on the internal validation datasets (wake versus REM versus NREM):** subject-wise performance metrics (mean \pm SD) are reported
528 using the internal validation data. REM: rapid-eye-movement, NREM: non-rapid-
529 eye-movement, Kappa score: κ 39
530
531 10 **Model characteristics on the internal validation datasets (wake versus REM versus NREM I, II, III):** subject-wise performance metrics (mean \pm SD)
532 are reported using the internal validation data. REM: rapid-eye-movement, NREM:
533 non-rapid-eye-movement, Kappa score: κ 40

534 **List of Figures**

535 1 **Sleep stage distribution for all the datasets used.** 27

536	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.	34
537			
538			
539			
540			
541	3	The distribution of non-wear time for all the participants from the UK Biobank.	35
542			
543	4	Receiver operating characteristics curves for two-class (wake/sleep) and three-class (wake/REM/NREM) settings on the internal validation dataset using our best performing model self-supervised SleepNet. REM: rapid-eye-movement sleep, NREM: non-rapid-eye-movement sleep.	37
544			
545			
546			
547	5	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).	42
548			
549			
550	6	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).	43
551			
552			
553	7	Agreement assessment via Bland-Altman plots for internal validation: non-rapid-eye-movement sleep (NREM) ratio, and rapid-eye-movement sleep (REM) ratio.	44
554			
555			
556	8	Agreement assessment via Bland-Altman plots for external validation: non-rapid-eye-movement sleep (NREM) ratio, and rapid-eye-movement sleep (REM) ratio.	45
557			
558			
559	9	Agreement assessment via Bland-Altman plots for internal validation: wake after sleep onset (WASO), and sleep efficiency (SE).	46
560			
561	10	Agreement assessment via Bland-Altman plots for internal validation: wake after sleep onset (WASO), and sleep efficiency (SE).	47
562			
563	11	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.	48
564			
565			
566			
567	12	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.	49
568			
569			
570			
571			

572	13	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.	50
573			
574			
575			
576			
577	14	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.	51
578			
579			
580			
581			
582	15	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.	51
583			
584			
585			
586			
587	16	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.	52
588			
589			
590			
591			
592	17	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.	54
593			
594			
595			
596	18	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.	55
597			
598			
599			
600	19	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.	56
601			
602			
603			
604	21	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.	57
605			
606			
607			

608	22	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). Areas of squares represent the inverse of the variance of the log risk. And the I bars denote the 95% confidence interval for the floated risks.	58
609			
610			
611			
612			
613			
614			
615			
616			
617	24	Associations of overnight sleep duration with all-cause mortality for groups with low and high sleep efficiency additionally adjusted for body mass index. The model used 1,642 events among 62,214 participants. We used age 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.	60
618			
619			
620			
621			
622			
623			
624			
625			
626	25	Associations of overnight sleep duration (a) and sleep efficiency (b) with all-cause mortality additionally adjusted for body mass index. The model used 1,642 events among 62,214 participants. We used age 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), 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.	61
627			
628			
629			
630			
631			
632			
633			
634			
635	26	Associations of device-measured overnight sleep duration and all-cause mortality with greater granularity. The model used 1,642 events among 62,214 participants. We used age 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), and 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.	62
636			
637			
638			
639			
640			
641			
642			

Table 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	✗	[1]
Raine Gen1	865	56.7 ± 5.6	Dom wrist	GT3X	✗	[2]
Raine Gen2	795	22.1 ± 0.6	Dom wrist	GT3X	✗	[2]
Newcastle	28	44.9 ± 14.9	Both wrists	GENEActiv	✓	[3]
Leicester	30	30.8 ± 6.7	Both wrists	Axivity	✗	[4]
Pennsylvania	22	22.8 ± 4.5	Non-dom wrist	Axivity	✗	[5]

643 5. Datasets

644 *Raine Study.* The Raine Study has followed up roughly 2900 children since 1989 in
645 Australia. A subset of children (Raine Gen2, 50% females) at the age of 22 and their
646 parents (Raine Gen1, 57% females) were invited to undergo one night of laboratory-
647 based polysomnography at Western Australia’s Center for Sleep Science [2, 6]. Every
648 participant was instructed to wear an ActiGraph GT3X device on the dominant
649 wrist. Earlier GT3X firmware would enter an idle mode to save the battery when no
650 sufficient movement was detected, so we only included participants with no missing
651 data for the Raine Gen2 cohort.

652 *Newcastle.* The Newcastle dataset recruited 28 adult patients (39% females) for a
653 one night laboratory-based polysomnography assessment in Newcastle upon Tyne,
654 UK, as part of their routine clinical visit [3]. During the polysomnography recording,
655 the participants wore two GENEActive devices, one on each wrist. The sampling
656 frequency for the wristbands was set to 85.7 Hz.

657 *Leicester.* Thirty healthy volunteers (63% females and 73% white) wore three de-
658 vices: GENEActive, Axivity AX3, and ActiGraph GT9X on each wrist during one
659 night of laboratory-based polysomnography assessment [4]. The relative position of
660 the devices was randomly allocated for each participant. The devices were set to
661 record at 100 Hz. During the lab visit, when the participants wished to go to bed,
662 the recording was started. The sleep episodes usually ended between 6 am and 7

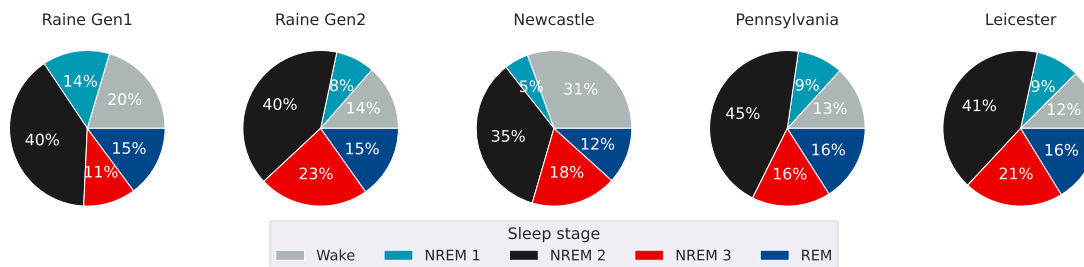


Figure 1: Sleep stage distribution for all the datasets used.

663 am the following morning. We cleaned up the recording sessions such that every
664 recording would start from "light off" and end at "light off" to ensure comparability.

665 *Pennsylvania.* The Pennsylvania dataset consists of 22 healthy sleepers who had one-
666 night of laboratory-based polysomnography assessment at the University of Penn-
667 sylvania Center for sleep [5]. The participants were asked to wear an Axivity device
668 on the non-dominant wrist during the polysomnography session.

669 *UK Biobank.* The UK Biobank is a longitudinal cohort study that recruited 500,000
670 adults from the UK [7]. A subset of the participants was invited to wear an Axivity
671 device on the dominant wrist for one week in a free-living environment [1]. The sam-
672 pling rate was set to 100 Hz. Roughly 100,000 participants (56% females) consented
673 and participated in the accelerometry study. Other than the accelerometry data, a
674 rich set of biomedical information was also collected on the study participants, such
675 as health record linkage, self-reported questionnaire and genetic data.

676 We preprocessed all the datasets by manual quality checks for unrealistic high
677 values for accelerometry (>200 mg), parsing successes, polysomnography alignment,
678 and visual inspection.

679 6. Model development

680 6.1. Self-supervised pre-training

681 To obtain a feature extractor by leveraging a large amount of unlabelled data
682 from the UK Biobank, we applied multi-task self-supervised learning following [8].
683 In self-supervision pre-training, the model was designed to discriminate whether a

684 set of binary transformations have been applied to the signal. We selected reversal,
685 permutation, and time-warping as potential self-supervised learning because they are
686 suitable for learning spatiotemporal patterns.

687 The feature extractor was built on top of ResNet-17 V2 [9] with 1D convolution,
688 in total, with 10M parameters. Each feature vector is of size 1024. We used cross-
689 entropy as the cost function, with each task having the same weight to balance the
690 features learned from each task. In the training procedure, we applied axis swap and
691 rotation as data augmentation to obtain a representation that is orientation invariant.
692 During training time, we used a batch size of 2000 as a larger batch size was found
693 to produce features with better quality. Adam [10] was used for optimisation with a
694 learning rate of 1e-3. We distributed the training across 4 Tesla V100-SXM2 GPUs
695 with 32GB. Early-stopping with a patience of five steps was used to avoid overfitting.
696 It took about 420 GPU hours for the model to converge. More details can be found
697 in [8].

698 *6.2. SleepNet training*

699 We used the pre-trained ResNet from self-supervision as the base model for fea-
700 ture extraction. Then, we appended two layers of Bi-directional Long-Short-Term-
701 Memory (LSTM) layers of 1024 units to learn the temporal dependencies of the
702 model [11]. In the end, we had two fully-connected layers of 512 units to generate the
703 sleep stages. The model was trained to discriminate five sleep stages directly (wake,
704 N1, N2, N3 and REM). To obtain the three-class output, we combined NREM I, II,
705 and III into the NREM class. Likewise, we combined NREM I, II, III and NREM
706 into the sleep class to obtain the two-class output.

707 The learning rate was set to be 1e-3. We also set the gradient clapping to 1 to
708 avoid exploding gradient for LSTM. We used weighted Cross-Entropy as the objective
709 function and weighted each class with the inverse of its frequency to account for the
710 imbalanced dataset. We also used rotation and axis swap to augment the input data
711 to obtain a direction-invariant model. Each training mini-batch consisted of five
712 participants. For each individual, we selected four 1.5-hour sequences with random
713 starting points to avoid overfitting to the study protocol, where the beginning and
714 the end of the sequence are always the “wake” class. The model was trained on a

715 Tesla V100-SXM2 with 32GB of memory. It took about 12 hours for the model to
 716 converge. The model performance was reported using five-fold subject-wise cross-
 717 validation. We first split the data into train/test with a ratio of 8:2. We further split
 718 the train set into train/validation with a ratio of 8:2. We used early stopping with a
 719 patience of ten steps to avoid overfitting on the validation set in each cross-validation
 720 fold.

Table 2: **Hand-crafted features**

Handcrafted features	Notes
Sleep features [12]	
ENMO	All sleep features have 12 derived variables: mean, std, min, max, entropy 20 bins (low resolution), entropy 200 bins (high resolution), median absolute derivation, and mean difference between neighbouring windows.
Angle Z	
Locomotor inactivity during sleep	
Axis features [13]	
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 [13]	$\text{norm} = \sqrt{x^2 + y^2 + z^2}$
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

Table 3: Model performance metric definitions (TP: true positive; TN: true negative; FP: false positive; FN: false negative)

Metric	Definition
Precision	$\frac{TP}{TP+FP}$
Sensitivity/Recall	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{TN+FP}$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
F1	$2 \times \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$
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 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. Our 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	$\frac{\text{NREM duration}}{\text{TSD}}$

721 7. UK Biobank analysis

Table 5: Code table for UK Biobank variables used in the study.

Variable	Code name
Month of birth	p52
Year of birth	p34
Device wear time	p90010
Sex	p31
Ethnicity	p21000
Smoking status	p20116
Alcohol consumption	p1558
Education qualification	p6138
Body mass index	p21001
Employment status	p6142
Overall health rating	p2178
Self-reported total sleep duration	p1160
Townsend Deprivation Index	p189
Overall accelerometry average	p90012
Self-reported trouble falling/ staying asleep	p1200

722 The UK Biobank variable codes are shown in Table 5. We used the month of birth
723 (p52) and year of birth (p34) along with device wear time (p90010) to compute the
724 age at wear time. Participants were asked about their insomnia symptoms history
725 (p1200) by “Do you have trouble falling asleep at night or do you wake up in the
726 middle of the night?”. Four responses were possible: “never/rarely”, “sometimes”,
727 “usually”, and “prefer not to answer”.

728 7.1. Sleep and all-cause mortality

729 The relationship between machine learning-derived sleep architecture estimates
730 and all-cause mortality was assessed using association analyses. The main analysis
731 split the participants into six groups stratified by sleep efficiency cut-off with clinical
732 relevance. Then, five groups were created based on exact hour cut-offs in line with
733 sleep recommendation guidelines for overnight sleep duration [14]. Four groups were
734 created based on percentage cut-offs of clinical relevance for sleep efficiency [15]. In

735 the sensitivity analysis, seven sleep groups were created on exact hour cut-offs to
736 capture the variations in participants with lower and higher sleep durations.

737 Mortality was determined using death registry data (obtained by UK Biobank
738 from NHS Digital for participants in England and Wales and from the NHS Central
739 Register, National Records of Scotland, for participants in Scotland). For survival
740 analyses, participants were censored at the earliest of UK Biobank's record censor-
741 ing date for mortality data (2021-09-30 for participants in England and Wales and
742 2021-10-31 for participants in Scotland, with country assigned based on baseline as-
743 sessment centre) and a record of loss to linked health record follow-up (field 191; 2
744 participants only).

745 In addition to the exclusions described for the analyses above, for prospective
746 analyses for incident mortality we further excluded the participants if they had a
747 prior hospitalisation for restless syndrome, any cardiovascular disease or cancer (a
748 hospital episode with primary diagnosis G473, I00-I99 or C00-C99).

749 Models used age as the timescale, and the main analysis was adjusted for sex
750 (male/female), ethnicity (white/non-white), Townsend Deprivation Index of baseline
751 address (split by quarter in the study population), educational qualifications (school
752 leaver, further education, higher education), smoking status (never smoker, ex-
753 smoker, current smoker), alcohol consumption (never, <3 times/week, 3+ times/week),
754 and overall activity (measured in milli-gravity units). An additional analysis further
755 adjusted for BMI (categorised as <18.5 kg/m², 18.5-24.9 kg/m², 25.0-29.9 kg/m²,
756 30+ kg/m²). See Supplementary Table 5 for UK Biobank fields).

757 Results are presented with their 95% confidence intervals. The Floating Absolute
758 Risk approach was used to calculate confidence intervals for the estimate in each
759 group, without contrast to a reference group [16, 17, 18].

760 In statistical testing using the Grambsch-Therneau test with the Kaplan-Meier
761 transformation, there was some evidence that the joint associations of overnight
762 sleep duration and sleep efficiency with incident mortality violated the proportional
763 hazards assumption (with age as the timescale). However, assessing associations
764 at younger (< 65 years) and older (\geq 65 years) ages did not suggest substantially
765 differing associations by age, and so the overall hazard ratios are presented.

766 7.2. Reliability assessment for device wear time exclusion criterion

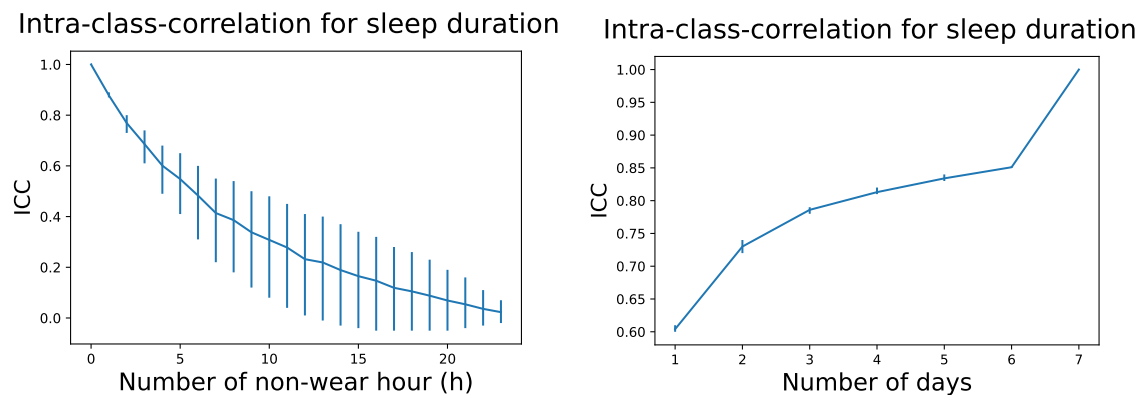


Figure 2: **How the intraclass correlation coefficient (ICC) changes with respect to the non-wear hours (h) (left) and the number of wear days (right) in a reliability simulation using data from 27,870 participants that had zero non-wear time across a seven-day period.** Mean and 95% confidence intervals are plotted.

767 We needed to discard participants with too much non-wear time to obtain a stable
768 sleep duration estimate. Ideally, all the participants would have perfect seven-day
769 device wear, which was not the case. Thus, we needed to determine the minimum
770 wear time for seven days so that there is a high agreement between sleep duration
771 computed for participants with perfect data and those computed for participants
772 with missing data. To do this, we first selected a subset of 27,870 participants who
773 did not have any non-wear time during the seven-day window. Then, we simulated
774 the missing data by randomly removing one hour from each day or one whole day of
775 data from each week from their recordings. We increased the amount of simulated
776 missing data step-wise until all the data was removed. Then, we compared weekly
777 mean sleep durations computed on data before and after removing the simulated
778 missing periods.

779 We used the intraclass correlation coefficient (ICC) to determine the acceptable
780 missing time threshold. We selected two-way random-effects, single rater with an ab-
781 solute agreement, ICC2, to reflect the reliability of our sleep duration measurement
782 if we have missing data in the measurements [19]. Supplementary Figure 2 depicts

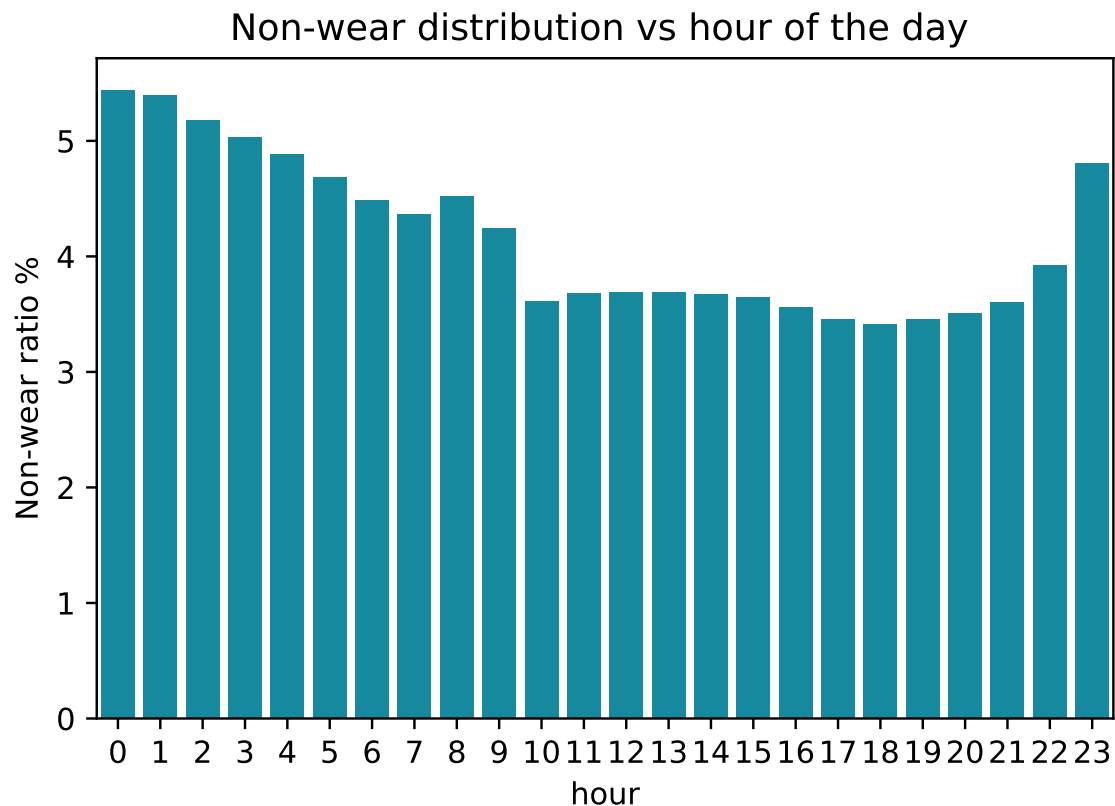


Figure 3: **The distribution of non-wear time for all the participants from the UK Biobank.**

783 the ICC mean and 95% confidence intervals for the missing non-wear hour (Supple-
784 mentary Figure 2 Left) and missing days (Supplementary Figure 2 Right). We used
785 an ICC of 0.75 threshold when deciding the acceptable device wear range. According
786 to the 0.75 cut-off, a maximum of two non-wear hours per day and a minimum of
787 three days per week are suitable for obtaining stable measurements of sleep duration.

788 8. Additional Results

789 8.1. Model performance

Table 6: **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 \pm 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		
	κ	Accuracy	F1	κ	Accuracy	F1
Random forest [13, 12]	0.472 \pm 0.192	0.756 \pm 0.102	0.729 \pm 0.104	0.290 \pm 0.149	0.507 \pm 0.072	0.464 \pm 0.072
SleepNet	0.468 \pm 0.193	0.757 \pm 0.103	0.727 \pm 0.105	0.313 \pm 0.162	0.576 \pm 0.112	0.535 \pm 0.106
SleepNet-SSL	0.511\pm0.196	0.775\pm0.105	0.750\pm0.107	0.375\pm0.163	0.625\pm0.116	0.573\pm0.116

790 Supplementary Table 6 shows the model performance comparison between the
791 random forest model that used hand-crafted features and our proposed SleepNet
792 on the internal validation. SleepNet pre-trained with self-supervision had the best
793 performance in both the two-class ($\kappa = 0.511 \pm 0.196$) and three-class settings ($\kappa =$
794 0.375 ± 0.163). In addition, the area under the receiver operating characteristic curve
795 for the best SleepNet model is 0.88 for the two-class setting and 0.81 for the three-class
796 setting (Supplementary Figure 4).

Table 7: **Subject-wise performance sleep classification validation using our 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

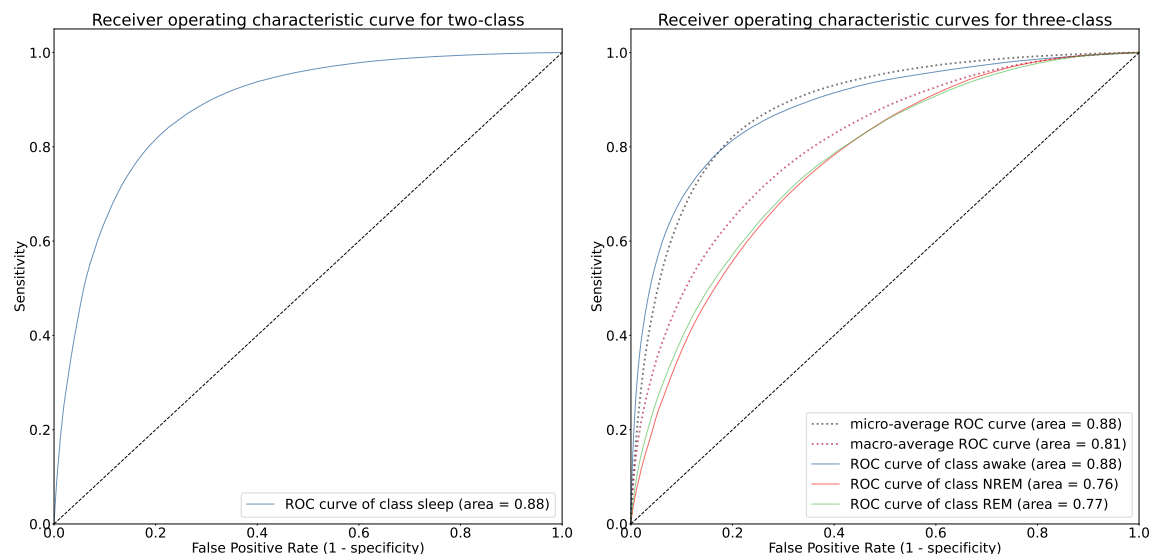


Figure 4: **Receiver operating characteristics curves for two-class (wake/sleep) and three-class (wake/REM/NREM) settings on the internal validation dataset using our best performing model self-supervised SleepNet.** REM: rapid-eye-movement sleep, NREM: non-rapid-eye-movement sleep.

Table 8: **Model characteristics on the internal validation datasets (wake versus sleep):** subject-wise performance metrics (mean \pm 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.

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

Table 9: **Model characteristics on the internal validation datasets (wake versus REM versus NREM):** subject-wise performance metrics (mean \pm SD) are reported using the internal validation data. REM: rapid-eye-movement, NREM: non-rapid-eye-movement, Kappa score: κ .

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

Table 10: **Model characteristics on the internal validation datasets (wake versus REM versus NREM I, II, III):** subject-wise performance metrics (mean \pm SD) are reported using the internal validation data. REM: rapid-eye-movement, NREM: non-rapid-eye-movement, Kappa score: κ .

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

797 8.2. *Cohort-specific performance against polysomnography using SleepNet*

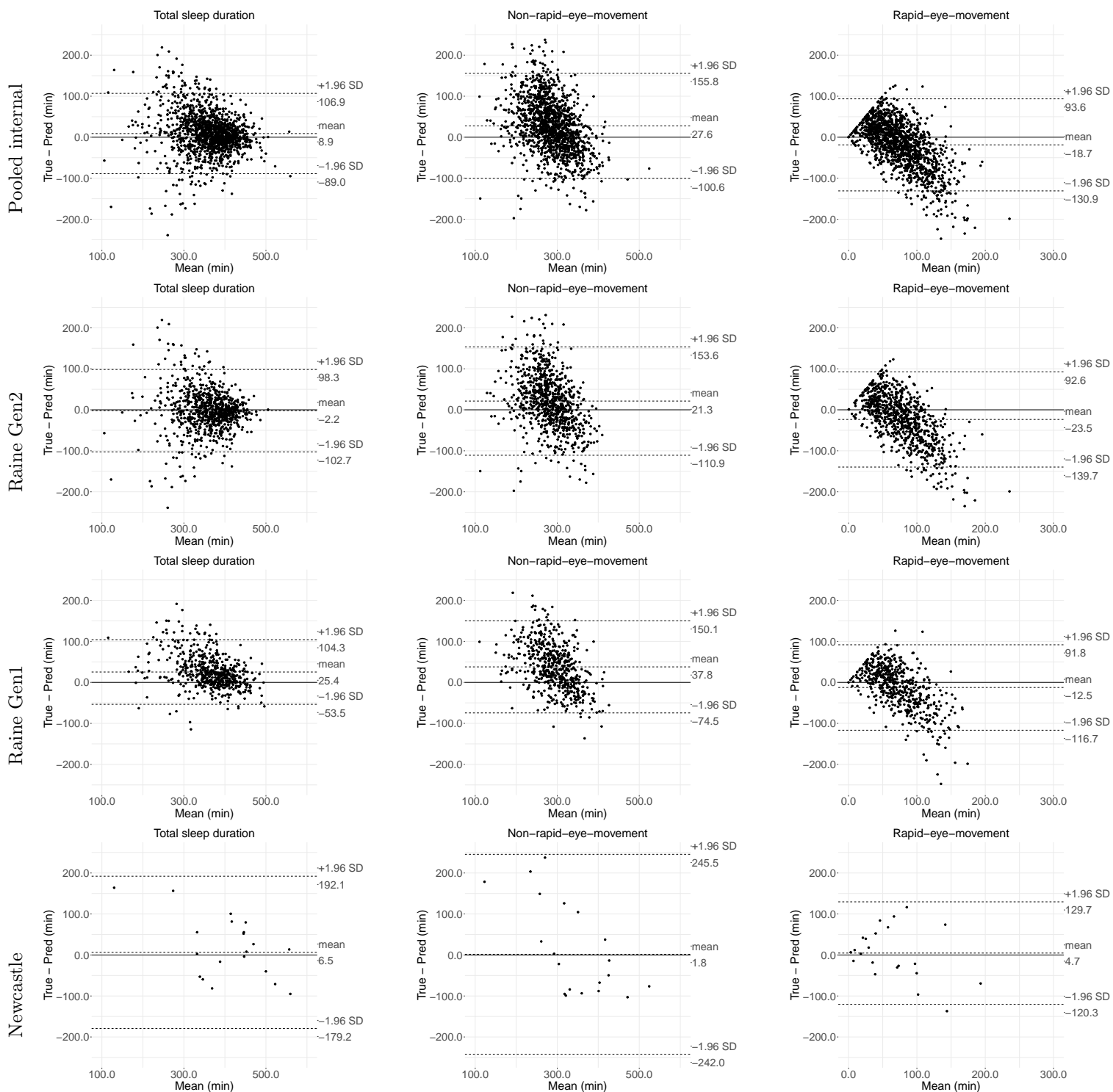


Figure 5: Agreement assessment via Bland-Altman plots for internal validation: total sleep duration (TSD), non-rapid-eye-movement sleep (NREM), and rapid-eye-movement sleep (REM).

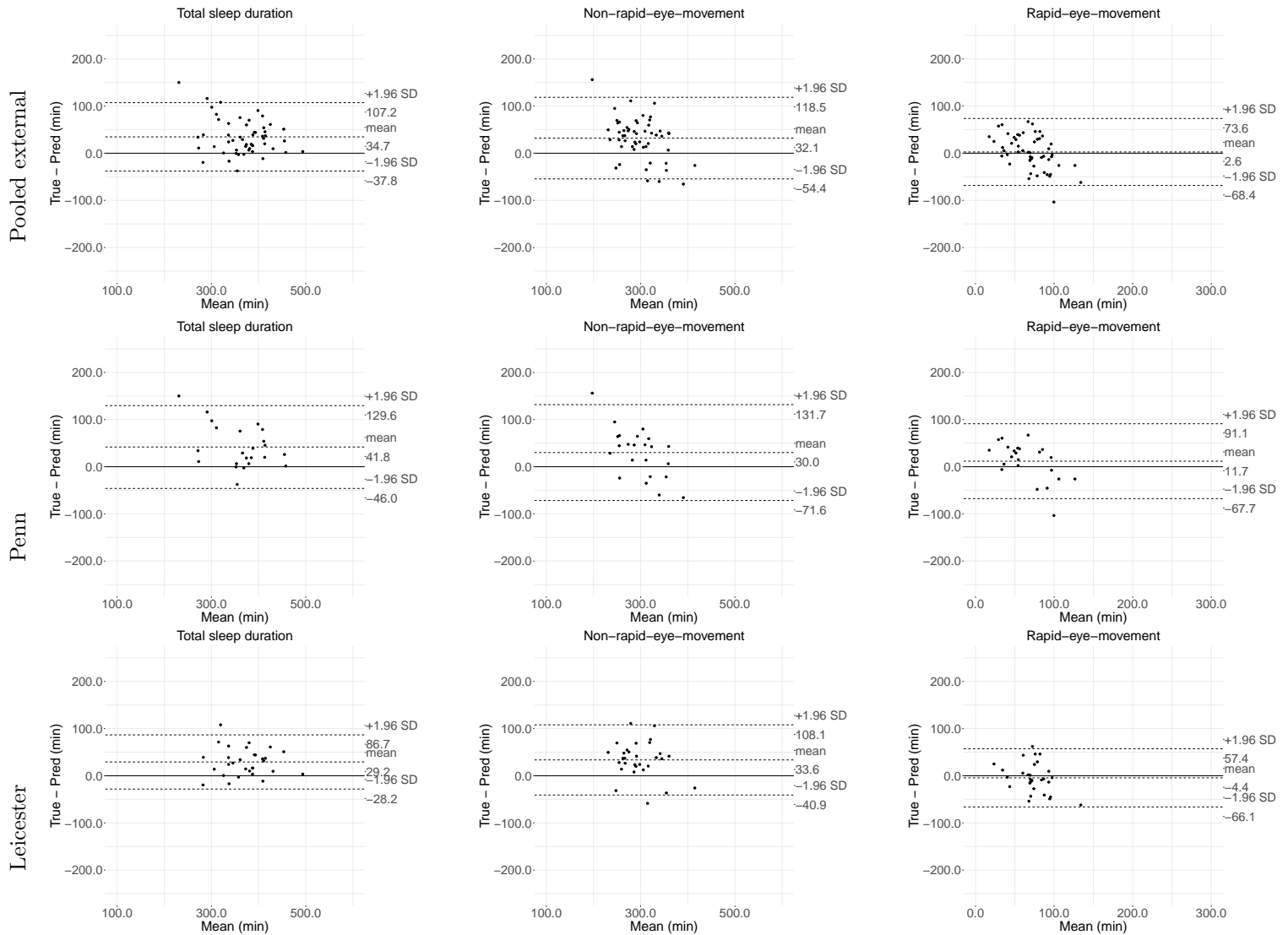


Figure 6: Agreement assessment via Bland-Altman plots for external validation: total sleep duration, wake after sleep onset (WASO), non-rapid-eye-movement sleep (NREM), and rapid-eye-movement sleep (REM).

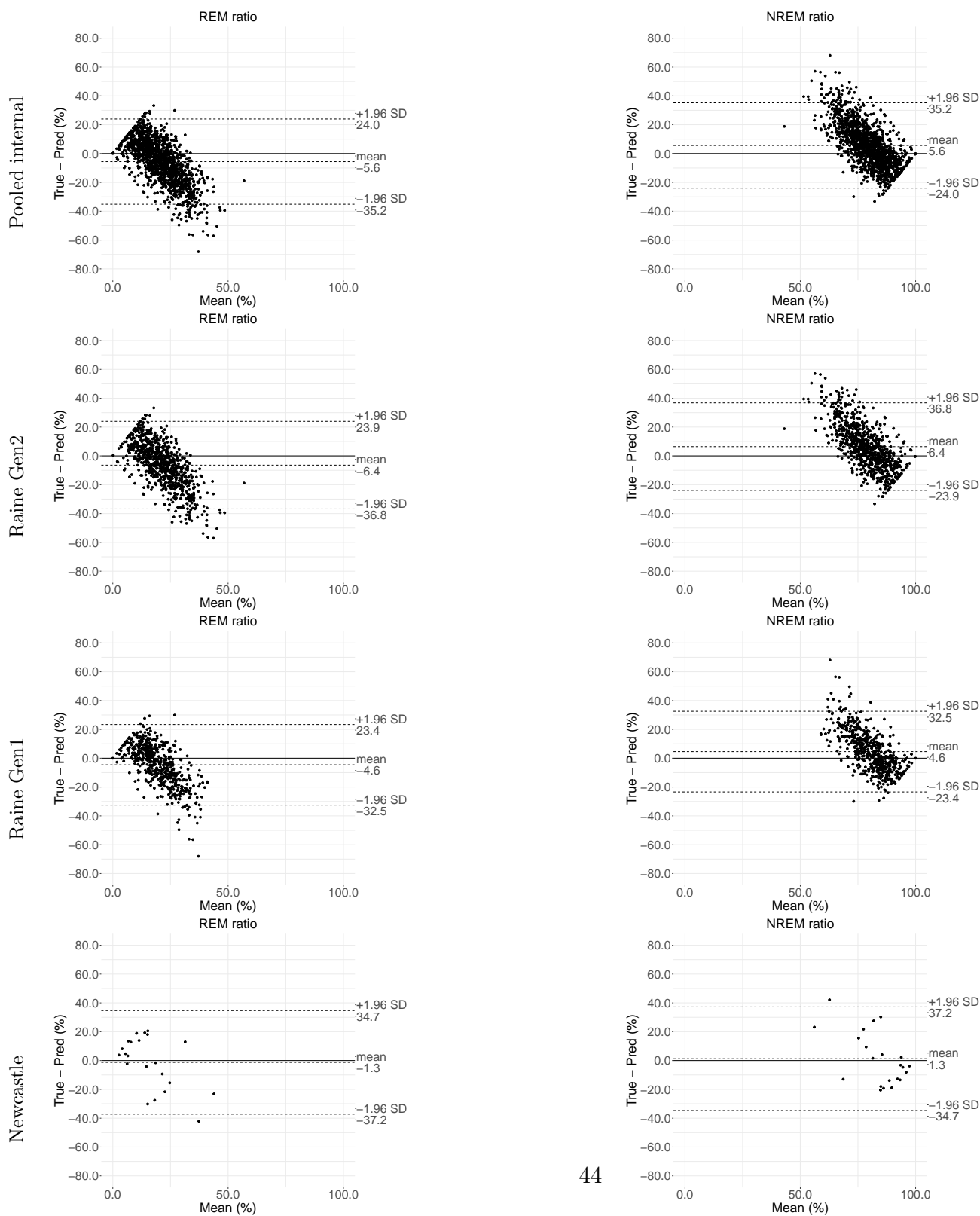


Figure 7: Agreement assessment via Bland-Altman plots for internal validation: non-rapid-eye-movement sleep (NREM) ratio, and rapid-eye-movement sleep (REM) ratio.

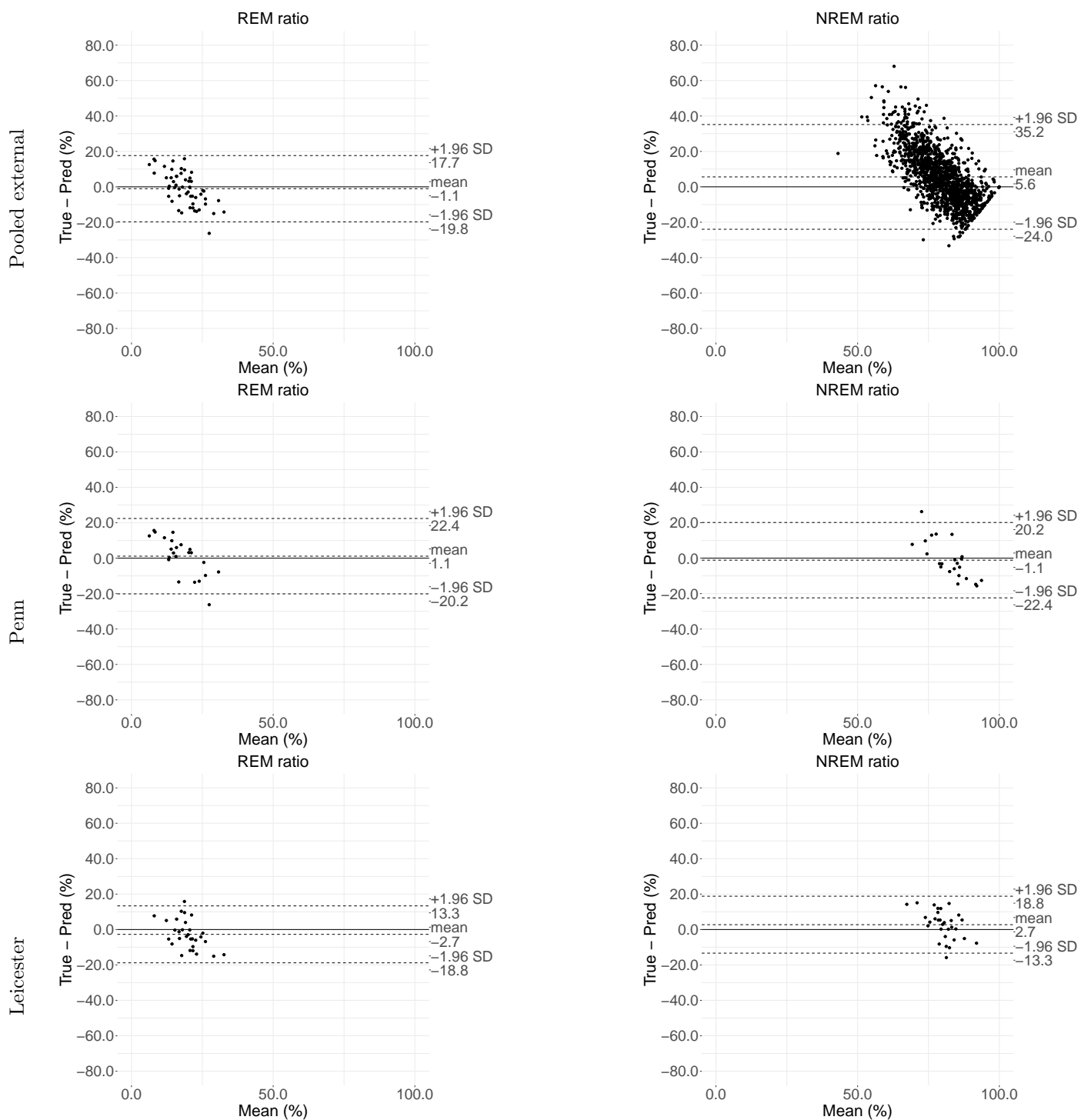


Figure 8: Agreement assessment via Bland-Altman plots for external validation: non-rapid-eye-movement sleep (NREM) ratio, and rapid-eye-movement sleep (REM) ratio.

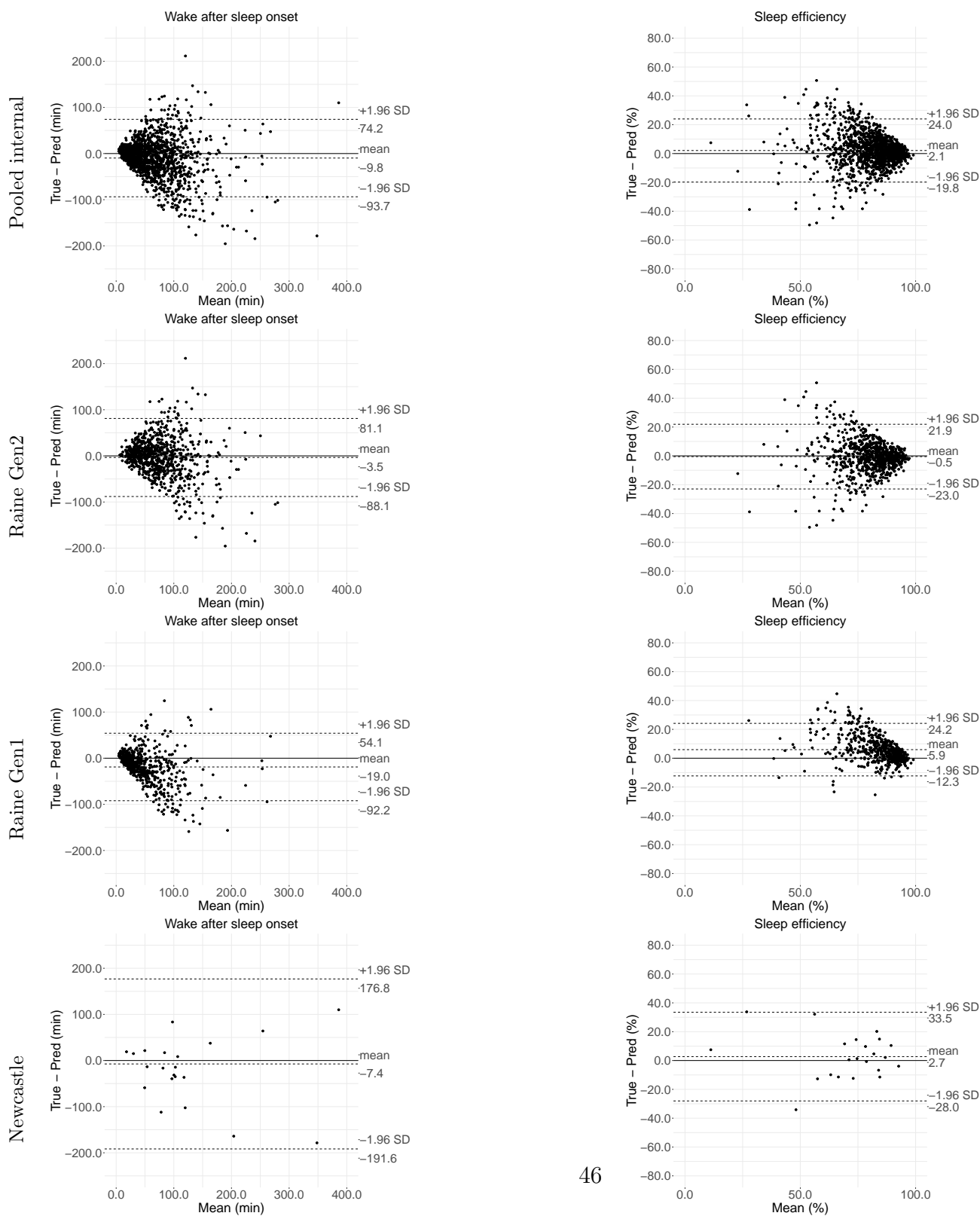


Figure 9: Agreement assessment via Bland-Altman plots for internal validation: wake after sleep onset (WASO), and sleep efficiency (SE).

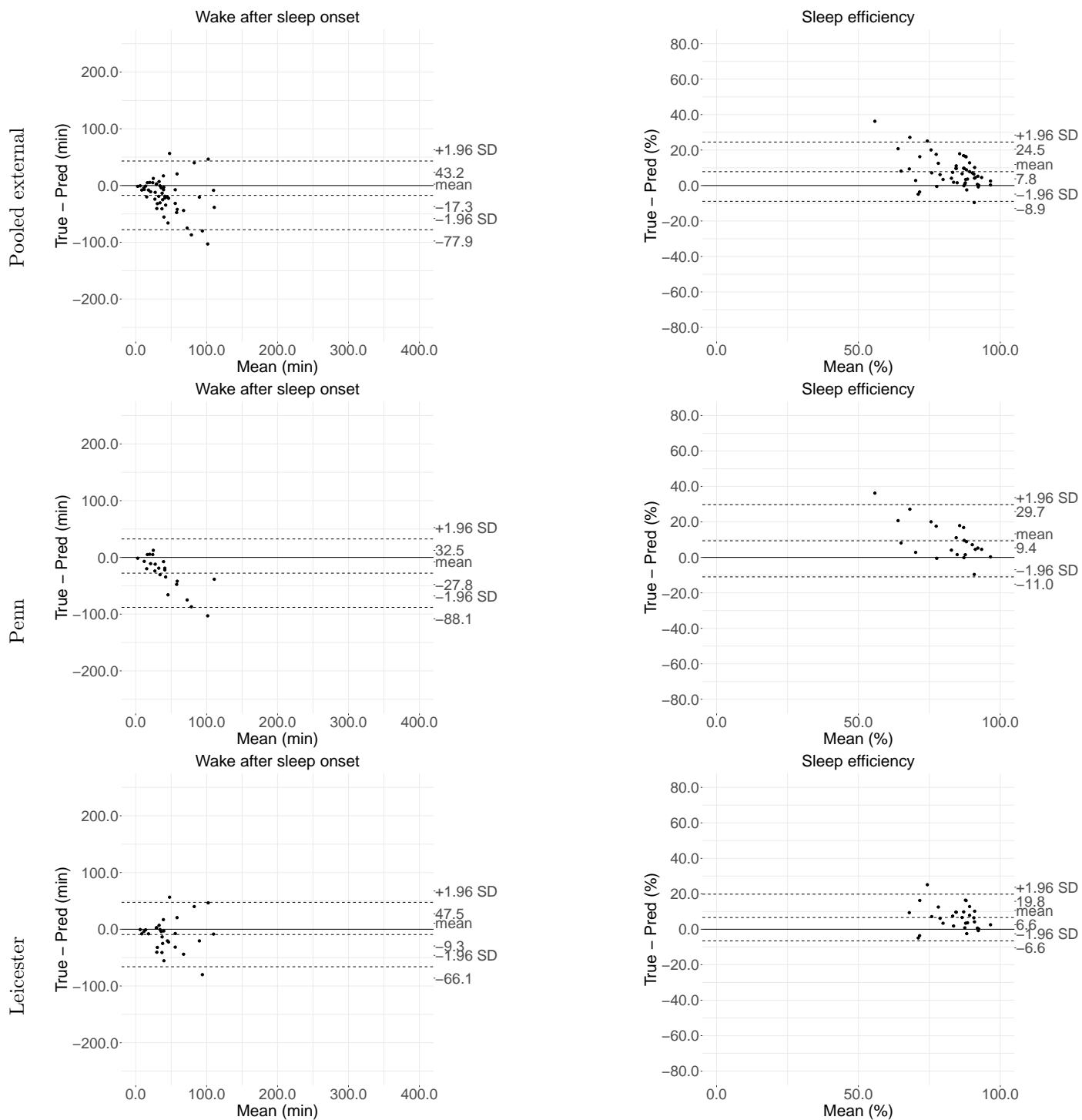


Figure 10: Agreement assessment via Bland-Altman plots for internal validation: wake after sleep onset (WASO), and sleep efficiency (SE).

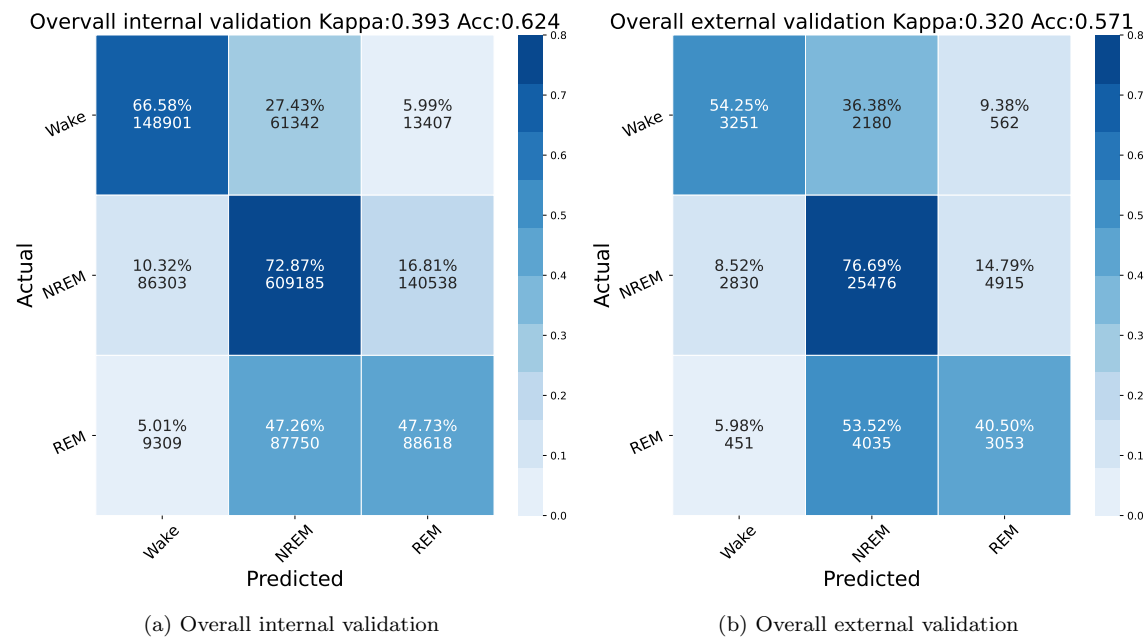


Figure 11: **Three class classification (wake/REM/NREM) confusion matrix:** epoch-to-epoch Kappa and balanced accuracies are shown. The number of predictions and proportion ratios are shown for each pair of ground-truth and prediction class. REM: rapid-eye-movement sleep; NREM: non-rapid-eye-movement sleep.

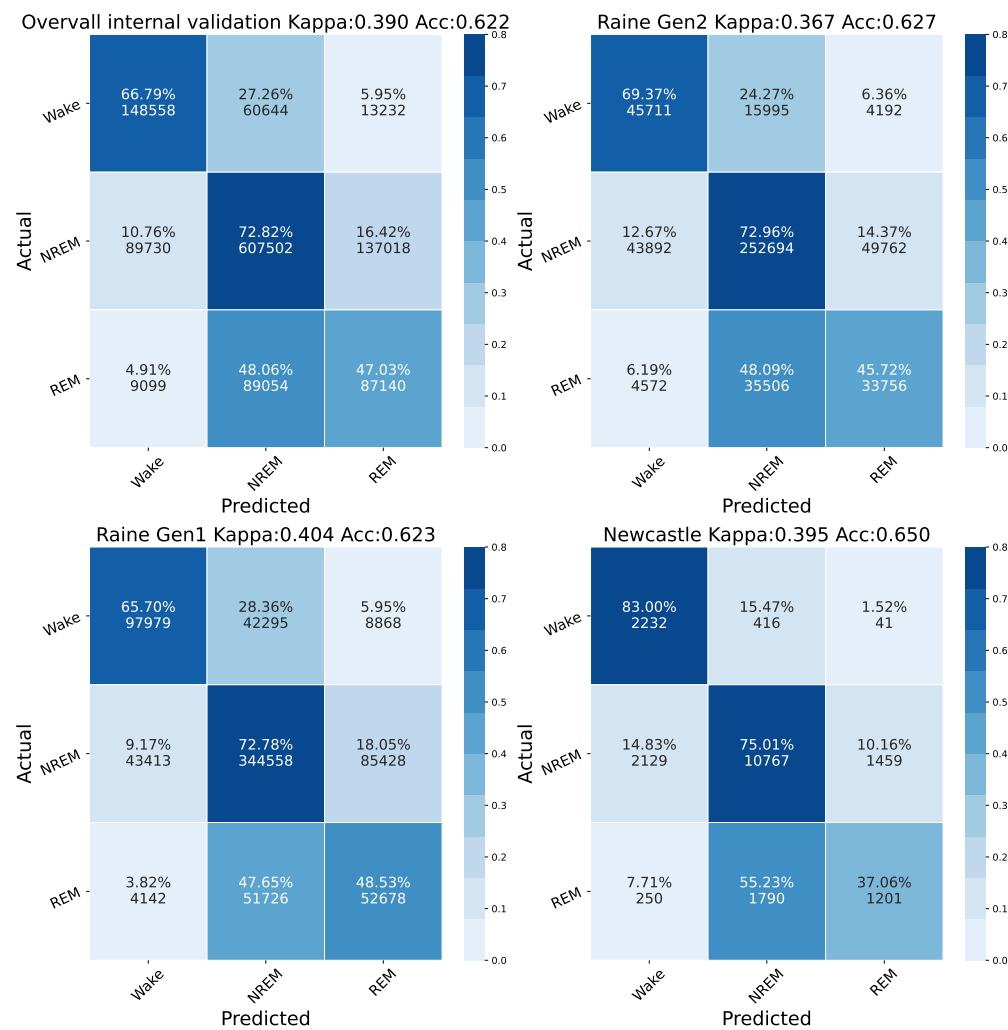


Figure 12: **Three-class sleep staging (wake/REM/NREM) for internal validation: epoch-to-epoch Kappa and balanced accuracies are shown.** The number of predictions and proportion ratios are shown for each pair of ground-truth and prediction class. REM: rapid-eye-movement sleep; NREM: non-rapid-eye-movement sleep.

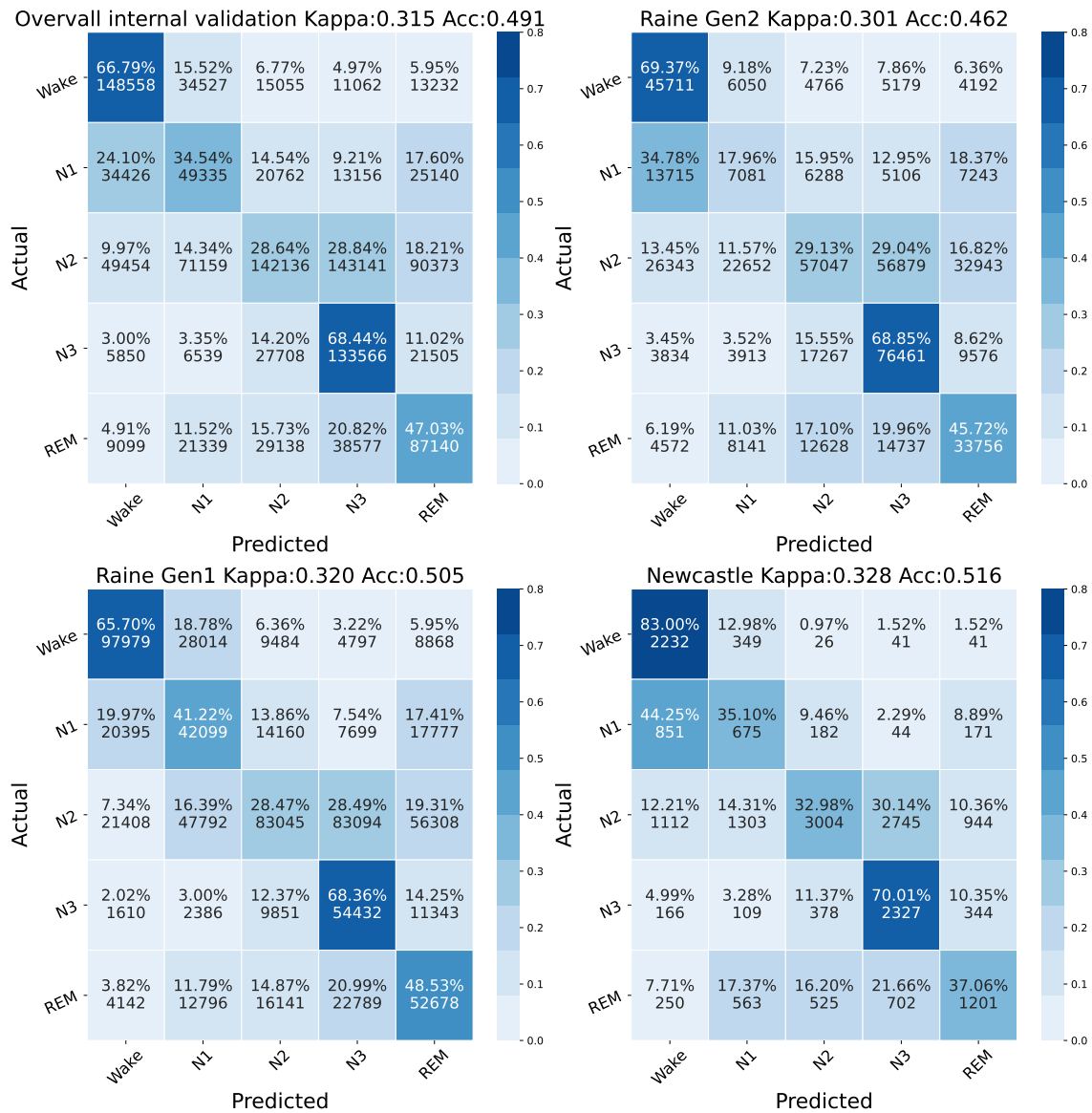


Figure 13: **Five-class sleep staging (wake/REM/N1/N2/N3) for internal validation: epoch-to-epoch kappa and balanced accuracies are shown.** The number of predictions and proportion ratios are shown for each pair of ground-truth and prediction class. REM: rapid-eye-movement sleep, N1, N2, N3: non-rapid-eye-movement sleep 1, 2, 3.

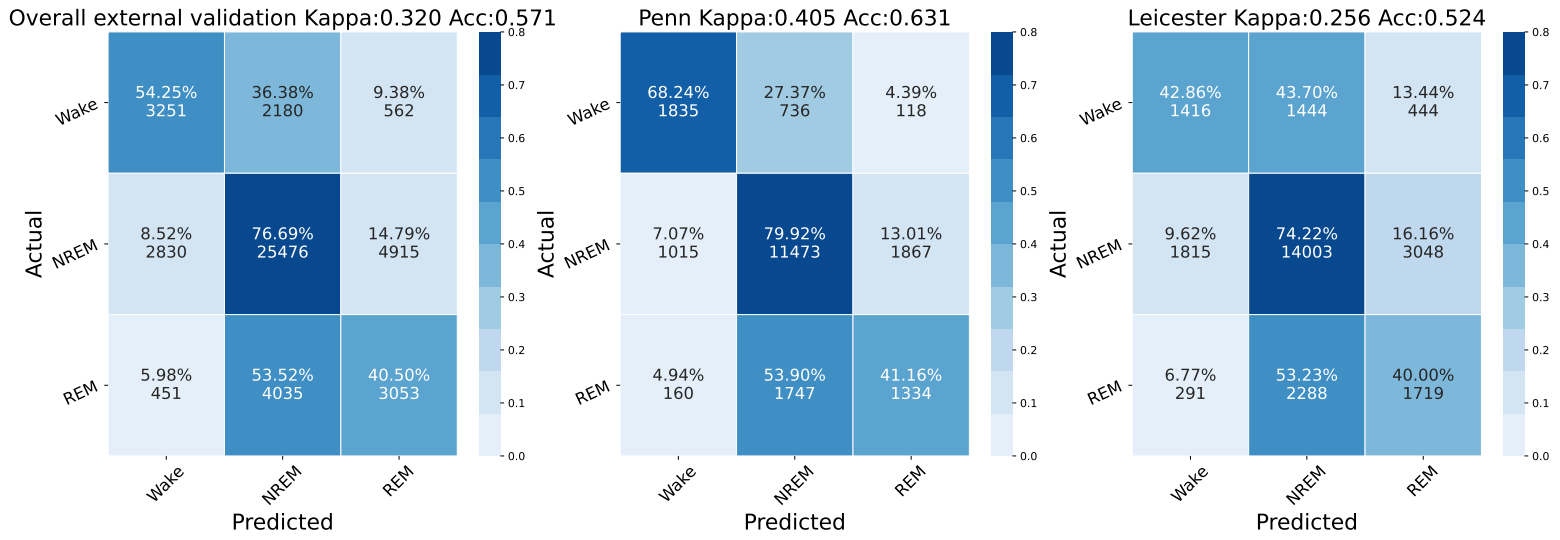


Figure 14: **Three-class sleep staging (wake/REM/NREM) for external validation: epoch-to-epoch kappa and balanced accuracies are shown.** The number of predictions and proportion ratios are shown for each pair of ground-truth and prediction class. REM: rapid-eye-movement sleep; NREM: non-rapid-eye-movement sleep.

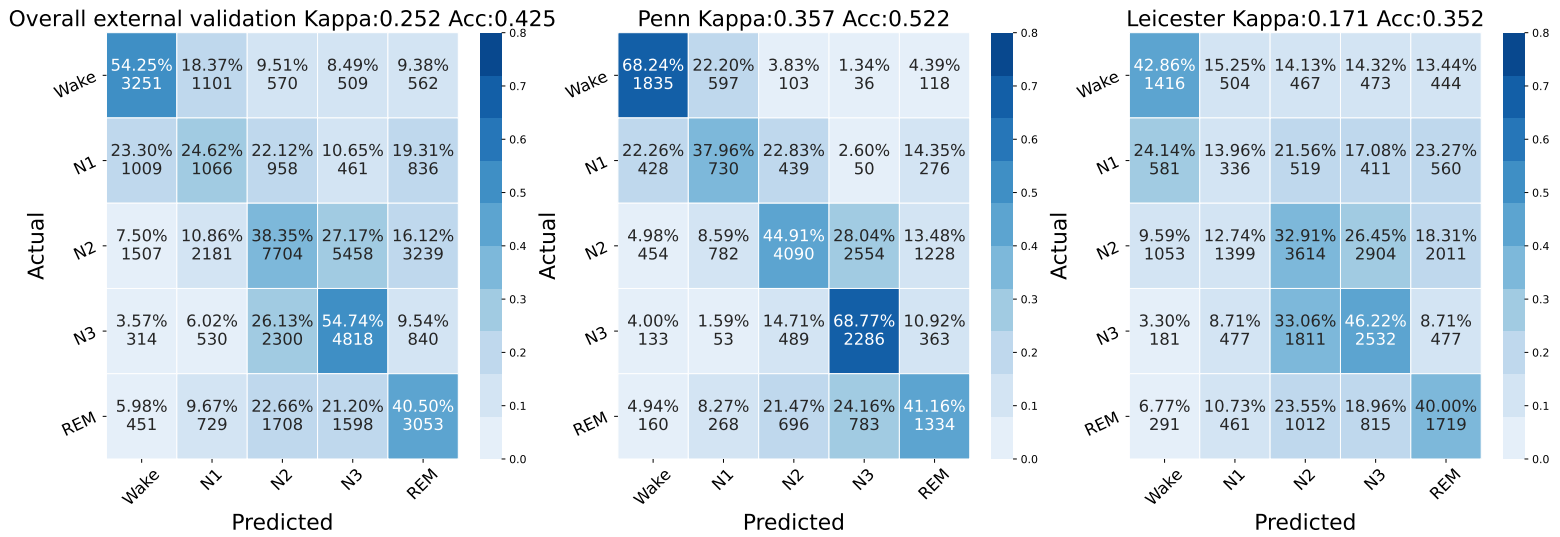


Figure 15: **Five-class sleep staging (wake/REM/N1/N2/N3) for external validation: epoch-to-epoch kappa and balanced accuracies are shown.** The number of predictions and proportion ratios are shown for each pair of ground-truth and prediction class. REM: rapid-eye-movement sleep, N1, N2, N3: non-rapid-eye-movement sleep 1, 2, 3.

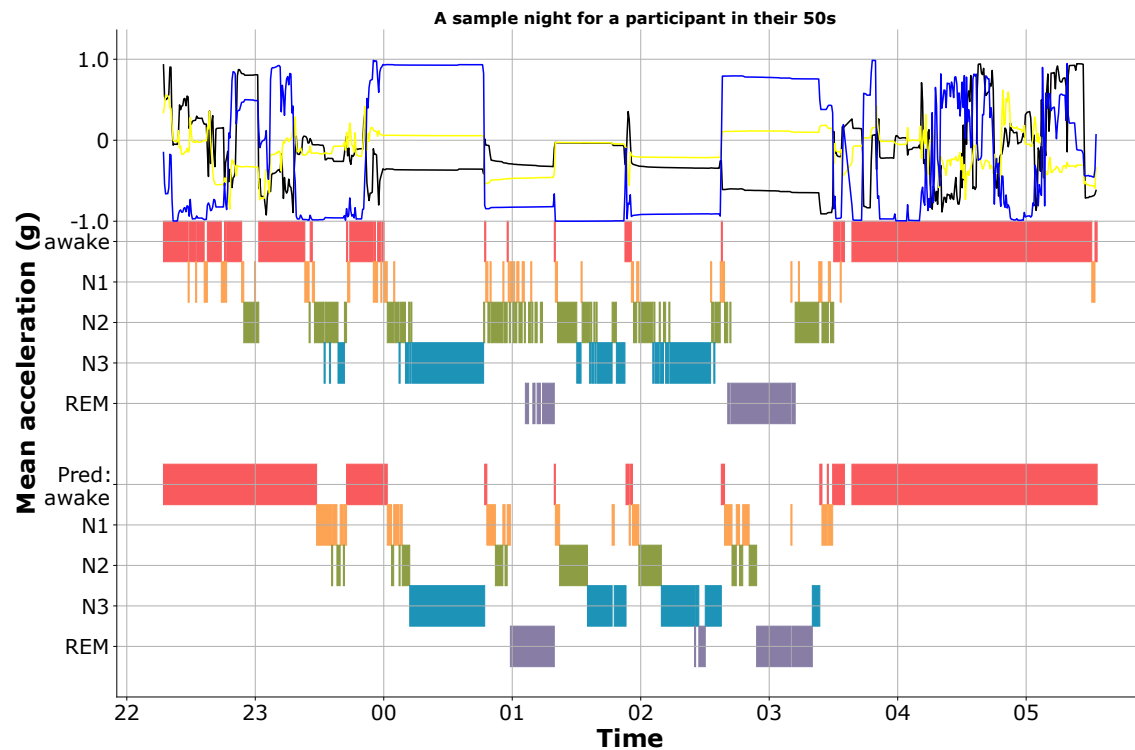
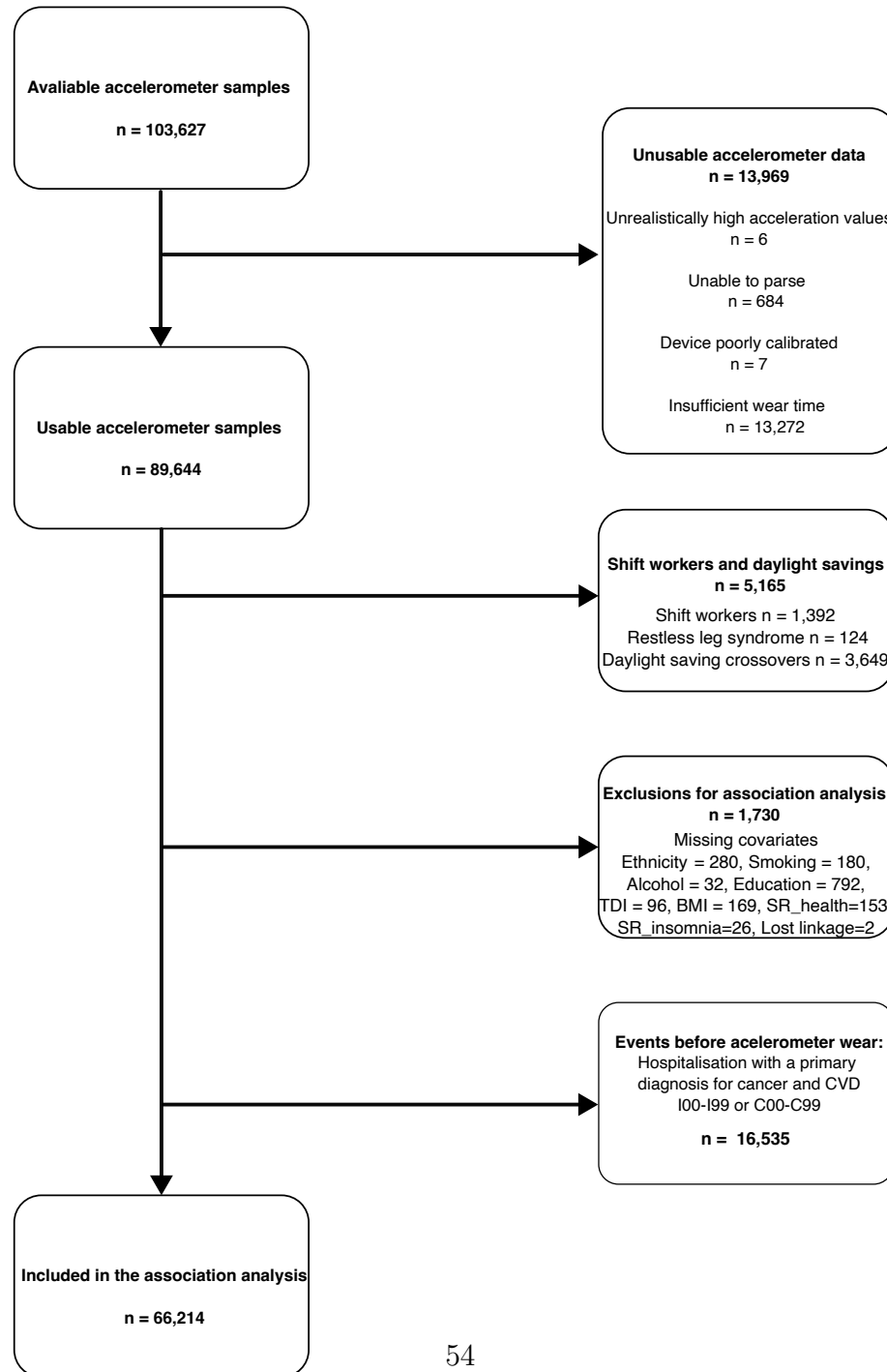


Figure 16: **A sample actigram, hypnogram ground truth and prediction for a participant whose sleep stages are well captured:** the **top** hypnogram is the ground-truth and the **bottom** hypnogram is the prediction generated by SleepNet based on the actigram. REM: rapid-eye-movement sleep, N1, N2, N3: non-rapid-eye-movement sleep 1, 2, 3.

798 *8.3. Additional results on the sleep variations for the UK Biobank participants*

Figure 17: **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.



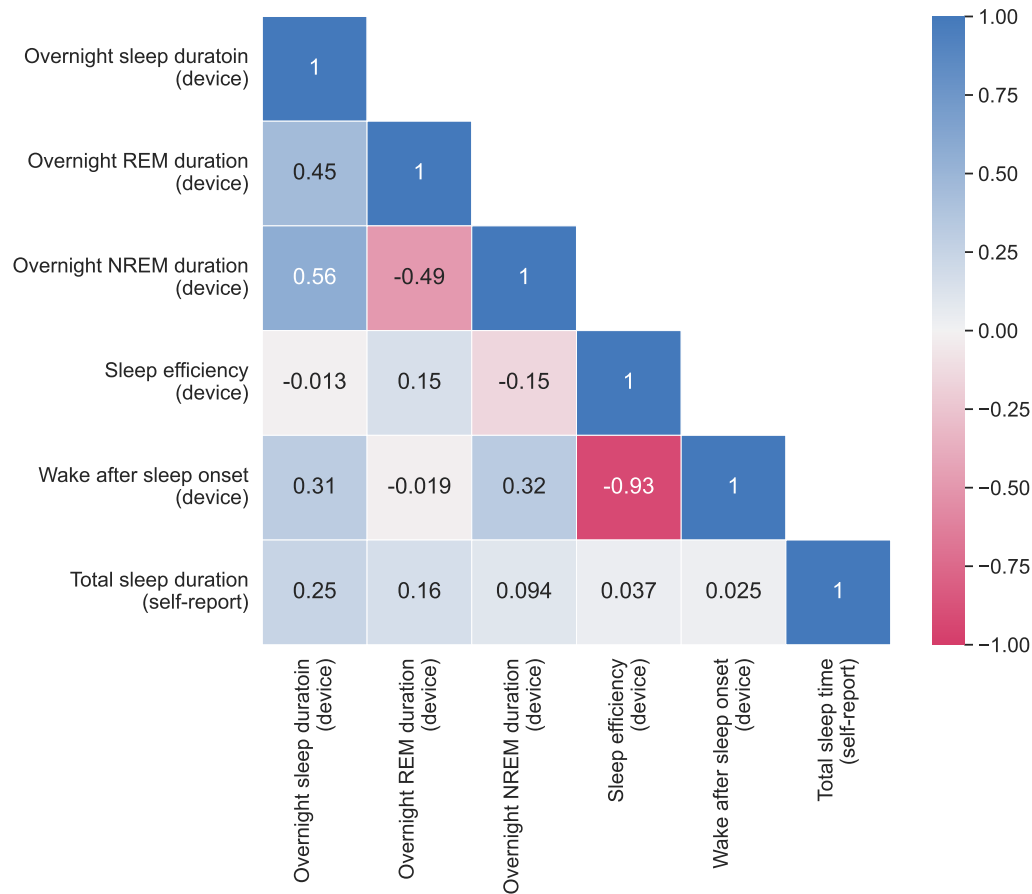


Figure 18: **Correlation matrix for device-measured and self-reported sleep parameters on the UK Biobank.** The self-reported total sleep duration was obtained via questionnaire at baseline assessment in the UK Biobank. REM: rapid-eye-movement sleep, NREM: non-rapid-eye-movement sleep.

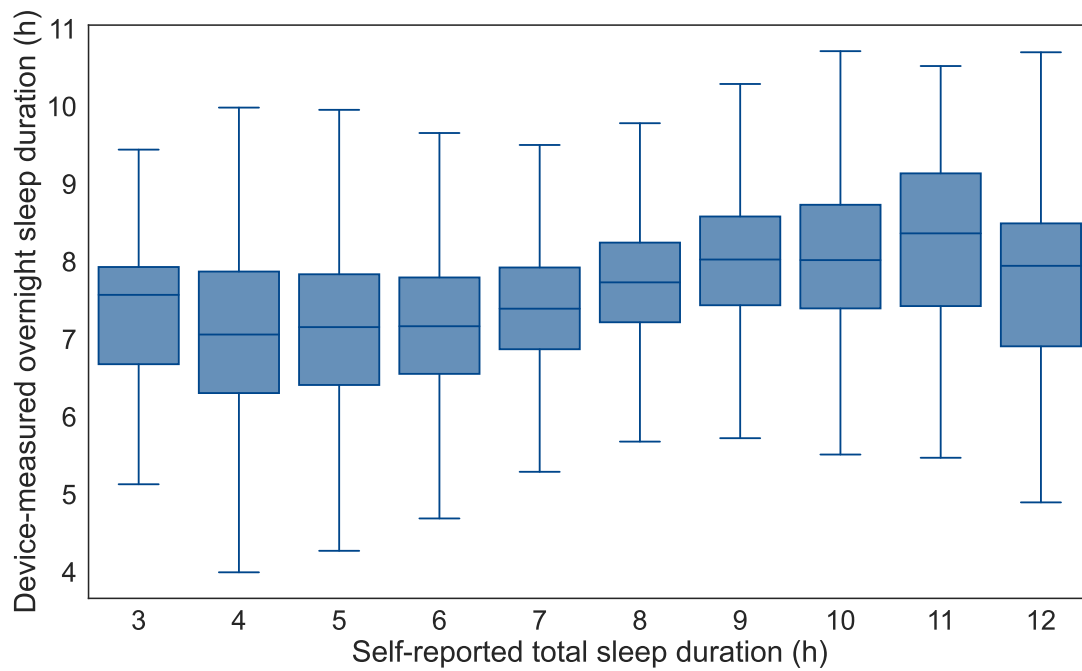


Figure 19: **Box plots showing the distributions of device-measured overnight sleep duration against self-reported total sleep duration.** The box whiskers reflect the lowest and highest data points that are 1.5 times of the inter-quartile-range from the median.

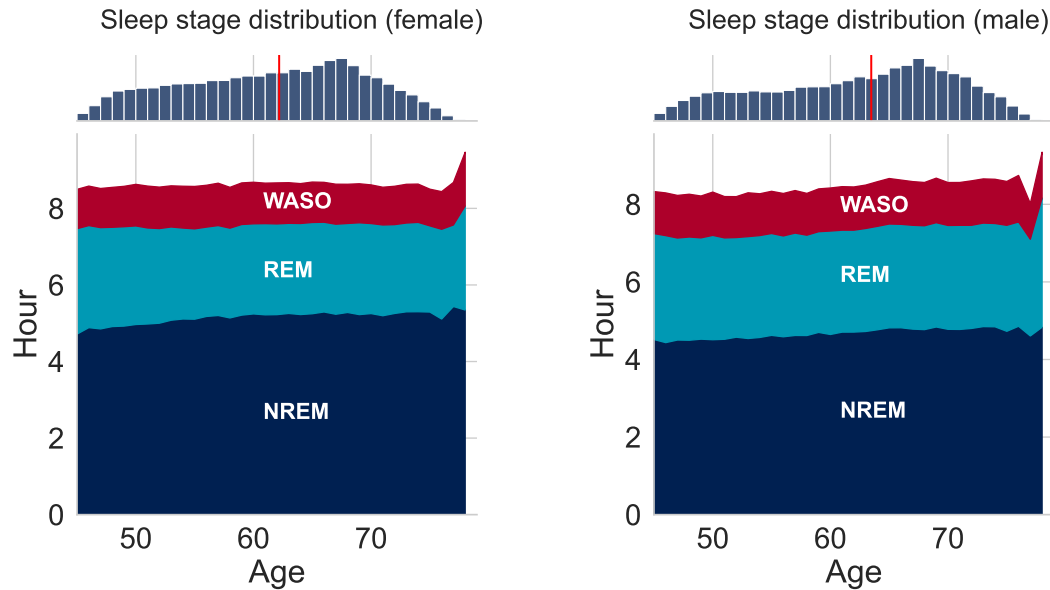


Figure 20: The average device-measured sleep stage distribution with respect to age for both females (left) and males (right) on the UK Biobank. The histograms on the top show the age distribution for the participants. The red vertical line denotes the median age for each sex. WASO: wake after sleep onset; REM: rapid-eye-movement sleep; NREM: non-rapid-eye-movement sleep.

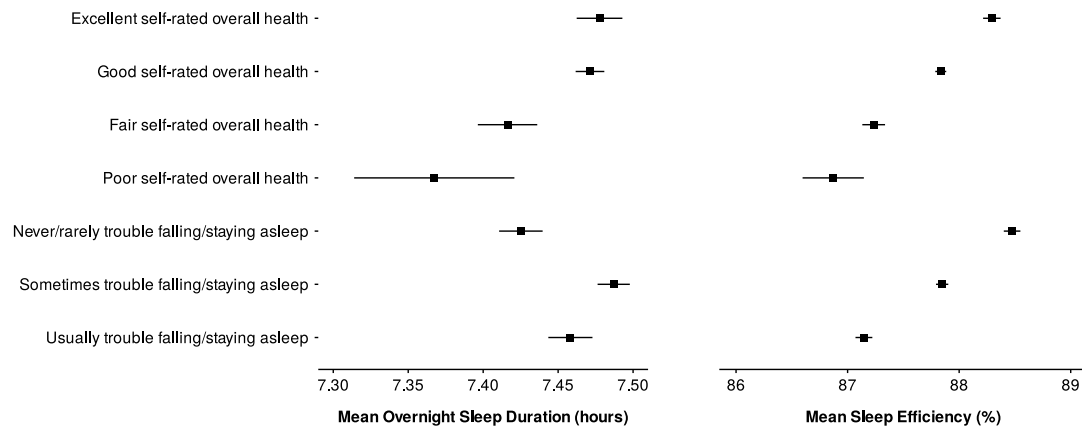


Figure 21: Adjusted marginal mean (95% confidence interval) device-measured mean overnight sleep duration and mean sleep efficiency by self-reported overall health status and insomnia history in the UK Biobank. Mean overnight sleep duration and sleep efficiency were adjusted for age and sex.

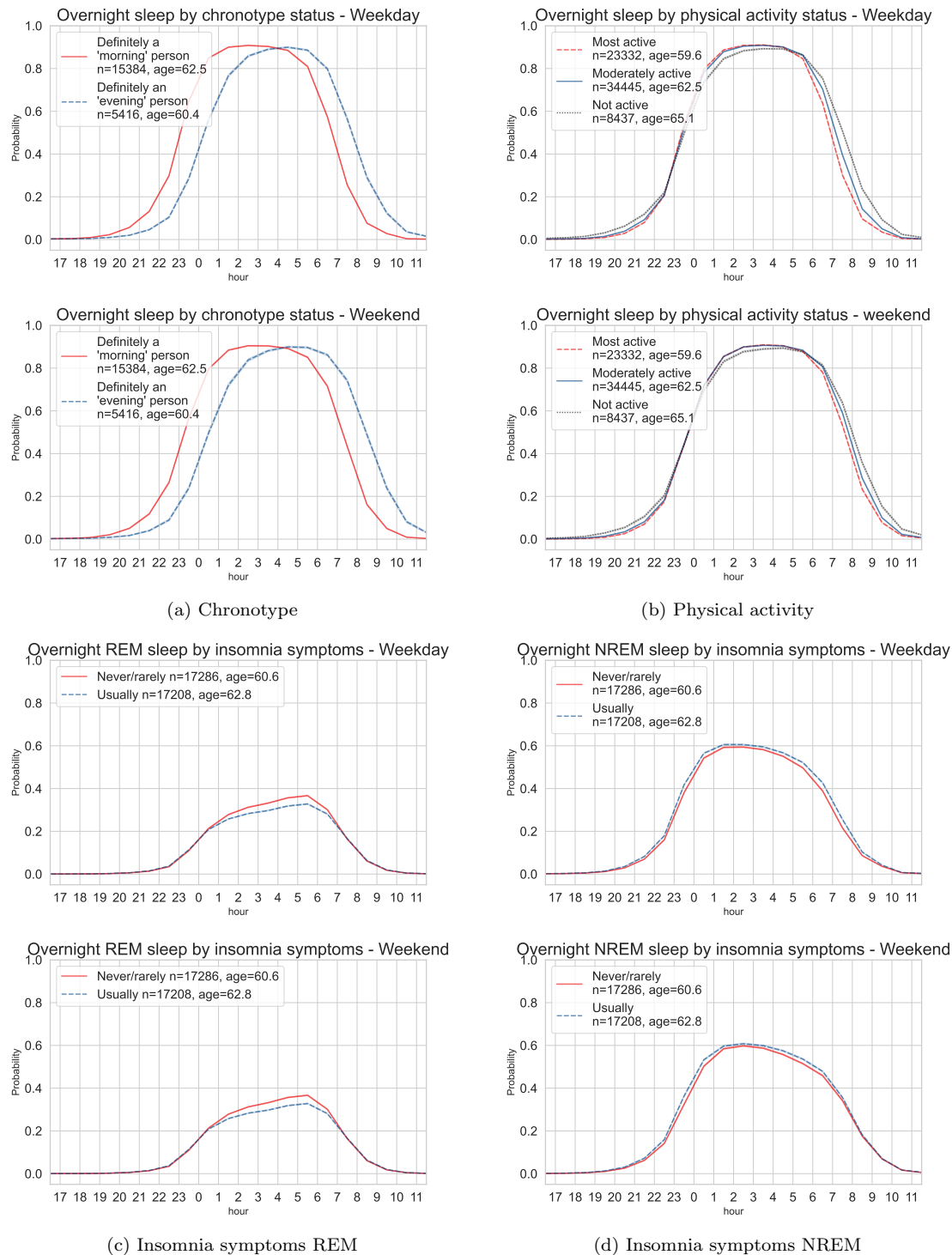


Figure 22: **Device-measured sleep probability trajectories throughout the day for the UK Biobank participants (weekday vs weekend).** Top: variations of the average overnight sleep probability for the participants with self-reported “morning” and “evening” chronotype (a) and the overnight sleep distributions 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). Areas of squares represent the inverse of the variance of the log risk. And the I bars denote the 95% confidence interval for the floated risks.

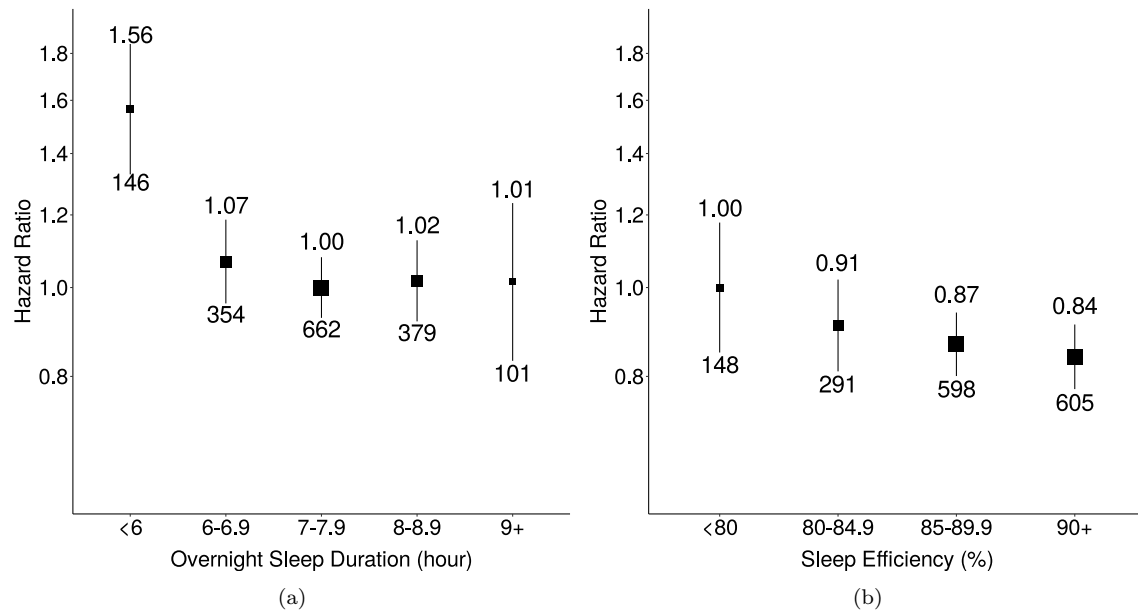


Figure 23: **Associations of overnight sleep duration (a) and sleep efficiency (b) with all-cause mortality.** The model used 1,642 events among 62,214 participants. We used age as the timescale and adjusted for sex, ethnicity, Townsend Deprivation Index of baseline address (split by quarter in the study 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.

799 8.3.1. Models additionally adjusted for body mass index

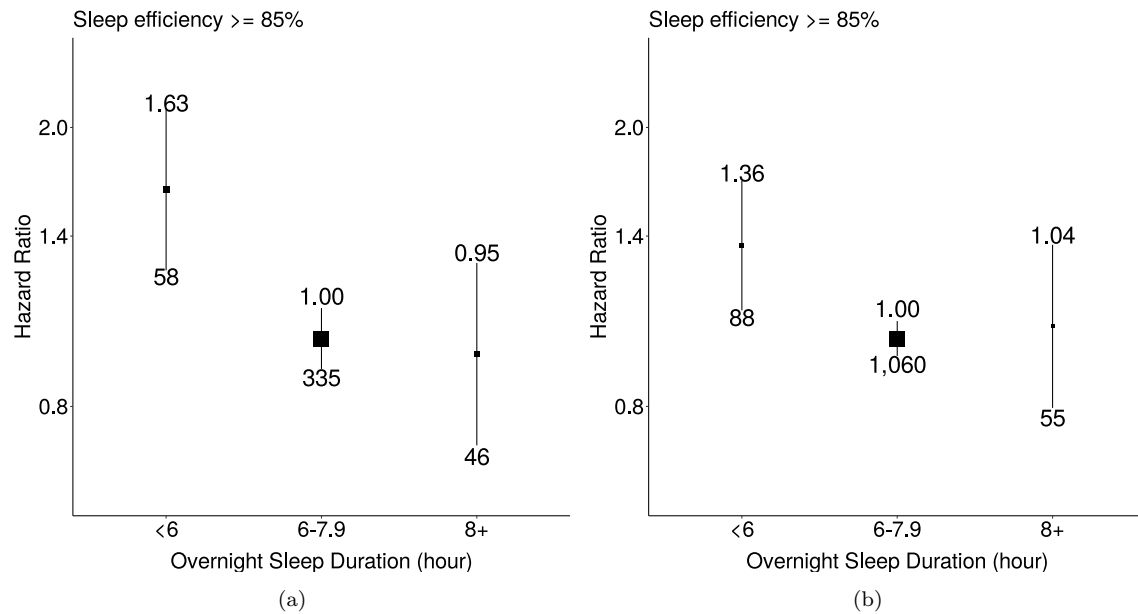


Figure 24: **Associations of overnight sleep duration with all-cause mortality for groups with low and high sleep efficiency additionally adjusted for body mass index.** The model used 1,642 events among 62,214 participants. We used age as the timescale and adjusted for sex, ethnicity, Townsend Deprivation 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.

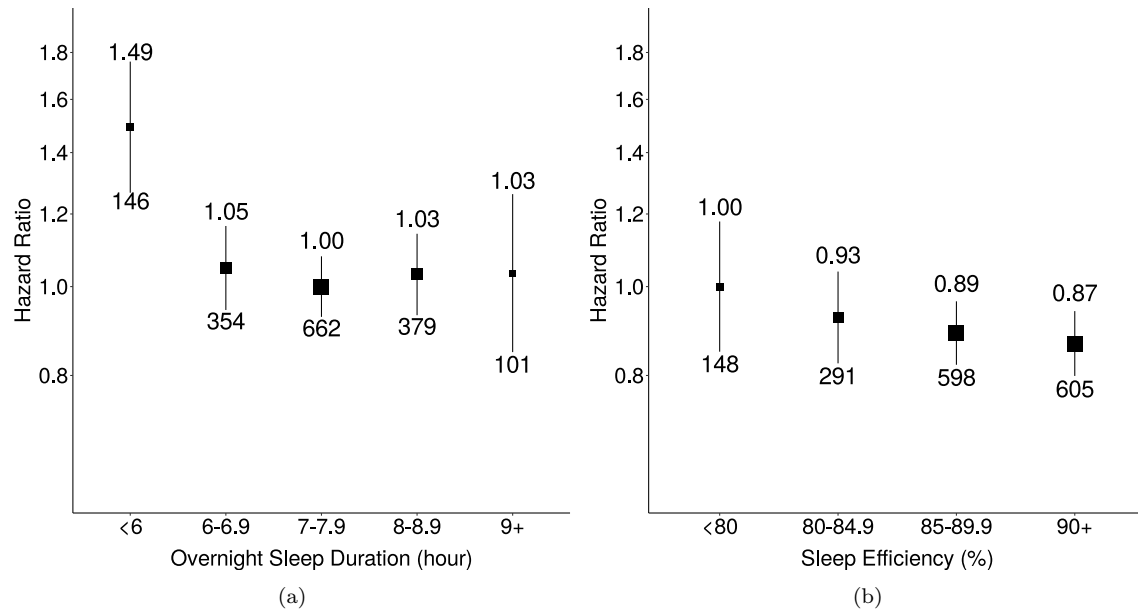


Figure 25: **Associations of overnight sleep duration (a) and sleep efficiency (b) with all-cause mortality additionally adjusted for body mass index.** The model used 1,642 events among 62,214 participants. We used age as the timescale and adjusted for sex, ethnicity, Townsend Deprivation Index of baseline 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), 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.

800 8.3.2. Sensitivity analysis for overnight sleep duration

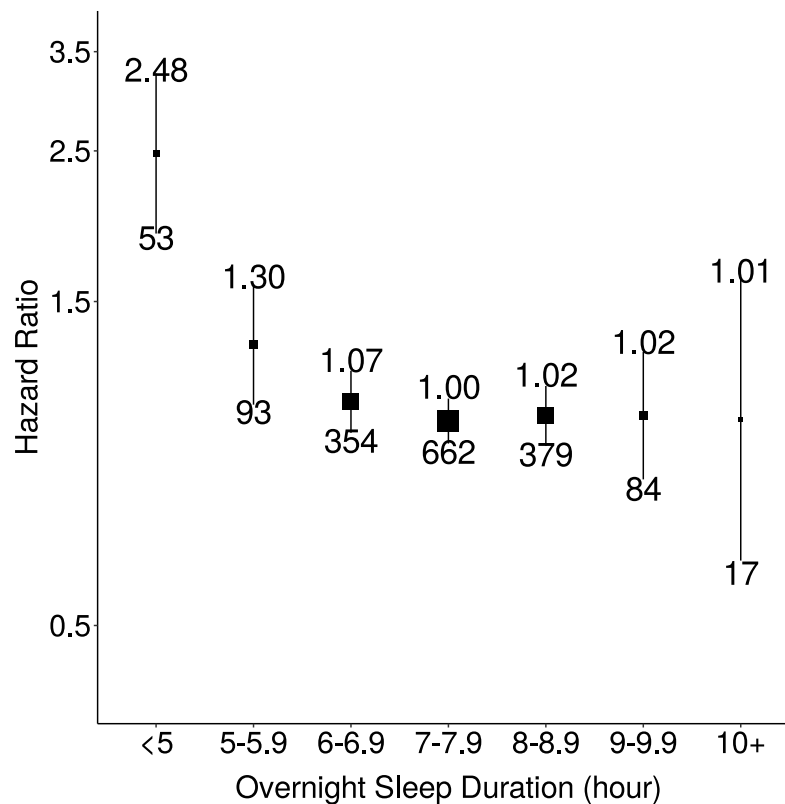


Figure 26: **Associations of device-measured overnight sleep duration and all-cause mortality with greater granularity.** The model used 1,642 events among 62,214 participants. We used age as the timescale and adjusted for sex, ethnicity, Townsend Deprivation Index of baseline address (split by quarter in the study population), educational qualifications, smoking status, alcohol consumption (Never, <3 times/week, 3+ times/week), and 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.

801 **References**

- 802 [1] Aiden Doherty et al. “Large scale population assessment of physical activity
803 using wrist worn accelerometers: the UK biobank study”. In: *PloS One* 12.2
804 (2017), e0169649.
- 805 [2] Leon Straker et al. “Cohort profile: the Western Australian pregnancy cohort
806 (Raine) study—Generation 2”. In: *International Journal of Epidemiology* 46.5
807 (2017), 1384–1385j.
- 808 [3] Vincent van Hees, Sarah Charman, and Kirstie Anderson. *Newcastle polysomnog-*
809 *raphy and accelerometer data*. Version 1.0. Zenodo, Jan. 2018. DOI: 10.5281/
810 [zenodo.1160410](https://doi.org/10.5281/zenodo.1160410). URL: <https://doi.org/10.5281/zenodo.1160410>.
- 811 [4] Tatiana Plekhanova et al. “Validation of an automated sleep detection algo-
812 rithm using data from multiple accelerometer brands”. In: *Journal of Sleep*
813 *Research* (2022).
- 814 [5] Enda M Byrne et al. “Genetic correlation analysis suggests association between
815 increased self-reported sleep duration in adults and schizophrenia and type 2
816 diabetes”. In: *Sleep* 39.10 (2016), pp. 1853–1857.
- 817 [6] Manon L Dontje, Peter Eastwood, and Leon Straker. “Western Australian preg-
818 nancy cohort (Raine) study: generation 1”. In: *BMJ open* 9.5 (2019), e026276.
- 819 [7] Cathie Sudlow et al. “UK biobank: an open access resource for identifying the
820 causes of a wide range of complex diseases of middle and old age”. In: *PLoS*
821 *Medicine* 12.3 (2015), e1001779.
- 822 [8] Hang Yuan et al. “Self-supervised Learning for Human Activity Recognition
823 Using 700,000 Person-days of Wearable Data”. In: *arXiv preprint arXiv:2206.02909*
824 (2022).
- 825 [9] Kaiming He et al. “Identity mappings in deep residual networks”. In: *European*
826 *Conference on Computer Vision*. Springer. 2016, pp. 630–645.
- 827 [10] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimiza-
828 tion”. In: *arXiv preprint arXiv:1412.6980* (2014).

- 829 [11] Zhiheng Huang, Wei Xu, and Kai Yu. “Bidirectional LSTM-CRF models for
830 sequence tagging”. In: *arXiv preprint arXiv:1508.01991* (2015).
- 831 [12] Kalaivani Sundararajan et al. “Sleep classification from wrist-worn accelerom-
832 eter data using random forests”. In: *Scientific Reports* 11.1 (2021), pp. 1–10.
- 833 [13] Rosemary Walmsley et al. “Reallocation of time between device-measured
834 movement behaviours and risk of incident cardiovascular disease”. In: *British
835 Journal of Sports Medicine* 56.18 (2022), pp. 1008–1017.
- 836 [14] Max Hirshkowitz et al. “National Sleep Foundation’s updated sleep duration
837 recommendations”. In: *Sleep health* 1.4 (2015), pp. 233–243.
- 838 [15] Bin Yan et al. “Objective sleep efficiency predicts cardiovascular disease in
839 a community population: the sleep heart health study”. In: *Journal of the
840 American Heart Association* 10.7 (2021), e016201.
- 841 [16] Douglas F Easton, Julian Peto, and Abdel GAG Babiker. “Floating absolute
842 risk: an alternative to relative risk in survival and case-control analysis avoiding
843 an arbitrary reference group”. In: *Statistics in Medicine* 10.7 (1991), pp. 1025–
844 1035.
- 845 [17] Martyn Plummer and Bendix Carstensen. “Lexis: An R Class for Epidemio-
846 logical Studies with Long-Term Follow-Up”. In: *Journal of Statistical Software*
847 38.5 (2011), pp. 1–12. URL: <https://www.jstatsoft.org/v38/i05/>.
- 848 [18] Martyn Plummer. “Improved estimates of floating absolute risk”. In: *Statistics
849 in Medicine* 23.1 (2004), pp. 93–104.
- 850 [19] Terry K Koo and Mae Y Li. “A guideline of selecting and reporting intra-
851 class correlation coefficients for reliability research”. In: *Journal of Chiropractic
852 Medicine* 15.2 (2016), pp. 155–163.