

1 **Development and Validation of a Machine Learning Wrist-worn Step Detection Algorithm with**

2 **Deployment in the UK Biobank**

3 Scott R. Small^{1,2,3}, Shing Chan^{1,2}, Rosemary Walmsley^{1,2}, Lennart von Fritsch³, Aidan Acquah^{1,2,4},

4 Gert Mertes^{1,2,4}, Benjamin G. Feakins^{1,2}, Andrew Creagh^{1,4}, Adam Strange⁵, Charles E. Matthews⁶,

5 David A. Clifton⁴, Andrew J. Price³, Sara Khalid³, Derrick Bennett¹, Aiden Doherty^{1,2}

6

7 **Affiliations:**

8 ¹Nuffield Department of Population Health, University of Oxford, UK

9 ²Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford,
10 UK

11 ³Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University
12 of Oxford, UK

13 ⁴Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford

14 ⁵SwissRe Institute, UK

15 ⁶ Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland,
16 USA

17

18 **Correspondence to:**

19 Aiden Doherty, PhD

20 Professor of Biomedical Informatics

21 University of Oxford

22 Nuffield Department of Population Health

23 Big Data Institute, Old Road Campus

24 Oxford, UK OX3 7LF

25 Email: aiden.doherty@ndph.ox.ac.uk

26 **Abstract**

27 **Background:** Step count is an intuitive measure of physical activity frequently quantified in a
28 range of health-related studies; however, accurate quantification of step count can be difficult in
29 the free-living environment, with step counting error routinely above 20% in both consumer and
30 research-grade wrist-worn devices. This study aims to describe the development and validation
31 of step count derived from a wrist-worn accelerometer and to assess its association with
32 cardiovascular and all-cause mortality in a large prospective cohort study.

33 **Methods:** We developed and externally validated a hybrid step detection model that involves
34 self-supervised machine learning, trained on a new ground truth annotated, free-living step
35 count dataset (OxWalk, n=39, aged 19-81) and tested against other open-source step counting
36 algorithms. This model was applied to ascertain daily step counts from raw wrist-worn
37 accelerometer data of 75,493 UK Biobank participants without a prior history of cardiovascular
38 disease (CVD) or cancer. Cox regression was used to obtain hazard ratios and 95% confidence
39 intervals for the association of daily step count with fatal CVD and all-cause mortality after
40 adjustment for potential confounders.

41 **Findings:** The novel step algorithm demonstrated a mean absolute percent error of 12.5% in
42 free-living validation, detecting 98.7% of true steps and substantially outperforming other recent
43 wrist-worn, open-source algorithms. Our data are indicative of an inverse dose-response
44 association, where, for example, taking 6,596 to 8,474 steps per day was associated with a 39%
45 [24-52%] and 27% [16-36%] lower risk of fatal CVD and all-cause mortality, respectively,
46 compared to those taking fewer steps each day.

47 **Interpretation:** An accurate measure of step count was ascertained using a machine learning
48 pipeline that demonstrates state-of-the-art accuracy in internal and external validation. The
49 expected associations with CVD and all-cause mortality indicate excellent face validity. This
50 algorithm can be used widely for other studies that have utilised wrist-worn accelerometers and
51 an open-source pipeline is provided to facilitate implementation.

52

53

54

55

56

57

58

59

60

61

62

63

64

65 **Funding Acknowledgements:** This research has been conducted using the UK Biobank Resource
66 under Application Number 59070. This research was funded in whole or in part by the
67 Wellcome Trust [223100/Z/21/Z]. For the purpose of open access, the author has applied a CC-
68 BY public copyright licence to any author accepted manuscript version arising from this
69 submission. AD and SS are supported by the Wellcome Trust. AD and DM are supported by
70 Swiss Re, while AS is an employee of Swiss Re. AD, SC, RW, SS, and SK are supported by HDR UK,
71 an initiative funded by UK Research and Innovation, Department of Health and Social Care
72 (England) and the devolved administrations. AD, DB, GM, and SC are supported by
73 NovoNordisk. AD is supported by the BHF Centre of Research Excellence (grant number
74 RE/18/3/34214). SS is supported by the University of Oxford Clarendon Fund. DB is further
75 supported by the Medical Research Council (MRC) Population Health Research Unit. DC holds a
76 personal academic fellowship from EPSRC. AA, AC and DC are supported by GlaxoSmithKline. SK
77 is supported by Amgen and UCB BioPharma outside of the scope of this work. Computational
78 aspects of this research were funded from the National Institute for Health Research (NIHR)
79 Oxford Biomedical Research Centre (BRC) with additional support from Health Data Research
80 (HDR) UK and the Wellcome Trust Core Award [grant number 203141/Z/16/Z]. The views
81 expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the
82 Department of Health.

83

84

85 Introduction

86 Physical activity has been associated with lower risk of a wide range of non-communicable
87 diseases and is a key feature of public health guidelines for cardiovascular health¹⁻³. While
88 researchers most commonly report device-measured activity in terms of overall acceleration or
89 time-use behaviours derived from intensity thresholds⁴, the reporting of steps is a more intuitive
90 measure of physical activity intrinsically linked to the key biomechanical feature of human gait⁵.
91 However, current methods to measure steps from wrist-worn monitors during free-living activity
92 are inaccurate⁶.

93 Most activity tracking devices with embedded step counting rely on proprietary step counting
94 methods without transparent evaluation⁷, and many popular open-source step counting
95 algorithms were not developed in accordance with, or lack validation against, direct observation
96 ground truth step counts in a free-living environment⁸⁻¹⁰. Current standards require commercial
97 activity trackers to estimate step counts with an error of less than 10% in laboratory-controlled
98 treadmill testing¹¹. Subsequently, many devices and algorithms perform well during scripted,
99 moderately paced walking in controlled conditions^{12,13}. However, step counting performance
100 substantially deteriorates in the real-world environment, wherein mean absolute percent error
101 (MAPE) is regularly well above 20% in both commercial and research-grade activity monitors
102 during free living⁶. As a consequence, uncertainty exists around the strength and shape of the
103 association of daily step count with all-cause mortality and cardiovascular mortality^{14,15}, where
104 recent studies have not used transparent or robustly validated free-living step counting
105 algorithms.

106 In response, we set out to develop and validate a method to accurately measure steps in free-
107 living environments. The purpose of this study was threefold: 1) to develop a novel self-
108 supervised learning step detection algorithm trained with free-living stepping data, 2) to
109 externally validate the algorithm alongside other open-source algorithms, and 3) to evaluate the
110 face validity of this method in a large scale prospective cohort study by associating step counts
111 with fatal CVD and all-cause mortality.

112 **Methods**

113 *Development of the Free-Living, Ground Truth Annotated OxWalk Dataset*

114 To develop the OxWalk¹⁶ dataset, participants contributed activity data during unscripted, free
115 living. Participants wore four triaxial accelerometers (AX3, Axivity, Newcastle, UK), two placed
116 side-by-side on the dominant wrist and two clipped to the dominant-side hip at the midsagittal
117 plane. Accelerometers were synchronised using the Open Movement GUI software (v.1.0.0.42),
118 with one recording at 100 Hz and the other at 25 Hz at each body location. Final accelerometer
119 data was resampled to the nominal sampling rate and calibrated to local gravity using the Open
120 Movement software package. Foot-facing video was captured using an action camera (Action
121 Camera CT9500, Crosstour, Shenzhen, China) mounted at the participant's beltline
122 (Supplemental Figure 1). Participants were instructed to wear the camera for one hour and could
123 remove the camera any time they felt uncomfortable or required additional privacy¹⁷. To create
124 a clear, easily distinguishable data point for video and accelerometer synchronisation in this
125 study, participants were asked to strike their accelerometers together with four forceful blows
126 within camera view at the start of data collection¹⁸.

127 Ground truth annotation of steps was conducted within video annotation software (Elan 6.0, The
128 Language Archive, Nijmegen, Netherlands) by two independent annotators (SS and LvF) blinded
129 to each other's results. Similar to Bassett et al., we identified the act of lifting a foot and placing
130 it in a new location as a central tenant of step identification⁵. This definition was used as the
131 framework for step annotation in the OxWalk dataset, with an annotated step being a
132 repositioned foot linked to a change in gross body position along the floor. Annotated steps did
133 not include foot shuffling, changing of foot alignment via pivoting, or shifting of weight from one
134 foot to the other. Ethical approval for participant recruitment was obtained from the Central
135 University Research Ethics Committee of the University of Oxford (Ref: R63137/RE001). Written
136 informed consent was obtained from adult volunteers (aged 18 and above) with no lower limb
137 injury within the previous six months and who were able to walk without an assistive device.

138

139 *Model Development and Evaluation*

140 To develop the proposed step count model, a hybrid machine learning and peak detection
141 algorithm was created wherein an activity classification model was first used to detect periods of
142 walking and non-walking, followed by step counting only on predicted walking data epochs
143 (Figure 1). Activity classification was performed using a self-supervised deep learning model
144 developed by Yuan et al²⁰ incorporating an 18-layer ResNet-V2²¹ pre-trained using self-
145 supervised tasks on the UK Biobank accelerometer dataset. This pre-training step has previously
146 demonstrated consistent performance improvement for downstream activity recognition tasks
147 against Random Forest activity classification²⁰. The pre-trained self-supervised learning model
148 was then trained for supervised gait classification using the OxWalk dataset, wherein training

149 data consisted of 10 second epochs of accelerometer data with ground-truth walk or non-walk
150 labels. In the OxWalk dataset, walking was defined as at least four steps within the 10 second
151 epoch. Ten-fold cross-validation was used to train and validate the walking activity classifier and
152 evaluate end-to-end performance of the step detection pipeline. The participant dataset was
153 divided into 10 equal random folds where one fold was left out for testing and the remaining
154 folds underwent a randomised 80%-20% split for training and validation, respectively. Folds were
155 stratified by class label and data was grouped by participant. The self-supervised learning model
156 was trained on the remaining set with an early-stopping mechanism on the validation set when
157 the loss stopped decreasing for 5 consecutive training epochs. The weights prior to early stopping
158 were used to perform activity prediction on the test and validation set. An additional data-
159 augmentation step was performed during training, whereby each triaxial training sample was
160 randomly transformed with a rotation along a random axis and the axes were switched in a
161 random order to make the model rotation invariant. The model was trained using PyTorch 1.12.1
162 and Adam optimisation²² with a learning rate of 0.0001. Weighted cross entropy loss was used,
163 with the class weights set in such a way that the balance of walking and non-walking segments
164 was 10% to 90%, respectively, bringing the class balance in line with 24-hour direct observation
165 during free living in a previously collected dataset²³. Finally, predictions on the validation set and
166 corresponding ground-truth labels were used to train a Hidden Markov Model smoother which
167 was then applied to the predictions in the test set.

168 Step counting was performed through peak detection on classified walking time windows using
169 the “find_peaks” method from the SciPy Python package²⁴. Euclidean norm of triaxial
170 acceleration, minus 1 *g* to remove the effect of gravity, was clipped between ± 2 *g* and lowpass

171 filtered at 5 Hz prior to use as the input signal for peak detection. The “find_peaks” method
172 detects local peaks using predefined heuristics including the minimum peak height (prominence),
173 maximum peak width (width), and minimum time between peaks (distance). These heuristics
174 served as detection hyperparameters for which optimal values would minimise the mean
175 absolute error for step count in the validation set. Detection parameters were iterated across a
176 pre-selected range of values (prominence: 0.1 to 1 g; distance: 0.2 to 2 s; width: 10 ms to 1 s).
177 Model performance metrics were calculated on participants within each test set; mean precision,
178 recall, F1, Cohen’s kappa, and accuracy were used to evaluate walking classification, while MAPE
179 and mean bias and Spearman’s rank correlation coefficient were calculated against ground truth
180 step annotations. Following internal model validation, the final activity prediction model was
181 retrained on the entire OxWalk dataset with an 80%-20% training-validation split prior to external
182 deployment.

183 *External Model Validation*

184 External model performance was assessed by applying the step detection algorithm to wrist-
185 worn accelerometer data to an open-source, step-annotated dataset from Clemson University¹⁹.
186 Within this external dataset, 30 participants contributed a mean of 37 minutes of activity, split
187 between three distinct sessions of regular walking (two laps around a predefined path),
188 semiregular walking (locating objects throughout a building), and irregular walking (collecting
189 and assembling building blocks distributed around a room). Participants were video recorded
190 throughout scripted activities, allowing timestamp-annotated steps while wearing Shimmer3
191 inertial measurement units (Shimmer, Dublin, Ireland) recording at 15 Hz. Researchers annotated

192 steps as well as “shifts”, foot movement not necessarily tied to a change in body position, though
193 these annotated shifts were not included in the current analysis¹⁹. Prediction error was quantified
194 by calculating MAPE and mean percent under/overcounting bias for each gait subtype and
195 overall, at the participant level, across all gait subtypes. Bland-Altman plots were created for
196 comparison between cumulative ground truth and predicted step counts for each participant.

197 *Open-source Step Count Algorithm Assessment*

198 In addition to assessment of the novel algorithm, two additional step counting approaches were
199 evaluated in this study using both the OxWalk and Clemson datasets: 1) a recently-published
200 acceleration-threshold algorithm by Ducharme et al.⁸, and 2) the Verisense algorithm, a popular
201 open-source peak detection algorithm developed from the Clemson dataset²⁵ and previously
202 applied to UK Biobank accelerometer data using integration with the GGIR package^{26,27}. Further
203 details for these algorithms are presented in Supplemental Note 1, while details of all datasets
204 used are presented in Supplemental Table 1.

205 *Model Implementation into the UK Biobank*

206 The UK Biobank is a prospectively recruited observational cohort of over 500,000 participants
207 aged 40–69 at the time of recruitment, from 2006–2010²⁸. From 2013–2015, participants were
208 invited to wear an Axivity AX3 accelerometer on their dominant wrist, recording at 100 Hz, for a
209 seven-day, 24 hours per day activity measurement window. In the current study, raw
210 accelerometer data was processed from 103,391 available participants, after which data was
211 excluded from participants with fewer than 72 hours of wear, those lacking data across the entire
212 diurnal cycle, with poor device calibration, or with unrealistic average acceleration (>100 mg)⁴.

213 The externally validated hybrid SSL step detection model was applied to raw accelerometer data
214 from the UK Biobank. Overall daily step count was reported as the median number of steps taken
215 across the seven-day measurement period. Missing step count data from non-wear was imputed
216 by averaging step count from the corresponding time of day in all other valid days, similar to the
217 imputation of vector magnitude acceleration during non-wear in the UK Biobank physical activity
218 cohort⁴. One-minute peak cadence was calculated as previously described by Saint-Maurice et
219 al²⁹.

220 *Statistical Analysis*

221 UK Biobank participants with prevalent cardiovascular disease or cancer as a primary diagnosis,
222 as identified by International Classification of Diseases (ICD) codes I00–I99 and C00–C97 in their
223 routine hospital data, were removed from analysis. Spearman’s rank correlation (r) was
224 calculated between step count, peak cadence, overall acceleration, and UK Biobank derived
225 activity time use activity classification²³. Daily step count and one minute peak cadence were
226 stratified across demographic and self-reported health variables as collected by the UK Biobank
227 at the time of enrolment. Analysis of variance and Tukey Honestly Significant Difference tests
228 were conducted to compare step count based on self-reported health and usual walking pace.

229 Multivariable adjusted estimates of the effect of quintiles of step count on the relative hazards
230 of cardiovascular mortality and all-cause mortality were derived using Cox proportional hazards
231 regression using age as the underlying timescale^{30,31}. Date and cause of death was gathered from
232 the UK Biobank linked death registry. Length of follow-up was calculated from censoring dates
233 from the data sources or date of death. Further detail is provided in Supplemental Notes 2-3.

234 Step count detection was deployed on the UK Biobank using the University of Oxford Biomedical
235 Research Computing cluster, while statistical analysis was completed using R (v.4.1.1) on the UK
236 Biobank Research Analysis Platform. Statistical code is available at
237 https://github.com/OxWearables/UKB_steps_mortality.

238 **Results**

239 *Step Count Validation in the OxWalk Dataset*

240 Accelerometer and ground truth camera data was collected from 39 participants (19 female, 20
241 male) with a mean age of 38.5 years (range 19.5 to 81.2 years), a mean wear time of 58 minutes,
242 and a median [interquartile range (IQR)] 863 [312–2,123] steps within the measurement period.
243 Thirty-three participants were annotated by both annotators, resulting in a corresponding step
244 count MAPE of 4.0% and interclass correlation coefficient of 1.0 between annotators. Internal
245 validation of the self-supervised learning model identified bouts of walking with a Cohen’s Kappa
246 performance of 0.79 (Supplemental Table 3). Overall cross-validation of step detection in the self-
247 supervised learning model resulted in a 12.5% MAPE, 1.3% underestimation of steps, and
248 correlation of $r = 0.98$ against ground truth in the free-living OxWalk dataset. For comparison,
249 external validation of the step counting of the 100 Hz OxWalk wrist-worn dataset using the
250 Ducharme acceleration-threshold algorithm⁸ resulted in a 69.1% overestimation of steps (231.3
251 % MAPE, $r = 0.91$) across all participants. External validation of the Verisense algorithm^{10,25},
252 incorporated into recent UK Biobank papers^{14,26}, produced a 63.5% MAPE, 7.2% underestimation
253 bias, and $r = 0.85$ against free-living ground truth step counts (Supplemental Table 2). Bland-
254 Altman plots for model comparisons against ground truth OxWalk step count are presented in

255 Figure 2, demonstrating lower variability and tighter agreement with ground truth using the
256 novel step detection algorithm in the free-living dataset.

257 *Step Count Validation in the Clemson Dataset*

258 Bland-Altman plots for the performance of each prediction method in the overall Clemson
259 dataset are also presented in Figure 2. This plot again demonstrates reduced variability and bias
260 against ground truth using the novel model compared to reference algorithms. In external
261 validation, the threshold model by Ducharme et al.⁸ performed well during sessions of regular
262 gait, but poorly irregular gait, culminating in an overall MAPE of 47.5% and a 46.9%
263 overestimation of steps at the participant-level, across all gait subtypes. The Verisense algorithm,
264 for which this dataset serves as an internal validation, demonstrated a 17.6% underestimation of
265 steps and a 17.3% per-participant MAPE over all gait subtypes, including 16.3% MAPE during
266 regular walking (Supplemental Table 4). External validation of our novel self-supervised learning
267 hybrid step algorithm performed best in the Clemson dataset, producing a 16.5% MAPE and
268 16.6% underestimation across all gait subtypes, including 9.2% MAPE during regular walking. Due
269 to superior performance in free-living and laboratory-based validation, the SSL step detection
270 model was selected for analysis of UK Biobank data.

271 *Step Counts in the UK Biobank Physical Activity Cohort*

272 Baseline data from 75,493 UK Biobank participants without prevalent CVD or cancer is presented
273 in Table 1 and Supplemental Figure 2. Peak step cadence demonstrated expected variations by
274 self-reported usual walking pace (Supplemental Figure 3) and our measurements of steps
275 demonstrated orthogonality to standard overall acceleration and time-use metrics
276 (Supplemental Figure 4). Participants that self-reported that their overall health was excellent

277 were more active than all other participants, taking 2,947 more steps [95% CI 2,678–3,215] ($p <$
278 0.001) than those reporting that their overall health was poor. Similarly, self-reported brisk
279 walkers had a peak one-minute cadence 11.2 steps per minute [95% CI 10.6–11.7] ($p < 0.001$)
280 higher than slow walkers. Adjusted mean daily step counts by self-reported health status and by
281 selected physician-diagnosed chronic conditions are presented in Figure 3.

282 *Association of Step Counts with All-Cause and Cardiovascular Mortality*

283 The Cox regression analysis cohort had a median follow-up of 6.9 [IQR 6.3–7.4] years, with 572
284 events in the CVD mortality analysis and 1,844 events in the all-cause mortality analysis (Figure
285 4). For CVD mortality, a curvilinear association was observed with a linear association observed
286 between the first and third fifths of the step count distribution and then a flattening of the
287 association for the top two fifths of the distribution. For example, a median daily step count of
288 8,474 to 10,284 steps per day was associated with a 56% [43–66%] lower risk of CVD mortality
289 compared to participants taking fewer than 6,596 steps per day, whereas taking 12,677 or more
290 steps was associated with a 56% [43–66%] lower risk on CVD mortality. Similar results were
291 observed in the analysis of all-cause mortality and median daily step count, with a 39% [30–47%]
292 and 43% [34–51%] lower risk of all-cause mortality in the middle and most active 20%,
293 respectively.

294 **Discussion**

295 We have developed a new open-source step counting method, informed by self-supervised
296 machine learning methods that substantially outperforms current wrist-worn step counting
297 algorithms in the free-living environment. The open data and code released with this manuscript
298 will provide the global research community access to a more transparent and well-validated

299 method to measure steps in large-scale wrist-worn accelerometer datasets. When applying the
300 algorithm and resulting step metric in epidemiological analysis, we demonstrated that a higher
301 daily step count is associated with a lower risk of all-cause and cardiovascular mortality.

302 Our novel approach of using a hybrid step detection model that involves self-supervised machine
303 learning outperformed existing wrist-worn step counting methods, producing a 12.5% MAPE and
304 1.3% step underestimation during free living. Wrist-worn step counting is highly popular in both
305 commercial and research applications, but valid step detection at the wrist can be associated
306 with high measurement error relative to ground truth. In 2018, Toth et al.⁶ assessed wrist-worn
307 step detection in free-living conditions, finding error rates between 18% and 120% across a range
308 of methodologies. We found similar performance in current open-source algorithms during free-
309 living testing, with a mean average percent error ranging from 64% to 231%. Even while analysing
310 data from a different device and sampling rate, external validation of the novel model in the
311 Clemson dataset demonstrated a 9.2% error during regular walking in the laboratory-based
312 setting, below the 10% MAPE threshold required during treadmill-based validation¹¹. External
313 validation of the novel model outperformed both reference algorithms, including the Verisense
314 algorithm, which was trained and tuned using the Clemson laboratory dataset^{10,25}.

315 This study demonstrates a strong inverse curvilinear association between increased step count
316 and lower risk of fatal CVD and all-cause mortality while highlighting the importance of accurate
317 step detection algorithms in epidemiological analysis. Our current results parallel those of Paluch
318 et al.¹⁵, who demonstrated higher daily step counts are associated with an incrementally lower
319 risk of all-cause mortality across 15 international longitudinal cohorts nearly exclusively using hip-
320 mounted devices. Using less accurate step-detection methods, another study has also indicated

321 a curvilinear association between daily steps and CVD mortality²⁷. Though the direction of
322 epidemiological associations may remain broadly similar across step detection algorithms, it is
323 important that algorithms derive step counts as accurately as possible. Accurate step counting
324 will be particularly important when translating results into target levels of physical activity in
325 guidelines compatible with device-measured activity³². Reporting of inaccurate step counts may
326 additionally be demotivating and counterproductive in terms of health metrics and behavioural
327 change for individuals monitoring their own physical activity³³.

328 Clear strengths of our study include the development of a step counting algorithm trained in a
329 large dataset of free-living, wrist-worn accelerometer data with doubly-annotated ground truth
330 video and demonstrated high accuracy. While this training data consisted of short 1-hour data
331 collection windows, it is important to note that the current study algorithm is trained on one of
332 the most complete free-living, open-source datasets to date. Some overestimation of step counts
333 may occur when applied to multiday protocols due to the lack of extended periods of sedentary
334 inactivity in the short training data, however; class rebalancing was utilised to minimise this
335 effect. In the future, it will be important to further assess the robustness of this method across a
336 variety of populations and against 24-hour free-living, ground truth annotated step count data.

337 **Conclusions**

338 We have developed a new, open, and transparent method that markedly improves the ability to
339 measure steps in large-scale wrist-worn accelerometer datasets. While using this validated step
340 detection method trained using free-living data, we demonstrate an inverse dose response of
341 daily step count with all-cause and cardiovascular disease mortality. This reinforces public health

342 messaging of “the more, the better” approaches toward step count guidelines, encouraging any
343 increase in physical activity, particularly in populations wherein a specific target number of daily
344 steps may be unrealistic or feel unreachable.

345 **Data Availability**

346 The OxWalk dataset generated in this study is available for download and free for use through
347 the Oxford University Research Archive ([https://ora.ox.ac.uk/objects/uuid:19d3cb34-e2b3-](https://ora.ox.ac.uk/objects/uuid:19d3cb34-e2b3-4177-91b6-1bad0e0163e7)
348 [4177-91b6-1bad0e0163e7](https://ora.ox.ac.uk/objects/uuid:19d3cb34-e2b3-4177-91b6-1bad0e0163e7)). An open-source accelerometer processing tool integrating the
349 hybrid machine learning step detection method derived in this study will be available for use at
350 <https://github.com/OxWearables/stepcount>.

351

352

353

354

355

356

357

358

359

360

361 References

- 362 1 World Health Organization. 2020 WHO guidelines on physical activity and sedentary
363 behavior. Geneva: World Health Organization, 2020.
- 364 2 Khurshid S, Weng L-C, Nauffal V, *et al.* Wearable accelerometer-derived physical activity
365 and incident disease. DOI:10.1038/s41746-022-00676-9.
- 366 3 Barker J, Smith Byrne K, Doherty A, *et al.* Physical activity of UK adults with chronic
367 disease: cross-sectional analysis of accelerometer-measured physical activity in 96 706
368 UK Biobank participants. *Int J Epidemiol* 2019; published online Feb 5.
369 DOI:10.1093/ije/dyy294.
- 370 4 Doherty A, Jackson D, Hammerla N, *et al.* Large scale population assessment of physical
371 activity using wrist worn accelerometers: The UK biobank study. *PLoS One* 2017; **12**: 1–
372 14.
- 373 5 Bassett DR, Toth LP, LaMunion SR, Crouter SE. Step Counting: A Review of Measurement
374 Considerations and Health-Related Applications. *Sport. Med.* 2017; **47**: 1303–15.
- 375 6 Toth LP, Park S, Springer CM, FEYERABEND MD, Steeves JA, Bassett DR. Video-Recorded
376 Validation of Wearable Step Counters under Free-living Conditions. *Med Sci Sports Exerc*
377 2018; **50**: 1315–22.
- 378 7 Johnston W, Judice PB, García PM, *et al.* Recommendations for determining the validity
379 of consumer wearable and smartphone step count: Expert statement and checklist of the
380 INTERLIVE network. *Br J Sports Med* 2020; **0**: 1–14.
- 381 8 Ducharme SW, Lim J, Busa MA, *et al.* A Transparent Method for Step Detection Using an
382 Acceleration Threshold. *J Meas Phys Behav* 2021; **1**: 1–10.
- 383 9 Femiano R, Werner C, Wilhelm M, Eser P. Validation of open-source step-counting
384 algorithms for wrist-worn tri-axial accelerometers in cardiovascular patients. *Gait*
385 *Posture* 2022; **92**: 206–11.
- 386 10 Maylor BD, Edwardson CL, Dempsey PC, *et al.* Stepping towards More Intuitive Physical
387 Activity Metrics with Wrist-Worn Accelerometry: Validity of an Open-Source Step-Count
388 Algorithm. *Sensors* 2022, Vol 22, Page 9984 2022; **22**: 9984.
- 389 11 ANSI/CTA-2056. Physical Activity Monitoring for Fitness Wearables: Step Counting.
390 Consumer Technology Association, 2016
391 <https://webstore.ansi.org/standards/ansi/cta20562016ansi>.
- 392 12 Feito Y, Bassett DR, Thompson DL, Y F, DR B, DL T. Evaluation of activity monitors in
393 controlled and free-living environments. *Med Sci Sports Exerc* 2012; **44**: 733–41.
- 394 13 Mora-Gonzalez J, Gould ZR, Moore CC, *et al.* A catalog of validity indices for step
395 counting wearable technologies during treadmill walking: the CADENCE-adults study. *Int*
396 *J Behav Nutr Phys Act* 2022 191 2022; **19**: 1–16.
- 397 14 Del Pozo Cruz B, Ahmadi MN, Lee IM, Stamatakis E. Prospective Associations of Daily
398 Step Counts and Intensity with Cancer and Cardiovascular Disease Incidence and
399 Mortality and All-Cause Mortality. *JAMA Intern Med* 2022; : 1DUMMY.
- 400 15 Paluch AE, Bajpai S, Bassett DR, *et al.* Daily steps and all-cause mortality: a meta-analysis
401 of 15 international cohorts. *Lancet Public Heal* 2022; **7**: e219–28.
- 402 16 Small SR, von Fritsch L, Doherty A, Khalid S, Price AJ. OxWalk: Wrist and hip-based activity
403 tracker dataset for free-living step detection and gait recognition. 2022.

- 404 DOI:10.5287/bodleian:ORQ2abnbR.
- 405 17 Kelly P, Marshall SJ, Badland H, *et al*. An ethical framework for automated, wearable
406 cameras in health behavior research. *Am J Prev Med* 2013; **44**: 314–9.
- 407 18 Fortune E, Lugade V, Morrow M, Kaufman K. Validity of using tri-axial accelerometers to
408 measure human movement - Part II: Step counts at a wide range of gait velocities. *Med*
409 *Eng Phys* 2014; **36**: 659–69.
- 410 19 Mattfeld R, Jesch E, Hoover A. Evaluating pedometer algorithms on semi-regular and
411 unstructured gaits. *Sensors* 2021; **21**: 13–6.
- 412 20 Yuan H, Chan S, Creagh AP, Tong C, Clifton DA, Doherty A. Self-supervised Learning for
413 Human Activity Recognition Using 700,000 Person-days of Wearable Data. *arXiv* 2022;
414 published online June 6. DOI:10.48550/arxiv.2206.02909.
- 415 21 He K, Zhang X, Ren S, Sun J. Identity mappings in deep residual networks. *Lect Notes*
416 *Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)* 2016; **9908**
417 **LNCS**: 630–45.
- 418 22 Kingma DP, Ba JL. Adam: A method for stochastic optimization. In: 3rd International
419 Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings.
420 International Conference on Learning Representations, ICLR, 2015.
421 <https://arxiv.org/abs/1412.6980> (accessed Dec 21, 2022).
- 422 23 Walmsley R, Chan S, Smith-Byrne K, *et al*. Reallocation of time between device-measured
423 movement behaviours and risk of incident cardiovascular disease. BMJ Publishing Group
424 Ltd and British Association of Sport and Exercise Medicine, 2021.
- 425 24 Virtanen P, Gommers R, Oliphant TE, *et al*. SciPy 1.0: fundamental algorithms for
426 scientific computing in Python. *Nat Methods* 2020 173 2020; **17**: 261–72.
- 427 25 Patterson MR. Verisense-Toolbox/Verisense_step_algorithm at master ·
428 ShimmerEngineering/Verisense-Toolbox.
429 [https://github.com/ShimmerEngineering/Verisense-](https://github.com/ShimmerEngineering/Verisense-Toolbox/tree/master/Verisense_step_algorithm)
430 [Toolbox/tree/master/Verisense_step_algorithm](https://github.com/ShimmerEngineering/Verisense-Toolbox/tree/master/Verisense_step_algorithm) (accessed Oct 20, 2022).
- 431 26 Del Pozo Cruz B, Ahmadi M, Naismith SL, Stamatakis E. Association of Daily Step Count
432 and Intensity with Incident Dementia in 78430 Adults Living in the UK. *JAMA Neurol*
433 2022. DOI:10.1001/jamaneurol.2022.2672.
- 434 27 Del Pozo Cruz B, Ahmadi MN, Lee ; I-Min, Stamatakis E. Prospective Associations of Daily
435 Step Counts and Intensity With Cancer and Cardiovascular Disease Incidence and
436 Mortality and All-Cause Mortality Supplemental content. *JAMA Intern Med* 2022;
437 published online Sept 12. DOI:10.1001/JAMAINTERNMED.2022.4000.
- 438 28 Sudlow C, Gallacher J, Allen N, *et al*. UK Biobank: An Open Access Resource for
439 Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS*
440 *Med* 2015; **12**: e1001779.
- 441 29 Saint-Maurice PF, Troiano RP, Bassett DR, *et al*. Association of Daily Step Count and Step
442 Intensity with Mortality among US Adults. *JAMA - J Am Med Assoc* 2020; **323**: 1151–60.
- 443 30 Cox DR. Regression Models and Life-Tables. *J R Stat Soc Ser B* 2016; **15**: 1–23.
- 444 31 Cologne J, Hsu WL, Abbott RD, *et al*. Proportional hazards regression in epidemiologic
445 follow-up studies: An intuitive consideration of primary time scale. *Epidemiology* 2012;
446 **23**: 565–73.
- 447 32 Thompson D, Batterham AM, Peacock OJ, Western MJ, Booso R. Feedback from physical

448 activity monitors is not compatible with current recommendations: A recalibration study.
449 *Prev Med (Baltim)* 2016; **91**: 389–94.

450 33 Zahrt OH, Evans K, Murnane E, *et al*. Effects of Wearable Fitness Trackers and Activity
451 Adequacy Mindsets on Affect, Behavior, and Health: Longitudinal Randomized Controlled
452 Trial. *J Med Internet Res* 2023;25e40529 <https://www.jmir.org/2023/1/e40529> 2023; **25**:
453 e40529.

454 34 Gu F, Khoshelham K, Shang J, Yu F, Wei Z. Robust and accurate smartphone-based step
455 counting for indoor localization. *IEEE Sens J* 2017; **17**: 3453–60.

456 35 Lenth R V. emmeans: Estimated Marginal Means, aka Least-Squares Means. 2022.
457 <https://cran.r-project.org/package=emmeans>.

458 36 Terry M. Therneau, Patricia M. Grambsch. Modeling Survival Data: Extending the Cox
459 Model. New York: Springer, 2000.

460 37 Easton D, Peto J, AG B. Floating absolute risk: an alternative to relative risk in survival and
461 case-control analysis avoiding an arbitrary reference group. *Stat Med* 1991; **10**: 1025–35.

462 38 Carstensen B, Plummer M, Laara E, Hills M. Epi: A Package for Statistical Analysis in
463 Epidemiology. 2022. <https://cran.r-project.org/package=Epi>.
464
465

466

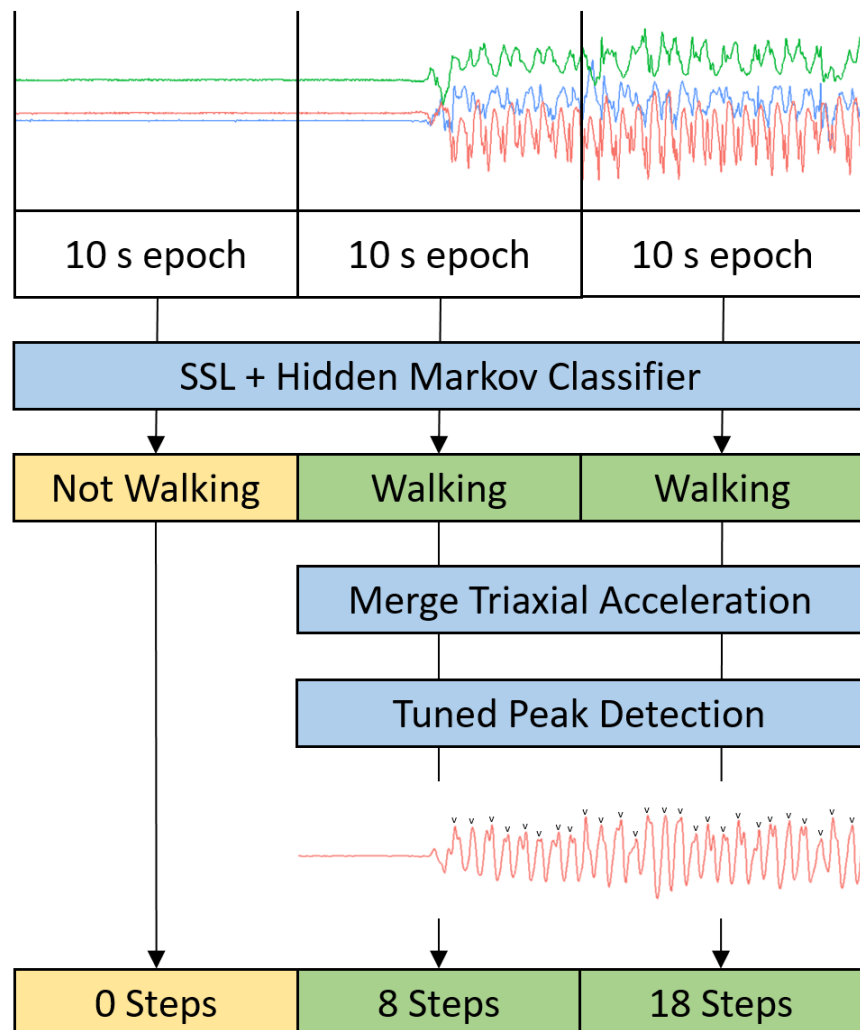
467

468

469

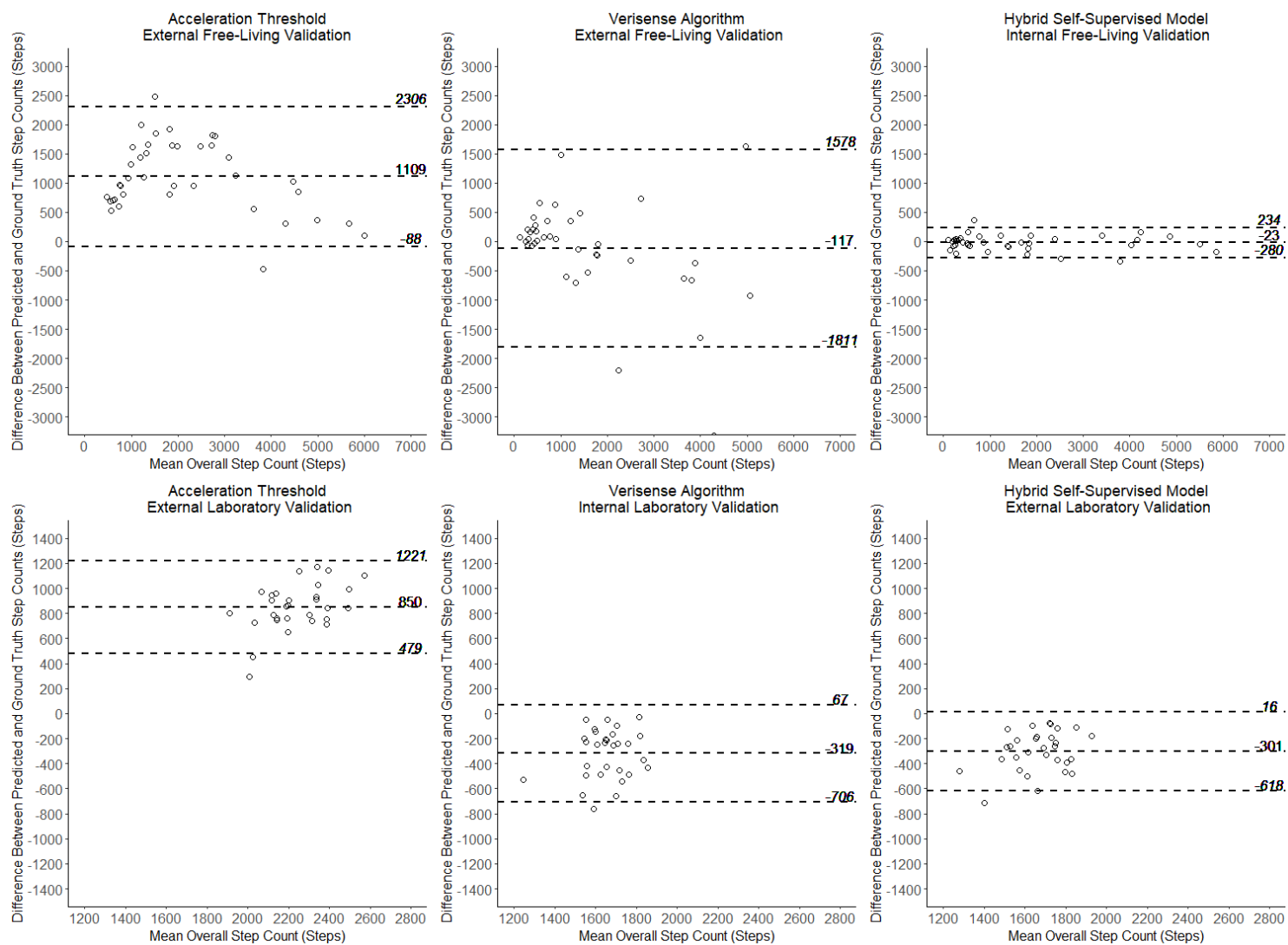
470

471



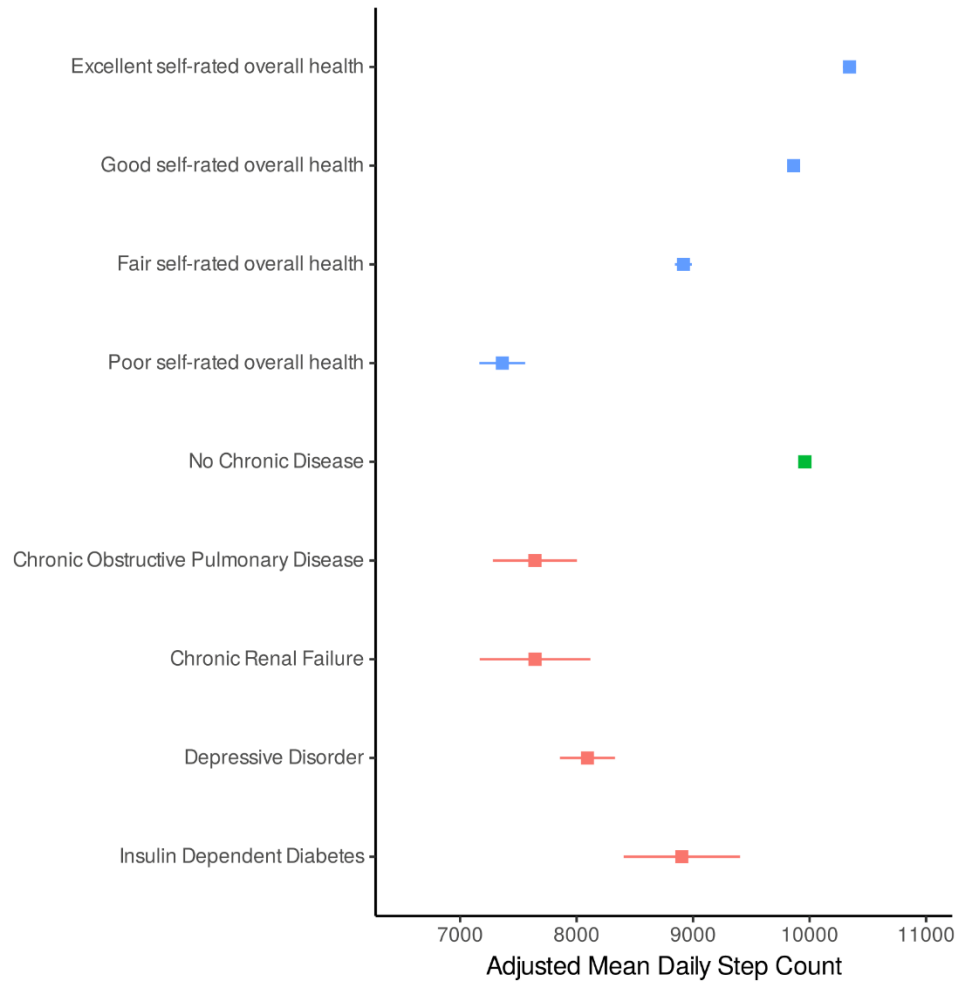
472

473 **Figure 1: Schematic of the process for generating step count from 30 seconds of raw triaxial**
474 **accelerometer data using a hybrid self-supervised learning (SSL) and peak detection step**
475 **counting model.**



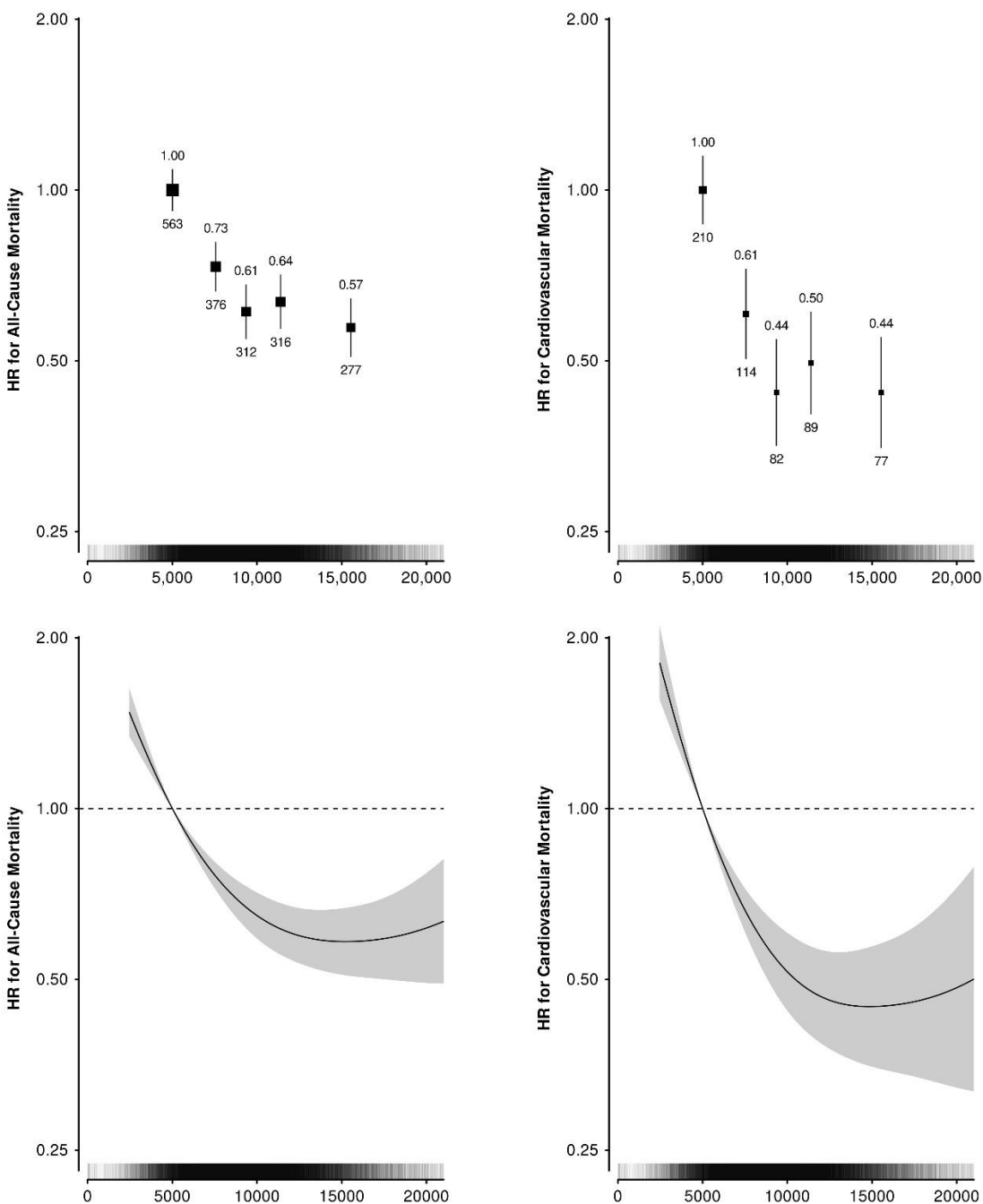
476

477 **Figure 2: Bland-Altman plots with dotted 95% limits of agreement for the comparison of step counting models in the (Top) OxWalk**
 478 **free-living dataset of 39 adults and (Bottom) Clemson laboratory-based dataset of 30 young adults. Left: baseline acceleration**
 479 **threshold model⁸, Centre: Verisense algorithm²⁵, and Right: the novel hybrid self-supervised learning model.**



480

481 **Figure 3: Adjusted estimated marginal mean (95% confidence interval) daily step count**
482 **according to self-reported overall health status, hospital data derived chronic disease status,**
483 **and select diagnoses for 75,493 UK Biobank participants. Mean daily step counts are adjusted**
484 **for age and sex.**



485

486 **Figure 4: (Top) Forest plots for all-cause mortality and cardiovascular disease mortality**
 487 **associations with quintiles of daily step count, (Bottom) continuous daily step count for**
 488 **75,493 UK Biobank participants.** Hazard ratios (HR) and 95% confidence intervals were
 489 calculated using age as a timescale, adjusted for sex, ethnicity, education, alcohol intake,
 490 smoking status, Townsend deprivation index, processed meat intake, fresh fruit intake, oily fish
 491 intake, and added salt intake. HR is above and number of events is plotted below each data
 492 point. Spline plot of hazard ratio and 95% confidence interval of the association of continuously
 493 modelled median daily step count. Vertical bars along the step axis indicate distribution of
 494 participant daily step counts.

Table 1: Overall Physical Activity Metrics by Demographic Characteristic in the UK Biobank

Characteristic	N (%)	Daily Steps	Peak Cadence (Steps per minute)	Overall Acceleration (mg)
Overall	75,493 (100.0)	9,352 [7,099-11,973]	117 [111-122]	27.5 [22.9-32.9]
Sex				
Female	43,802 (58.0)	9,267 [7,050-11,840]	118 [113-124]	27.8 [23.3-33.2]
Male	31,691 (42.0)	9,468 [7,182-12,159]	114 [109-119]	27.0 [22.3-32.6]
Age, years				
40-49	7,229 (9.6)	9,412 [7,181-12,047]	119 [113-125]	30.4 [25.5-36.5]
50-59	23,390 (31.0)	9,348 [7,136-12,014]	118 [113-124]	29.1 [24.4-34.8]
60-69	32,903 (43.6)	9,497 [7,222-12,122]	116 [110-122]	26.9 [22.5-32.1]
70-79	11,971 (15.9)	8,925 [6,667-11,374]	114 [108-120]	24.6 [20.5-29.2]
Ethnicity				
Nonwhite	2,380 (3.2)	9,050 [6,764-11,691]	118 [111-124]	28.7 [23.9-34.2]
White	73,113 (96.8)	9,362 [7,115-11,981]	116 [111-122]	27.4 [22.9-32.9]
Body Mass Index				
Underweight (<18.5 kg/m ²)	444 (0.6)	10,000 [7,972-13,120]	121 [115-127]	31.3 [25.2-36.8]
Normal weight (18.5-24.9 kg/m ²)	30,312 (40.2)	9,923 [7,696-12,526]	119 [113-125]	29.5 [24.7-35.1]
Overweight (25.0-29.9 kg/m ²)	30,791 (40.8)	9,363 [7,174-11,943]	116 [110-121]	27.0 [22.7-32.1]
Obese (30+ kg/m ²)	13,946 (18.5)	7,956 [5,896-10,457]	113 [107-119]	24.4 [20.3-29.2]
Education				
School Leaver	16,710 (22.1)	8,933 [6,749-11,532]	116 [110-122]	27.0 [22.3-32.5]
Further Education	25,052 (33.2)	9,172 [6,908-11,846]	116 [110-122]	27.5 [22.9-32.9]
Higher Education	33,731 (44.7)	9,679 [7,454-12,248]	117 [112-123]	27.7 [23.2-33.1]
Smoking Status				
Never	44,231 (58.6)	9,422 [7,200-12,007]	117 [112-123]	27.8 [23.2-33.2]
Former	26,107 (34.6)	9,329 [7,043-11,984]	116 [110-122]	27.3 [22.7-32.7]
Current	5,155 (6.8)	8,784 [6,532-11,547]	114 [109-120]	26.3 [21.5-31.8]
Alcohol Consumption				
Never	4,086 (5.4)	8,906 [6,448-11,620]	116 [109-122]	26.8 [21.8-32.4]
< 3 Days Per Week	34,319 (45.5)	9,031 [6,810-11,622]	117 [111-122]	27.3 [22.6-32.7]
3+ Days Per Week	37,088 (49.1)	9,685 [7,475-12,281]	117 [111-122]	27.8 [23.3-33.2]
Townsend Deprivation				
Least Deprived (<-3.8)	18,854 (25.0)	9,332 [7,200-11,945]	116 [110-122]	27.6 [23.1-32.9]
Second Least Deprived (-3.8 to -2.2)	18,892 (25.0)	9,326 [7,145-11,860]	116 [111-122]	27.5 [23.0-32.9]
Second Most Deprived (-2.5 to -1.2)	18,869 (25.0)	9,362 [7,080-11,951]	117 [111-122]	27.5 [22.9-32.9]
Most Deprived (≥-0.2)	18,878 (25.0)	9,384 [6,974-12,124]	117 [111-124]	27.4 [22.6-32.9]

495

Characteristic	N (%)	Daily Steps	Peak Cadence (Steps per minute)	Overall Acceleration (mg)
Self-Reported Usual Walking Pace				
Brisk	36,733 (48.7)	9,787 [7,567-12,405]	118 [113-124]	29.0 [24.4-34.6]
Steady	35,727 (47.3)	9,078 [6,890-11,653]	115 [110-121]	26.4 [22.0-31.4]
Slow	2,905 (3.8)	6,889 [4,582-9,495]	109 [102-116]	22.3 [18.1-27.2]
None of the above	61 (0.1)	5,773 [3,371-10,301]	107 [99-112]	22.5 [18.4-28.4]
Missing	67 (0.1)	3,148 [868-5,758]	96 [65-105]	18.9 [14.1-24.2]
Self-Reported Overall Health				
Excellent	17,781 (23.6)	9,855 [7,708-12,432]	118 [113-124]	29.2 [24.5-35.0]
Good	45,755 (60.6)	9,397 [7,178-12,017]	116 [111-122]	27.5 [23.0-32.7]
Fair	10,520 (13.9)	8,468 [6,208-11,099]	114 [108-120]	25.3 [20.9-30.4]
Poor	1,437 (1.9)	6,939 [4,394-9,626]	110 [103-117]	22.9 [18.4-28.1]
Wear Season				
Spring	17,327 (23.0)	9,479 [7,193-12,130]	117 [111-123]	27.9 [23.2-33.4]
Summer	20,014 (26.5)	9,812 [7,525-12,520]	116 [110-121]	28.1 [23.4-33.6]
Autumn	22,342 (29.6)	9,236 [7,040-11,804]	117 [111-123]	27.4 [22.9-32.8]
Winter	15,810 (20.9)	8,774 [6,647-11,288]	117 [111-123]	26.6 [22.1-31.7]

Activity metrics reported as unadjusted median [interquartile range]

496

497

498

499

500

501

502

503

504

505

506

507