

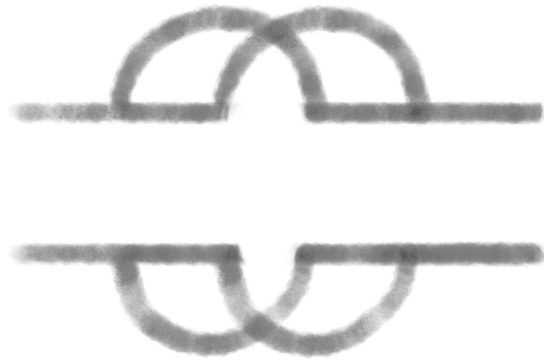
Use of Wrist-Worn Accelerometers to Predict Parkinson's Disease



Aidan Acquah
Linacre College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Hilary 2025

Acknowledgements



"Nyansapo wosane no badwenma" – the adinkra symbol shown above, signifies the untying of the knot by the wise. As Uncle Iroh said, "It is important to draw wisdom from different places. If you take it only from one place, it grows rigid and stale." Embodied within these pages is my four-year journey in the search for greater wisdom. However, this path was not taken alone, and I would like to thank all those who aided me along the way.

My friends were an incredible source of support, especially during the isolation of the COVID-19 lockdowns. With Sayap joining me for transcontinental calls and Astra for late-night Dungeons & Dragons sessions, I was able to stay connected and keep my spirits up. Rejoice and Thugz4Lyf were also there when I needed a break, joining me for spontaneous trips to London whenever the workload felt overwhelming.

My family were also there with me through it all. Our early morning calls became a source of positivity and motivation, helping me start each day with a fresh outlook.

Finally, I would like to thank my colleagues and collaborators, particularly members of the OxWearables group, CHI lab, and GSK digital biomarkers team. I would particularly like to thank my supervisors, Dr. Andrew Creagh, Prof. Aiden Doherty, and Prof. David Clifton for their continued support throughout this research. I would also like to thank Dr. Valentin Hamy, my GSK supervisor, for his advice and experience throughout these years.

To all of you, I extend my deepest gratitude for being part of this journey.

Abstract

Parkinson’s disease (PD) is a growing public health concern with rising diagnosis rates and no preventive treatments. Early-stage monitoring is crucial for understanding disease progression and smartwatches offer great potential in early detection to enable this. This DPhil thesis investigates how wrist-worn accelerometers can be used to help predict incident PD, through the detection of early markers of the disease.

In the initial phase of this work, I reviewed literature on PD and wearable monitoring technologies. This review highlighted research into walking pattern monitoring and proposed risk prediction models using accelerometer data. To conduct this research, a variety of datasets were used, including the UK Biobank Physical Activity Monitoring, PPMI-Verily, Capture-24, OxWalk, and MJFF Levodopa Response studies.

Wrist-worn accelerometer data depended on the development of reliable activity classification models for this thesis’ aims. Existing models were improved to detect labels of activity intensity (sleep, sedentary, light, and moderate-vigorous activity) using a modified ResNet-18 model, pre-trained using self-supervised learning. This was trained and evaluated on the Capture-24 dataset, using both a single train-test split, as seen in prior research, and using 5-fold group cross-validation. Training on the first 100 participants, and testing on the remaining 51 participants, the new ActiNet model outperformed the existing Accelerometer model with a mean per-participant macro F1 score of 0.806 and 0.786, respectively. When trained and evaluated using 5-fold group cross-validation, the ActiNet model achieved a mean macro F1 score of 0.827, outperforming the existing model’s score of 0.773. Walking recognition models were also enhanced to be trained on both healthy and PD datasets. Patterns showed that walking recognition models trained only on healthy participants performed worse in participants with increasing severity of PD. However, training the same models on both healthy and PD data improved general performance, with no drop in performance for higher severities of PD.

An epidemiological analysis was then conducted to investigate the association between device-measured daily step count and PD incidence in the UK Biobank, to assess low daily steps as a risk factor for PD. A 1,000-step increase in median daily steps resulted in a hazard ratio of 0.92 for PD diagnosis, consistent after accounting for various confounders. However, this association attenuated towards

null after the removal of the first six years of follow-up, giving an indication of reverse-causation driving this association. Therefore, daily steps appear to be a predictor of but not a risk factor for PD.

Subsequently, machine learning models trained to categorise activity intensity and walking were used to compare active walking windows among participants with different PD statuses. In the PPMI-Verily dataset, a self-supervised ResNet-18 model distinguished between diagnosed PD, prodromal PD, and healthy controls, achieving a mean macro F1 of 0.586 and an area under the receiver operator characteristic curve (AUROC) of 0.856 using the active walking windows. This finding establishes that wrist-worn accelerometers can detect differences in walking behaviour related to PD impairment.

The final phase of research combined a variety of digital measures extracted from wrist-worn accelerometers to build risk prediction models for PD in the UK Biobank. These wrist-worn accelerometer-derived digital measures were compared with traditional risk factors. The risk model with the highest discrimination performance incorporated all risk factors, including demographic, polygenic risk score, overall activity levels, activity composition, and a PD-like gait score, which compared detected active walking windows of participants to those of diagnosed PD participants in the PPMI-Verily dataset. After accounting for shrinkage, this model produced a C-index of 0.902, as compared to 0.778 when using non-accelerometer-based traditional demographic risk factors for PD. These results highlight the added benefit of wrist-worn accelerometry, particularly PD-specific digital measures, for improving PD prediction models.

In summary, this thesis leverages wrist-worn accelerometry and machine learning to predict PD, aiming to enhance early diagnosis. The results across the different studies support the promising use of wearable technology for PD.

Contents

List of Figures	xi
List of Tables	xv
List of Abbreviations	xvii
1 Introduction	1
1.1 Parkinson’s disease	1
1.1.1 Predictive measures of Parkinson’s disease	5
1.2 Wearable devices	9
1.2.1 Wrist-worn accelerometers	10
1.3 Wrist-worn accelerometers and Parkinson’s disease	12
1.3.1 Population health studies	13
1.4 Challenges	16
1.5 Thesis aims and objectives	17
1.5.1 Thesis research questions	17
1.5.2 Structure of the thesis	18
2 Datasets	19
2.1 Introduction	19
2.2 UK Biobank Physical Activity Monitoring Study	22
2.3 Parkinson’s Progression Markers Initiative Verily study	28
2.4 Capture-24	31
2.5 OxWalk	32
2.6 Michael J. Fox Foundation Levodopa Response study	34
2.7 Data generalisability	35
2.8 Dataset preparation	37
3 Improving Human Activity Recognition Models	39
3.1 Introduction	39
3.2 Actinet: Classification of Activity Intensity Using Self-Supervised Deep Learning	40
3.2.1 Introduction	40

3.2.2	Methods	44
3.2.3	Results	49
3.2.4	Discussion	57
3.2.5	Summary	60
3.3	Walking Recognition in Parkinson's Disease Populations	60
3.3.1	Introduction	60
3.3.2	Methods	62
3.3.3	Results	64
3.3.4	Discussion	67
3.3.5	Summary	72
4	Investigating the association between daily steps counts and incident Parkinson's disease	73
4.1	Introduction	74
4.2	Methods	75
4.2.1	Study Cohort	75
4.2.2	Accelerometer-derived physical activity processing	75
4.2.3	Ascertainment of Parkinson's disease	76
4.2.4	Statistical Analysis	76
4.3	Results	80
4.4	Discussion	83
4.5	Summary	86
5	Use of active walking windows to categorise Parkinson's disease	87
5.1	Introduction	88
5.2	Methods	89
5.2.1	Study population	90
5.2.2	Pre-processing	90
5.2.3	Human activity recognition	91
5.2.4	Parkinson's disease classification using active walking windows	94
5.2.5	Model Evaluation and Explainability	94
5.3	Results	96
5.3.1	Study population	96
5.3.2	Human activity recognition	97
5.3.3	PD Classification	98
5.3.4	Explainability	101
5.4	Discussion	103
5.4.1	Comparison with published literature	103
5.4.2	Strengths and Limitations	105
5.5	Summary	106

6	Use of wrist-worn accelerometers to improve Parkinson's disease risk models in the UK Biobank	107
6.1	Introduction	107
6.2	Methods	109
6.2.1	Study design and outcomes	109
6.2.2	Predictor variables	110
6.2.3	Model derivation and validation	112
6.3	Results	113
6.3.1	Study population and Baseline Characteristics	113
6.3.2	Incidence of Parkinson's disease	116
6.3.3	Predictor variables	116
6.3.4	Validation	116
6.4	Discussion	118
6.5	Summary	121
7	Discussion	125
7.1	Summary of findings	125
7.2	Ethical considerations	128
7.2.1	Ethical research conduct	128
7.2.2	Ethical implications of this research	129
7.3	Limitations	131
7.4	Future Work	133
7.5	Conclusion	135
Appendices		
A	Comparison of Accelerometer and ActiNet	139
A.1	HMM matrices	139
B	Median daily step count and incident Parkinson's disease	153
B.1	Power Calculations	174
C	Parkinson's disease risk prediction	175
	References	191

List of Figures

1.1	Illustration of markers of Parkinson’s disease related neurodegeneration	3
1.2	Illustration of Parkinson’s disease symptoms	5
1.3	Illustration of wrist-worn accelerometer device	10
2.1	Comparison of research and consumer grade device outputs	36
2.2	Sample accelerometer data from AX3 Axivity and GENEActiv watches	38
3.1	Illustration of HAR models	42
3.2	UMAPs showing the separation of different activities in Capture-24	43
3.3	Activity intensity HAR model confusion matrix comparison, 100:51 split	52
3.4	Activity intensity HAR model confusion matrix comparison, cross-validation	52
3.5	Bland-Altman plot for Capture-24 activity intensity classifications .	53
3.6	Performance of HAR models within subpopulations of Capture-24 .	54
3.7	Exclusion flow diagram for activity intensity HAR deployment in UK Biobank	55
3.8	Observed rate of HAR predicted activity intensities in UK Biobank	56
3.9	Bland-Altman plot for UK Biobank activity intensity classifications	57
3.10	Grouping of MJFF Levodopa Response participants by severity of PD	62
3.11	Performance of walking classification model trained on healthy population evaluated on participants with increased severity of PD . . .	65
3.12	Performance of walking classification models given an increased severity of PD	68
4.1	Illustration of hybrid step count algorithm	76
4.2	Exclusion flow diagram for steps PD association in UK Biobank . .	78
4.3	Association between quintiles of median daily step count and incident Parkinson’s disease	82
4.4	Association between incident PD and median daily steps by period of follow-up	84
5.1	Method flowchart for active walking classification	89

5.2	Activity calendar for example PPMI-Verily participant	91
5.3	Illustration of activity, walking and PD status classification	92
5.4	Exclusion flowchart for PPMI Verily study	97
5.5	Histogram for detected active walking in PPMI-Verily population	98
5.6	Plot of detected active walking of sample participant in PPMI-Verily population	99
5.7	Confusion matrix for SSL ResNet-18 model to classify PD status	101
5.8	t-SNE plot for active walking in PPMI-Verily population	102
5.9	Active walking samples	104
6.1	Exclusion flow diagram for PD risk analysis	114
6.2	Discrimination of PD risk models	117
A.1	Comparison of Accelerometer and ActiNet models deployed on sample Capture-24 participants	141
A.2	Difference in performance of HAR models for activity intensity	142
A.3	Confusion matrices for Capture-24 activity intensity classifications by age group	143
A.4	Confusion matrices for Capture-24 activity intensity classifications by sex	144
A.5	Bland-Altman plot for Capture-24 activity intensity classifications by age group	145
A.6	Bland-Altman plot for Capture-24 activity intensity classifications: females	146
A.7	Bland-Altman plot for Capture-24 activity intensity classifications: males	146
A.8	Sample activity intensity plots	148
A.9	Observed rate of predicted activity intensities by age	149
A.10	Observed rate of predicted activity intensities by sex	150
A.11	Bland-Altman plots for predicted activity intensity by age	151
A.12	Bland-Altman plots for predicted activity intensity: Females	152
A.13	Bland-Altman plots for predicted activity intensity: Males	152
B.1	Kaplan-Meier plot for incident Parkinson's disease	162
B.2	Schoenfeld residuals plot	163
B.3	The sequential adjustment of the association between median daily steps and incident PD	164
B.4	The association between median daily steps and incident PD after adjustment for potential confounding factors	165
B.5	The association between median daily steps and incident PD within various subpopulations	166

B.6	Comaprison of outputs of step count algorithms	167
B.7	PD trained algorithm: Schoenfeld residuals plot	168
B.8	PD trained algorithm: Association between quintiles of median daily step count and incident Parkinson's disease	169
B.9	PD trained algorithm: The sequential adjustment of the association between median daily steps and incident PD	170
B.10	PD trained algorithm: The association between median daily steps and incident PD after adjustment for potential confounding factors	171
B.11	PD trained algorithm: The association between median daily steps and incident PD within various subpopulations	172
B.12	PD trained algorithm: Association between incident PD and median daily steps by period of follow-up	173
C.1	Illustration PD risk analysis	178
C.2	Incidence rate for PD risk analysis	179
C.3	Correlation of accelerometer-derived predictors	179
C.4	Calibration curve: Demographic factors	180
C.5	Calibration curve: Demographic and genetic factors	180
C.6	Calibration curve: Overall activity	181
C.7	Calibration curve: Activity composition	181
C.8	Calibration curve: Overall activity and activity composition	182
C.9	Calibration curve: PD-like gait score	182
C.10	Calibration curve: All factors	183

List of Tables

2.1	Comparison of different datasets used within this thesis	21
2.2	UK Biobank censoring dates across different data providers and countries[66, 67, 68]	23
2.3	Characteristics of UK Biobank Physical Activity Monitoring study population	25
2.4	Estimation of prevalent PD cases in the UK Biobank Physical Activity Monitoring Cohort Data	26
2.5	Estimation of incident PD cases in the UK Biobank Physical Activity Monitoring cohort	27
2.6	Characteristics of the PPMI Verily population	30
2.7	MJFF-LR study activities	35
3.1	Name and descriptions of extracted features for accelerometer package	48
3.2	Internal validation on test participants	51
3.3	Internal validation using cross-validation	51
3.4	Summary of walking and non-walking frames in OxWalk and MJFF Levodopa Response datasets.	66
3.5	Performance of walking classification models	69
3.6	P-value from statistical comparison of walking classification models	70
4.1	UK Biobank Physical activity monitoring study population characteristics	81
5.1	List of models used in PD active walking windows analysis	94
5.2	Description of various features and their units	95
5.3	PD status classification performance	100
5.4	Relative cross tabulation of GMM cluster and PD predicted status - Ratio of each predicted PD status per cluster	102
6.1	Summary of predictive models	111
6.2	PD risk analysis baseline characteristics	115
6.3	Univariate Cox model hazard ratios and C-index for each predictor	122
6.4	Summary of Model Evaluation Metrics for PD risk	123

A.1	Test Participant IDs for each fold of the stratified group cross-validation on the Capture-24 dataset.	147
B.1	Definition of variables in UK Biobank data	159
B.2	STROBE outline for study investigating the association between median daily steps and incident PD	161
C.1	Variables used in Parkinson’s disease risk prediction analysis.	177
C.2	Parameters for trained PD risk models	187
C.3	TRIPOD-AI Checklist	190

List of Abbreviations

AUROC		Area Under the Receiver Operating Characteristic Curve
CI		Confidence Interval
ENMO		Euclidean Norm Minus One
HAR		Human Activity Recognition
HMM		Hidden Markov Model
ICD-10		International Classification of Diseases, 10th Revision
IQR		Interquartile Range
MJFF-LR	. .	Michael J. Fox Foundation Levodopa Response
MVPA		Moderate-Vigorous Physical Activity
PD		Parkinson’s Disease
PPMI		Parkinson’s Progression Markers Initiative
RF		Random Forest
SSL		Self-Supervised Learning
t-SNE		t-distributed Stochastic Neighbour Embedding
UKBB		UK Biobank

Contributions

The work presented in this thesis is my own work, under the supervision of Dr. Andrew Creagh, Prof. Aiden Doherty and Prof. David Clifton. This work was supported by GSK, and I had regular communication with Dr. Valentin Hamy, my GSK supervisor. The project was conceptualised by a shared effort between myself and all supervisors. I conducted all the data analysis, production of figures and tables, and the writing of the thesis.

Several others contributed to part of the work presented:

- My co-authors guided me in the data analysis and interpretation of the published results.
- Prof. Sarah Floud and Tonima Trisa provided guidance on the statistical approaches to addressing reverse causation.
- Adam Sturge laid out the groundwork for risk analysis in the UK Biobank.
- Prof. Angeliki Kerasidou provided feedback regarding the ethical considerations and implications of the work.

Outputs

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

- **Aidan Acquah**, Andrew Creagh, Valentin Hamy, Alaina Shreves, Charilaos Zisou, Charlie Harper, Stefan Van Duijvenboden, Chrystalina Antoniadou, Derrick Bennett, David Clifton, Aiden Doherty. "Daily steps are a predictor of, but perhaps not a modifiable risk factor for Parkinson's disease: Findings from the UK Biobank." *medRxiv* (2024). <https://doi.org/10.1101/2024.08.13.24311539>

I have also contributed to the following manuscripts:

- Scott R. Small, Shing Chan, Rosemary Walmsley, Lennart von Fritsch, **Aidan Acquah**, Gert Mertes, Benjamin G. Feakins, Andrew Creagh, Adam Strange, Charles Matthews, David Clifton, Andrew Price, Sara Khalid, Derrick Bennett, Aiden Doherty. "Self-supervised machine learning to characterise step counts from wrist-worn accelerometers in the UK Biobank." *Medicine & Science in Sports & Exercise*, 15 May 2024. https://journals.lww.com/acsm-msse/fulltext/2024/10000/self_supervised_machine_learning_to_characterize.9.aspx
- Hang Yuan, Shing Chan, Andrew P. Creagh, Catherine Tong, **Aidan Acquah**, David A. Clifton, and Aiden Doherty. "Self-supervised learning for human activity recognition using 700,000 person-days of wearable data." *npj Digital Medicine* 7, no. 1 (12 April 2024): 1–10. <https://doi.org/10.1038/s41746-024-01062-3>
- D. Mcgagh, **A. Acquah**, L. Portas, C. Zisou, A. Shreves, C. Harper, S. Fenton, A. Doherty, and L. C. Coates. "Ab0214 association between accelerometer-measured daily step count and incident rheumatoid arthritis: A prospective cohort study in the UK Biobank." *Annals of the Rheumatic Diseases* 83, no. Suppl 1 (1 June 2024): 1345–1345. <https://doi.org/10.1136/annrheumdis-2024-eular.2245>

- Shing Chan, Hang Yuan, Catherine Tong, **Aidan Acquah**, Abram Schonfeldt, Jonathan Gershuny, and Aiden Doherty. "CAPTURE-24: A large dataset of wrist-worn activity tracker data collected in the wild for human activity recognition." *Nature Scientific Data*, 16 October 2024. <https://www.nature.com/articles/s41597-024-03960-3>

I have authored/contributed to seven publicly available software packages:

- Author: `pypi:actinet`, available at: <https://pypi.org/project/actinet/>
- Author: `OxWearables/pd_walking`, available at: https://github.com/OxWearables/pd_walking
- Author: `OxWearables/Wearables_Activity_Recognition_R`, available at: https://github.com/OxWearables/Wearables_Activity_Recognition_R
- Contributor: `pypi:actipy`, available at: <https://pypi.org/project/actipy/>
- Contributor: `pypi:stepcount`, available at: <https://pypi.org/project/stepcount/>
- Contributor: `pypi:asleep`, available at: <https://pypi.org/project/asleep/>
- Contributor: `OxWearables/ssl-wearables`, available at: <https://github.com/OxWearables/ssl-wearables>
- Contributor: `OxWearables/Oxford_Wearables_Activity_Recognition`, available at: https://github.com/OxWearables/Oxford_Wearables_Activity_Recognition
- Contributor: `OxWearables/rap_wearables`, available at: https://github.com/OxWearables/rap_wearables

I have also presented my work externally:

- "Walking Recognition in Parkinson's Disease Populations Using Wrist-Worn Accelerometers" presented at the 2024 International Conference on Ambulatory Monitoring of Physical Activity and Movement in Rennes, France.
- "Patient-Specific Training Data to Improve Walking Recognition in Clinical Parkinson's Populations" presented at the International Society for the Measurement of Physical Behaviour 2023 Virtual Fall Symposium.
- "Use of wrist-worn accelerometry to predict Parkinson's disease" presented at Health Research from Home Webinar 5, December 2024.

1

Introduction

Contents

1.1 Parkinson's disease	1
1.1.1 Predictive measures of Parkinson's disease	5
1.2 Wearable devices	9
1.2.1 Wrist-worn accelerometers	10
1.3 Wrist-worn accelerometers and Parkinson's disease . .	12
1.3.1 Population health studies	13
1.4 Challenges	16
1.5 Thesis aims and objectives	17
1.5.1 Thesis research questions	17
1.5.2 Structure of the thesis	18

1.1 Parkinson's disease

Parkinson's disease (PD) is a progressive neurodegenerative disorder that impacts both motor and non-motor systems of those affected[1]. It is primarily caused by the degradation of a specific region of the brain, in which dopamine-producing neurons begin to deteriorate over time[2]. This gradual neuronal loss leads to the characteristic motor symptoms of PD, such as tremors and muscle rigidity, as well as a host of non-motor symptoms including cognitive impairment and mood disorders. Current treatments for PD primarily aim to alleviate symptoms,

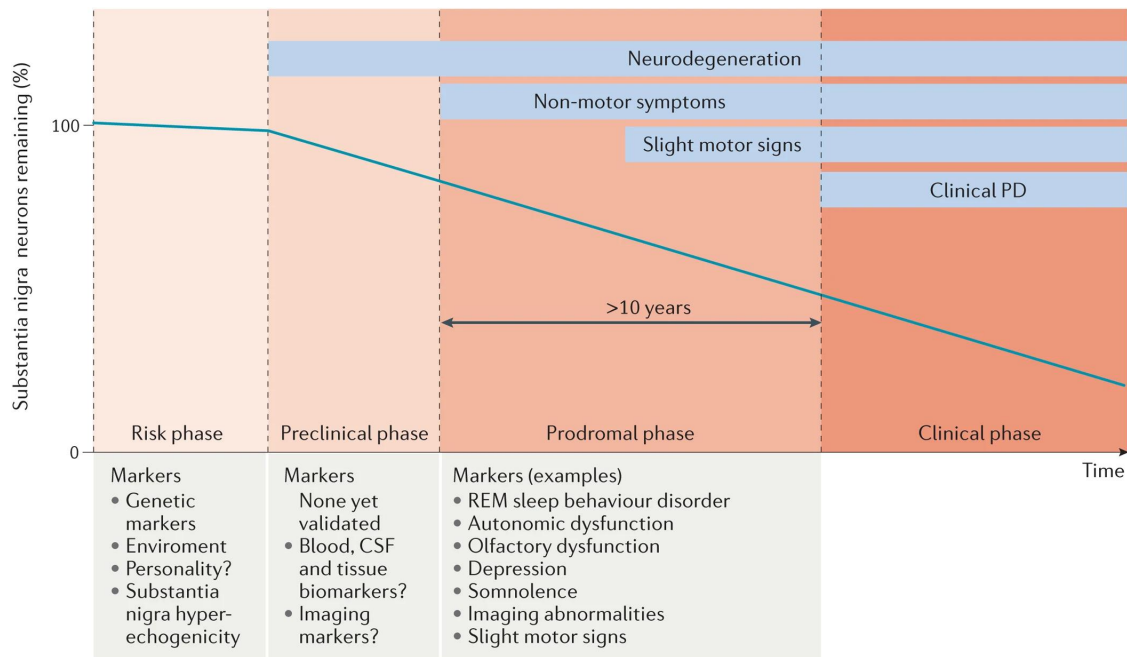
rather than halt or reverse the progression of the underlying neurodegenerative processes, causing their effectiveness to diminish over time. Furthermore, there is no known cure for the disease[1, 3].

The underlying pathology of PD is driven by the gradual degeneration of dopaminergic neurons residing within the substantia nigra pars compacta (SNpc) region of the brain[2]. These neurons are pivotal for orchestrating smooth and coordinated movements by releasing dopamine, a neurotransmitter that modulates various neural circuits involved in motor control. Specifically, dopamine exerts its influence by binding to specific receptors in target regions of the brain, notably the striatum, where it fine-tunes the balance between excitatory and inhibitory signals, ensuring precise motor initiation, execution, and coordination. The loss of dopaminergic neurons disrupts this delicate balance, leading to the hallmark motor symptoms observed in PD[1, 2, 3].

PD is the second most common, and fastest-growing neurodegenerative disease worldwide[4, 5]. The global burden of the disease was estimated at 9.4 million (95% CI: 7.7M, 11.2M) in 2020, marking a significant rise from the 6 million cases reported just four years earlier, in 2016[6]. This burden extends beyond those directly diagnosed with PD, affecting the lives of carers and family members as well. In the UK, the lifetime risk of developing PD is relatively low, estimated at just 2.7%[7]. However, it is projected that the prevalence of PD in the UK will rise by 23.2% by 2025[8]. Researchers have thus warned of an impending "Parkinson's pandemic"[5], driven by an ageing population and the absence of a cure. This underscores the urgent need for research efforts in PD, to better understand, prevent and treat the disease.

As dopaminergic neurons decline, Parkinson's disease progresses through distinct phases over many years. This transition is displayed in Figure 1.1 taken from Postuma et al. (2016)[9]. Of note are 4 phases of PD; risk phase, preclinical phase, prodromal phase and clinical phase.

The risk phase of PD is the state of the majority of the global population, which is before any amount of PD-specific neurodegeneration. During this phase,



Nature Reviews | Neurology

Figure 1.1: Illustration of neurodegeneration during phases of Parkinson's disease taken from Postuma et al. (2016)[9].

various markers/behaviours can predispose individuals to have a higher risk of a PD diagnosis. These risk factors are outlined in Section 1.1.1.

Individuals transition into the PD preclinical phase at the start of neurodegeneration[9, 10]. This is a difficult phase to identify, as individuals in this phase experience no noticeable change to their motor/non-motor function. It is currently hypothesised that advanced imaging scans or blood biomarkers may be able to identify individuals in the preclinical phase. However, it may be difficult to carry these out on large populations, due to the extensive testing required. It would therefore be most useful to selectively recruit the highest-risk populations for this testing. In addition, early research into specific blood/imaging biomarkers for the detection of this phase have shown mixed results[9, 11].

Next, the prodromal phase features ongoing neurodegeneration alongside the development of first non-motor, then motor symptoms of PD. The prodromal window for PD can last over a decade, with the emergence of symptomatic prodromal PD behaviour, that does not escalate to the severity of clinical PD. Some examples of

the non-motor symptoms that may appear are a loss of smell (anosmia), depression, and REM sleep behaviour disorder (RBD)[9, 12]. Following this, slight changes to motor behaviour also occur, including reduced arm swing, stooped posture, and diminished dexterity[9, 13].

The clinical phase of PD begins with the emergence of the more commonly known PD symptoms, leading to a diagnosis. This occurs after 20-50% of dopaminergic substantia nigra neurons have already been lost[9]. Figure 1.2 illustrates a range of symptoms typical of clinical PD, divided into motor and non-motor symptoms. Some motor symptoms of clinical PD include slowness of walking, resting tremor and rigid muscles. Examples of non-motor PD symptoms include those during the prodromal window, but also include constipation and issues with speech and vision, among others. Not all of these symptoms are present in each PD case, and symptoms can fluctuate while generally worsening over time. The Unified Parkinson's Disease Rating Scale (UPDRS)[14] attempts to measure the severity of both motor and non-motor symptoms of PD, allowing for the monitoring of the progression of the disease[9, 14, 15].

In the absence of a cure, current therapeutic modalities for PD predominantly focus on reducing symptomatic manifestations, with levodopa and deep brain stimulation emerging as common interventions. Levodopa augments depleted dopamine levels, providing a measure of reprieve from the progressive nature of the disease. However, the use of levodopa is accompanied by a spectrum of adverse effects, including gastrointestinal disturbances, neuropsychiatric complications, and the emergence of jolting in movements (dyskinesia), posing challenges to its long-term efficacy[16]. In contrast, deep brain stimulation represents a surgical modality involving precise electrical modulation, selectively targeting dysfunctional neural circuits[17]. While demonstrating sustained alleviation of motor symptoms over extended periods, deep brain stimulation is not devoid of risks, with potential surgical complications overshadowing its therapeutic potential[11, 17, 18].

For these reasons, there is great research interest in identifying Parkinson's disease risk markers, for the early identification of PD neurodegeneration[9, 18].

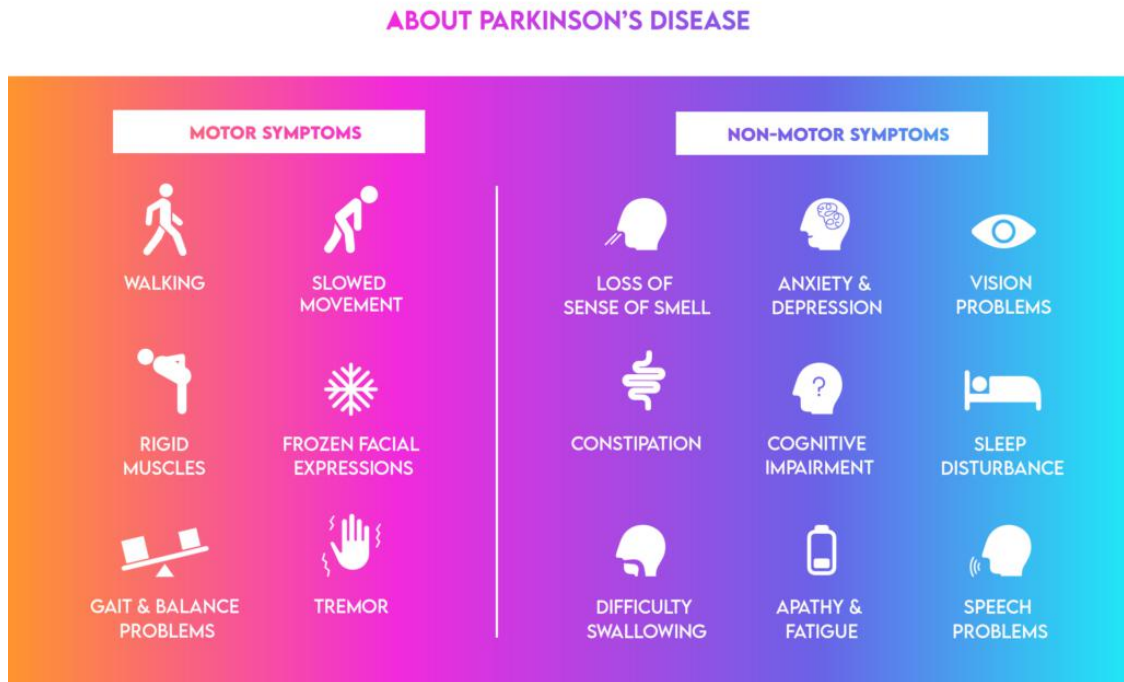


Figure 1.2: Illustration of Parkinson's disease symptoms. The image is courtesy of the Australian Parkinson's Genetics Study, developed by Health Haus in 2022, and commissioned by Associate Professor Miguel E. Rentería at the QIMR Berghofer Medical Research Institute, Brisbane, Australia[15].

There are many positive benefits to this. Firstly, this would allow for earlier intervention in these individuals before substantial neurodegeneration. It would also allow for the selective recruitment of at-risk populations into critical PD studies, aiding in the better understanding of the disease and the search for a cure. Furthermore, the identification of modifiable risk factors for PD could allow the possibility of public health messaging to lessen the emerging PD burden[2, 18]. To realise these benefits, better PD risk prediction algorithms must first be developed, using more predictive measures of PD.

1.1.1 Predictive measures of Parkinson's disease

It can be somewhat unclear to distinguish between two main types of predictive measure: risk factors and early markers. PD risk factors refer to exposures measured before any Parkinsonian neurodegeneration, that might be causally associated with an increased risk of PD. Early markers, on the other hand, refer to measured

exposures during the preclinical or prodromal phase of PD that are caused by the neurodegeneration. Note that other measures that are associated with future PD diagnosis but lack a known causal mechanism are referred to in this thesis generically as predictive measures. In the context of this thesis, PD prediction refers to the use of predictive measures to identify individuals with a high risk of PD.

Old age is the biggest known risk factor for idiopathic Parkinson's disease, as evidenced by the significant increase in its prevalence with advancing age[1, 19]. In the UK, between the ages of 45 and 69, the incidence of PD almost doubles for every 5-year increase in age. Over 1% of the population over the age of 60 and 5% of individuals over the age of 85 are affected by PD[7]. This increased risk is primarily due to the age-related decline in the health and function of dopaminergic neurons in the substantia nigra. These neurons are more susceptible to damage from slight changes in protein metabolism or mitochondrial function, leading to their catastrophic death and the onset of motor symptoms characteristic of Parkinson's disease[19].

Biological sex is also an important risk factor for incident Parkinson's disease[20, 21, 22]. Males are at a greater risk of PD diagnosis, as compared to females. The degree of this increased risk varies greatly in different studies and countries. Cerri et al. (2019)[20] reported that males have twice the risk, while Zirra et al. (2023)[22], a meta-analysis of studies around the world, reported a male-to-female prevalence ratio of 1.18 (95% CI: 1.03, 1.36). In 2015, this male-to-female prevalence ratio in the UK was 1.4[7]. PD and sex-based differences extend beyond incidence rates to include differences in disease progression, symptomatology, and response to treatments. Women with PD generally have a higher age at onset, but experience faster disease progression and higher mortality rates compared to men. Environmental factors such as occupational pesticide exposure, which historically has been higher among males, may contribute to the higher prevalence of PD in men. Additionally, healthcare access disparities and societal factors may further influence the observed gender differences in PD prevalence. Genetic factors also

contribute to these differences, with certain genetic mutations affecting the onset and progression of PD differently in males and females[20, 21, 23].

Some individuals can also have a genetic predisposition to a higher risk of PD. In the past, genetic risk factors were estimated by having a family history of Parkinson's disease, with an estimated 15-25% of people with PD having a familial link[12]. However, it is relatively rare for PD to be caused by a single genetic mutation, which accounts for about 30% of familial cases and only 3-5% of sporadic cases[12]. Genome-wide association studies (GWAS) have revolutionised modern genetic risk analysis by identifying common genetic variants associated with a higher risk of PD. These studies identified numerous genetic loci independently associated with PD risk[12]. Researchers use this data to create polygenic risk scores, estimating an individual's genetic risk for developing Parkinson's disease. For example, Jacobs et al. (2020)[24] reported an odds ratio of 3.37 (95% CI: 2.41, 4.70) for individuals in the highest decile of polygenic risk compared to those in the lowest decile.

Cigarette smoking appears to be a preventative risk factor for PD. Multiple epidemiological studies have consistently shown a strong inverse association between smoking and the risk of developing PD, with smokers having a substantially lower relative risk compared to non-smokers. This association appears to be dose-dependent, with increased smoking duration and intensity correlating with even lower PD risk. While the exact mechanisms underlying this association are not fully understood, several hypotheses have been proposed, including the neuroprotective effects of nicotine and other components of tobacco smoke, such as carbon monoxide (CO) and monoamine oxidase-B (MAO-B) inhibitors[25, 26].

High caffeine intake similarly appears to be a preventative risk factor for PD. Caffeine is a psychoactive substance that is widely consumed, found in coffee, tea and other beverages. Several large prospective studies have consistently shown an inverse relationship between coffee (and therefore caffeine) consumption and the risk of developing PD. Moreover, experimental studies using animal models of PD, induced by mitochondrial toxins and α -synuclein expression, have provided further

evidence supporting the neuroprotective effects of caffeine against dopaminergic neurodegeneration[27].

Lower physical activity levels are a known predictive measure of PD; however, it remains unclear whether it is a risk factor or an early marker[28, 29, 30, 31]. Individuals with PD are noted to have lower physical activity levels several years before a diagnosis, during the prodromal window. Van Nimwegen et al. (2011)[28] reported a difference of 29% in comparing clinical PD patients to controls, by minutes of daily physical activity. This may be due to a variety of different factors, including overall disability, poorer walking performance, fear of falls/anxiety/depression, among others. However, several studies propose exercise as a preventative risk factor for PD. This may be explained by Cohen et al. (2003)[30] and Neeper et al. (1996)[31] who indicate through animal experimentation that increased exercise can lead to the secretion of neurotrophic factors encouraging neuroplasticity and survival of dopamine neurons. The measurement of activity levels has historically suffered from self-recall bias, however, modern accelerometers offer the ability to objectively measure activity levels, explored further in Section 1.2.

Changes in sleep behaviour are a known early marker of PD, during the prodromal phase. There are many sleep-related prodromal PD markers, including REM sleep behaviour disorder, poorer overnight sleep efficiency and duration, and higher rates of daytime sleepiness and naps[9, 32].

There are several other potential risk measures for PD, that have no established causal link, however, are used to varying degrees throughout this thesis. Firstly, Body Mass Index (BMI) was initially reported from a meta-analysis to have a significant association with incident PD[33]. However, after adjustment of confounding factors, this association was no longer significant[34]. Alcohol has similarly been found to have a weak association with incident PD, however, Bettiol et al. (2015)[35] proposed that this may be due to selection and recall biases. Ethnicity and indicators of socio-economic deprivation are demographic factors often used as risk measures. While these factors do not have clear independent causal links to

Parkinson’s disease, they remain useful covariates for adjustment with association analyses within this thesis[36, 37].

These current predictive measures, however, fail to fully capture the range of motor changes occurring during the prodromal phase of PD. Numerous early indicators manifest during routine daily activities, offering enhanced prospects for identifying individuals at pre-diagnostic stages. Wrist-worn accelerometers facilitate the capture and quantification of these free-living motor behaviours, for the improvement of PD risk algorithms.

1.2 Wearable devices

Wearable devices are technology worn or attached to the body, equipped with sensors for capturing information on an individual’s biometric or physiological data. These devices can track a wide range of measures, including heart rate, steps taken, sleep quality, and more. By seamlessly integrating into daily life, wearable devices provide continuous, real-time data collection, making them invaluable tools for both personal health management and scientific research[38, 39, 40, 41, 42].

Wearable devices continue to grow in mainstream popularity. The global wearable medical devices market is projected to grow from \$73.77 billion in 2023 to \$428.92 billion by 2030, at a compound annual growth rate of 28.6%. Studies report that in 2023, around 35% of US adults used wearable healthcare devices. This rapid growth underscores the increasing reliance on wearable technologies in both consumer health and clinical research[43].

These devices have seen increased usage within research studies over the past decade. Huhn et al. (2022)[41] performed a scoping review of affordable consumer-grade wearables used in research studies. The findings of this review revealed that between 2013 and 2020, the number of studies and study participants continually increased. Research-grade devices, however, are wearables that are specifically intended for research use. These devices allow for the extraction of raw sensor data, useful for bespoke extraction of useful measures.



Figure 1.3: Illustration of the use of wrist-worn accelerometer devices, to capture information on sleep, gait, gestures and activity.

The benefits of using wearable devices in research are numerous. First, wearable devices are generally more affordable compared to traditional medical equipment, making them accessible for a wide range of studies and participants. Second, wearables allow for continuous monitoring of individuals in their natural environments, providing more accurate data compared to in-clinic assessments. Third, these devices can provide objective measurements of various physiological and behavioural parameters, reducing the bias often associated with self-reported data.

1.2.1 Wrist-worn accelerometers

Wrist-worn accelerometers are a specific type of wearable device that record acceleration from the wrist of the wearer. Fitness trackers, which are often wrist-worn accelerometers with extra sensors, and wrist-worn accelerometers are reported by Huhn et al. (2022)[41] as the most popular types of consumer-grade wearable device used in research. Their use in research studies, compared to all other wearable device locations, has steadily grown in popularity in recent years. In addition, the majority of research studies using research-grade wearables use wrist-worn accelerometers[44].

Many modern wrist-worn devices come equipped with multiple sensors, not just limited to accelerometers. This can include photoplethysmography (PPG)

sensors, gyroscopes, magnetometers, light or temperature sensors. However, across different studies, accelerometers remain the most common sensor in wrist-worn wearable devices. This allows for the use of more biomedical datasets within this thesis. In addition, the use of all these sensors has a detrimental effect on the battery life of the device. Wrist-worn accelerometer devices however can last over 7 days of raw data capture on a full battery.

Raw acceleration data extracted from these sensors are typically in 1 to 3 orthogonal planes known as axes. This raw data is typically sampled at over 30Hz, to ensure meeting the Nyquist criterion of 15Hz for which 99% of the energy is contained below [45]. This is particularly necessary for studying PD, for which resting tremors can exist at high frequencies, but remain below this cap [46]. This raw data is typically measured in either meters per second squared, or units of gravity ($g=9.81\text{ms}^{-2}$). The acceleration data is also calibrated, to ensure that at rest, the device outputs a magnitude of acceleration equal to Earth's gravity[47].

Data captured from wrist-worn accelerometers can offer a variety of useful indicators/measures, as illustrated in Figure 1.3. To do so, the raw signal data must be processed into measures that have understandable meaning. Various research studies have made attempts at extracting this information from the raw signal data. For example, Van Hees et al. (2013)[48] used signal processing calculations called the Euclidean norm minus one (ENMO), to measure overall average acceleration during the period of monitoring. This calculation was found to be well correlated with physical activity energy expenditure (PAEE) of the individual during this time[48].

More complex measures of activity, giving a more descriptive profile of behaviour, can also be extracted. However, this relies on the development of machine learning models capable of classifying unseen accelerometer data. This was completed by Willetts et al. (2018)[49] and Walmsley et al. (2021)[50], the latter of which looks to separate participant behaviour into labels of sleep, sedentary, light and moderate-vigorous physical activity (MVPA) using a balanced random forest model. Such approaches to the processing of accelerometer data have grown more popular with increased capture of accelerometer data, and of machine learning, both

shallow and deep learning models, recognise human activity from accelerometer data. This offers the possibility of more explainable metrics of behaviour, specific to the disease of interest.

1.3 Wrist-worn accelerometers and Parkinson's disease

Several studies have investigated using wrist-worn accelerometer devices in PD research. These devices can provide a range of purposes for the disease. This includes the long-term monitoring of the progression of clinical PD symptoms, the detection of PD, the observation of associations between PD and device measures, and risk prediction for PD using accelerometer data.

Many PD studies use wrist-worn accelerometer data to detect or monitor clinical PD. Typically, these studies selectively recruit individuals with a PD diagnosis and sometimes include a healthy control population for comparison. Participants are equipped with wrist-worn accelerometers for a period, allowing for continuous monitoring of their behaviour. Various metrics are then extracted from the accelerometer data to identify patterns of movement that are specific to the PD population. These studies demonstrate the potential of wearable technology in providing critical insights into the motor symptoms associated with PD.

Sotriakis et al. (2023)[51] aimed to determine if wearable sensor data combined with machine learning could estimate clinical rating scales and monitor motor symptom progression in PD patients. The study involved seventy-four patients who were monitored over 15 months, with seven visits at 3-month intervals. Their walking and postural sway were recorded using six Inertial Measurement Unit (IMU) sensors, including 2 wrist-worn accelerometers. The researchers employed various machine learning algorithms, including random forests[52], to estimate the Movement Disorder Society - Unified Parkinson's Disease Rating Scale part III (MDS-UPDRS-III)[14]. The random forest model provided the most accurate estimation among the seven models used. Significant motor symptom progression was detected within 15 months, which the conventional clinical rating scale did not

capture. This study demonstrated that wearable sensors and machine learning could track motor symptom progression in PD more effectively than traditional clinical rating scales, highlighting the potential for these methods to improve diagnostic and prognostic accuracy. It should be noted, however, that this full suite of six wearable devices used to monitor participants is not feasible for a larger population[51].

Similarly, Atri et al. (2022)[53] explored the feasibility of using wearable sensors to detect PD in an unconstrained, real-world setting. The study used inertial sensor data from Verily Study Watches worn by seven PD subjects and four controls for up to 23 hours per day over several months. Human activity recognition techniques identified walk-like activities within the unconstrained, unlabelled data, and one-dimensional convolutional neural networks (1D-CNNs)[54] were trained to detect PD. The study achieved nearly 90% classification accuracy on single 5-second walk-like events and 100% accuracy with the majority vote over a day's events. While results are promising, this was a proof-of-concept study on just 11 participants, hence, these approaches should be attempted on a larger scale[53].

Small clinical studies, however, have clear limitations concerning relevance to this thesis' aims. Firstly, the selective recruitment of PD populations introduces complications when applying these findings to the general population. For instance, many observed patterns in the more severe PD population, as noted by Sokratis et al. (2023)[51], such as reduced arm swing and slower walking speed, can also be attributed to other diseases or simply to ageing. Furthermore, such studies primarily focus on individuals already diagnosed with clinical PD. Therefore, these study designs are not well positioned to investigate risk measures during the early stages of PD, particularly the preclinical or prodromal stages. During these phases, the profile of accelerometer-derived measures associated with the disease may be substantially different.

1.3.1 Population health studies

Population health studies are large-scale observation studies of the general population, tracking health outcomes over time. In these studies, it is anticipated

that a proportion of healthy participants at the time of first observation will eventually receive a disease diagnosis. In the context of PD, large observational studies with long follow-up times are required, to observe a sufficient number of PD cases, due to the relative rarity of PD. Note that all existing population health studies on Parkinson's disease and wrist-worn accelerometers use the UK Biobank Physical Activity Monitoring study[55]. This is because the study monitored over 100,000 participants in 2013-2015, meeting both conditions for sufficient expected PD cases[55].

Williamson et al. (2021)[56] analysed the wrist-worn accelerometry from participants with a PD diagnosis against age-matched healthy controls. They extracted measures of movement dispersion from gait data and movement dimensionality from low-movement data to apply machine learning classifiers to detect PD. The study achieved an area under the receiver operator characteristic curve (AUROC)[57] of 0.69 for gait data, 0.84 for low-movement data, and 0.85 when combining both activities. The results suggested that wrist-worn accelerometers could be a viable, low-cost sensor for detecting PD-related movement abnormalities. However, the study did not distinguish between the incident and prevalent PD cases, creating uncertainty around whether the observed markers are that of clinical or prodromal PD. In addition, the reliance on an age-matched control group, may not fully represent the general population, limiting the usefulness of the results. Finally, these results were not reproducible, with a reported incidence of walking falling far below expected levels in the general population[56].

Schalkamp et al. (2023)[58] leveraged the UK Biobank data to explore the predictive value of accelerometry in identifying prodromal PD. They compared accelerometry data with genetic, lifestyle, blood biochemistry, and prodromal symptoms data, finding that accelerometry provided better performance in distinguishing both clinically diagnosed PD and prodromal PD up to seven years pre-diagnosis. Accelerometry achieved an area under the precision-recall curve (AUPRC) of 0.14 for clinically diagnosed PD and 0.07 for prodromal PD, outperforming other modalities tested. However, the study's limitations include the relatively low AUPRC values,

indicating that while accelerometry is promising, it is not yet highly reliable on its own. Additionally, the study did not account for potential confounding factors such as comorbid conditions or medication use, which could affect accelerometer readings. Furthermore, this study used a metric called vector magnitude to summarise activity levels, however, this is not an interpretable measure of activity[58].

Liu et al. (2023)[59] examined the association between physical activity and PD risk using accelerometer data from the UK Biobank. They analysed data from 96,422 participants, observing that higher levels of light physical activity and moderate-to-vigorous physical activity were inversely associated with PD risk, while higher sedentary time was positively associated. After adjusting for potential confounders, comparing to those with both low light activity (<3.89 h/day) and low moderate-vigorous activity (<0.27 h/day), the adjusted hazard ratio (95% CI) of PD risk was 0.49 (0.36–0.66), 0.19 (0.36–0.66) and 0.13 (0.09–0.18), respectively, for participants with high moderate-vigorous activity only, high light activity only, and both high light and high moderate-vigorous activity. This study, however, makes no mention of reverse causation and its impact on the observed association[59].

Reverse causation is a type of bias that occurs when investigating the relationship between an exposure and an outcome, where the outcome actually causes the exposure rather than the other way around. This can create a misleading appearance that the exposure leads to the outcome when this is not the case. A common example of reverse causation is the delayed identification of incident cases. For instance, individuals diagnosed immediately after a study begins might have been prevalent cases before the study's start. To mitigate this bias, researchers often exclude data from the first few years of follow-up (both deaths and diagnoses) on the premise that these individuals were likely already in a compromised state at the study's commencement. This approach is especially crucial in studies of incident PD, due to the disease's long prodromal period. During this period, early markers of PD might affect associations with incident PD, leading to potential misinterpretation of risk factors from the preclinical phase.

Zhao et al. (2023)[60] compared the prediction performance of accelerometry-derived physical activity measures with traditional risk factors for incident PD in the UK Biobank. They analysed data from 92,352 participants, finding that accelerometry measures, particularly total acceleration, outperformed traditional predictors such as age and sex in predicting PD. This, however, used single predictor models and hence did not account for other lifestyle factors that might impact the risk of PD, such as age, sex and smoking[60].

Lin et al. (2024)[61] investigated the association between different physical activity patterns and Parkinson’s disease risk using data from the UK Biobank. They analysed data from 89,400 participants, categorising them into inactive, regularly active, and weekend warrior groups. Both active groups had a lower risk of developing PD compared to the inactive group, with no significant difference between the active groups. The study suggested that regular and weekend warrior exercise patterns are equally effective in reducing PD risk[61].

1.4 Challenges

Despite much promise, several challenges persist in translating movement and activity-related phenotypes into meaningful health insights, particularly in the context of PD.

Firstly, many studies rely on self-reported data to measure activity levels and health outcomes, which can introduce significant bias and inaccuracies. Self-reported data may not accurately reflect true activity levels, especially in individuals with PD, who may have impaired motor function and cognitive challenges that affect their ability to accurately report their activities. This thesis will evaluate the utility of device-measured activity data, which can provide a more objective and reliable measure of physical activity and movement patterns, for predicting and monitoring PD.

Furthermore, there are difficulties in extracting accurate, reliable and meaningful behaviour measures from wrist-worn accelerometers using machine learning. The device’s location at the wrist offers the possibility of a wide range of motion that

does not directly correspond to physical exercise. This is particularly noted in the difficulty of the classification of light physical activity[44]. There are also difficulties in ensuring the generalisability of models used to classify activities. This is particularly important for the PD population, for whom changes in physical behaviour are expected.

Another challenge is addressing reverse causation bias when investigating associations with various exposures to incident PD. This is essential for distinguishing between risk factors and early markers of Parkinson’s disease. Many existing studies have failed to appropriately account for reverse causation, and make little attempt to distinguish between markers of clinical PD, markers of prodromal PD and risk factors for incident PD.

Finally, there are challenges in identifying at-risk populations of PD during the early stages of neurodegeneration. While many traditional risk factors for PD have been identified, the effect size for these factors is often not very large. Predictive algorithms for PD remain rather unused, with key markers for prodromal PD such as REM sleep behaviour disorder used for the selective early recruitment into PD-related studies[62], rather than predictive algorithms.

By addressing these challenges, this research aims to enhance the understanding and predictive capabilities of wrist-worn accelerometers in the context of PD.

1.5 Thesis aims and objectives

1.5.1 Thesis research questions

In light of the challenges above, this thesis aims to address the following research questions;

- Can improved activity recognition models be developed, generalisable to both the healthy and PD populations?
- Are reduced activity levels a risk factor or early marker of incident PD?
- Can wrist-worn accelerometry distinguish between the free-living gait of healthy controls, clinical PD and prodromal PD populations?

- Can measures extracted from wrist-worn accelerometers during 7 days of free living improve PD risk prediction models?

1.5.2 Structure of the thesis

This thesis is organised into 7 chapters. Chapter 2 lists the wrist-worn accelerometer datasets used throughout this thesis and some key properties of these datasets. Chapter 3 outlines the development and improvement of activity recognition models deployed on wrist-worn accelerometer data. In Chapter 4, I investigate the epidemiological association between median daily step count and incident PD. Chapter 5 explores the classification of PD status using detected active walking windows of wrist-worn accelerometer data. Chapter 6 conducts a risk analysis, investigating how the use of accelerometer-derived features can help improve PD risk models. Finally, Chapter 7 reflects on the work completed throughout this thesis and outlines some potential future work.

2

Datasets

Contents

2.1	Introduction	19
2.2	UK Biobank Physical Activity Monitoring Study . . .	22
2.3	Parkinson’s Progression Markers Initiative Verily study	28
2.4	Capture-24	31
2.5	OxWalk	32
2.6	Michael J. Fox Foundation Levodopa Response study	34
2.7	Data generalisability	35
2.8	Dataset preparation	37

2.1 Introduction

This chapter outlines the datasets used throughout this work, investigating the use of wrist-worn accelerometry to predict Parkinson’s disease (PD). At the start of this research project, I searched for existing wrist-worn accelerometer datasets that would meet the research aims of this thesis. Primarily, these datasets needed to be:

- Easily accessible, and open to researchers.
- Capture a range of activities in daily living.
- Be relevant to both healthy and PD populations.

The datasets chosen typically fell into one of two main purposes, either for model inference or model training.

Inference datasets are large-scale datasets, capturing data from many individuals, used to observe patterns in behaviour between populations. Due to this large scale, more limited data is collected from participants to encourage higher recruitment rates. In the case of large-scale accelerometer datasets, participants are typically provided with just the accelerometer device. This means that while accelerometer data is captured from many individuals, the activities performed by participants during this period of data capture are uncertain. This reduces the participation burden, allowing for a wider variety of participants, which is essential for recruiting individuals with motor diseases like PD, who are typically under-recruited in research studies[63]. Due to the large amount of data collected, there is greater statistical power, offering greater confidence in observed patterns, allowing for large-scale, population-based epidemiological studies, and risk analyses.

Training datasets, on the other hand, are smaller-scale datasets used to train supervised machine learning models. To do so, the raw data must contain labels as ground truth. There are various approaches to producing ground truth labels, such as using instructed tasks or extra monitoring devices. The training of human activity recognition (HAR) models is a primary use of small-scale accelerometer datasets. HAR models are used to predict individuals' activity using only accelerometer data. Trained HAR models can therefore be deployed on inference datasets and offer insights into participant behaviour despite a lack of ground truth capture.

During this thesis, I elected to use five distinct datasets. These datasets are outlined in the following subsections and summarised in Table 2.1.

Characteristic	UK Biobank Physical Activity Monitoring Study	PPMI-Verily dataset	Capture-24	OxWalk study	MJFF Levodopa Response Study
Number of participants	103,660	356	151	39	28
Sensor	Axivity AX3	Verily watch	Axivity AX3	Axivity AX3	GENEActiv
Raw sampling rate	100Hz	80Hz	100Hz	100Hz	50Hz
Body location	Dominant wrist	Wrist, dominance not specified	Dominant wrist	Dominant wrist	Wrist of most affected limb
Activity protocol	Free-living	Free-living	Free-living	Free-living	Scripted tasks
Measurement period	7 days	Up to 2 years	Up to 24 hours	1 hour	2 hours
Ground truth capture	None	None	Chest worn, front facing, camera logger	Waist worn, foot facing, video camera	Task diaries/ inspection
Mean Participant Age	62.4 (SD 7.8) Range: 43 - 79	65.6 (SD 8.1) Range: 39 - 88	42 Range: 19 - 81	38.5 (SD 14.0) Range: 19 - 81	62.5 (SD 8.8) Range: 46 - 80
Participant Sex	58,267 Female, 45,393 Male	181 Female, 175 Male	99 Female, 52 Male	19 Female, 20 Male	9 Female, 19 Male
Participant PD status	>100,000 healthy, 376 Diagnosed PD*, 565 Prodromal PD*	161 Diagnosed PD, 158 Prodromal PD, 35 Healthy controls	151 Healthy	39 Healthy	28 Diagnosed PD

Table 2.1: Comparison of different datasets used within this thesis. *Estimates of predicted case numbers.

2.2 UK Biobank Physical Activity Monitoring Study

The UK Biobank is a large-scale biomedical database launched in 2006, aiming to improve the prevention, diagnosis, and treatment of many diseases. This dataset comprises information collected from over 500,000 participants aged between 40 and 69 across Great Britain (England, Scotland, and Wales). Baseline data collection occurred from 2006 to 2010, using a series of health questionnaires, physical measurements, and biological samples[64].

Health information for this population is regularly collected through linkages to primary care, hospital inpatient and death registry data. Primary care data consists of various notes captured from participants' visits to General Practitioners (GPs). Hospital inpatient data contains records of diagnoses codes (ICD-10)[65], for any hospital admissions. Death registry data contains notifications of the death of participants, as well as cause(s) of death identified using diagnoses codes (ICD-10)[65]. UK Biobank regularly collects and updates this health outcome data for research use[66]. The latest update of this data provides health outcomes from hospital episode records up to 2022, and death data up to 2023, for all participants[67]. Primary care data however has data up to 2016/17, for roughly 45% of participants[67, 68]. These censoring dates are dependent on the data provider, summarised in Table 2.2.

Numerous studies on PD have leveraged the UK Biobank dataset[24, 58, 59, 60, 61, 69, 70]. These investigations commonly adhere to the methodology outlined by Bush et al. (2018)[71], which combines self-reported diagnoses, hospital inpatient records, and death certificate data to identify both prevalent and incident PD cases. This approach has demonstrated robust performance, achieving a positive predictive value of 0.91[71]. A limitation of this strategy is its exclusion of primary care data. As highlighted by Clifton et al. (2024)[70], incorporating primary care data can potentially double the number of observed cases. This exclusion is however unavoidable, as primary care data is only available up to 2016/17 and for less than half of UK Biobank participants[67, 68].

Country	Data Provider	Participant coverage (approx.)	Censoring date
Primary care data			
England	TPP	165,000	31-May-16
England	Vision	18,000	31-May-17
Scotland	Vision/EMIS	27,000	31-Mar-17
Wales	Vision/EMIS	21,000	31-Aug-17
Hospital inpatient data			
England	NHS England	445,000	31-Oct-22
Scotland	Information and Statistics Division (ISD), Scotland	34,000	31-Aug-22
Wales	Secure Anonymised Information Linkage (SAIL), Wales	21,000	31-May-22
Death Data			
England	NHS England	445,000	30-Nov-23
Scotland	NHS Central Register, National Records of Scotland	34,000	31-Dec-23
Wales	NHS England	21,000	30-Nov-23

Table 2.2: UK Biobank censoring dates across different data providers and countries[66, 67, 68]

The UK Biobank conducts sub-studies, to collect further information from subsets of participants that were not collected at baseline. The UK Biobank Physical Activity Monitoring study is one such sub-study, conducted between 2013 and 2015. Over 200,000 participants were initially approached to participate in this study, with roughly 100,000 participants agreeing to participate[55]. This population is also referred to in this thesis as the UK Biobank Accelerometer cohort. The general demographics of these populations are outlined in Table 2.3.

Within this study, the activity levels of participants were monitored using an AX3 Axivity device, worn on their dominant wrist, for 7 days, while going about their regular activities. This dataset remains the largest available collection of accelerometer data (in person-years) to date. The AX3 Axivity device collected tri-axial accelerometer data non-stop for 7 days at 100Hz sampling frequency and a

dynamic range of $\pm 8g$. There was no ground truth capture of activities performed by participants, nor instructions to demand participant activities.

Due to the large population in the accelerometer cohort, and the considerable period of follow-up of health data linkage, I expected an appreciable number of PD cases. This is despite the relatively low prevalence and incidence of the condition. Given the age and sex distribution of these participants, the expected number of prevalent cases in this cohort was estimated to be 376 (95% CI: 340-415) and incident cases using all follow-up was 565 (95% CI: 503-634). A breakdown of this calculation is attached in Tables 2.4 and 2.5. It should be noted that these numbers may be slightly inflated, due to healthy volunteer bias, as those with PD are less likely to participate in the study.

The UK Biobank physical activity monitoring study therefore serves as an inference dataset for this research project. The large population allows the opportunity to carry out epidemiological association analyses observing patterns in activity and Parkinson's disease. In addition, I can develop risk models deployed on this population. I do however rely on the development of supervised classification models to be deployed on this data, to extract greater insights on participant behaviour during the period of monitoring.

Characteristic	Female (N=58,267)	Male (N=45,393)	Overall (N=103,660)
Age (years)			
40-49	5,164 (8.9%)	3,698 (8.1%)	8,862 (8.5%)
50-59	18,118 (31.1%)	11,792 (26.0%)	29,910 (28.9%)
60-69	25,720 (44.1%)	20,244 (44.6%)	45,964 (44.3%)
70-79	9,265 (15.9%)	9,659 (21.3%)	18,924 (18.3%)
Ethnicity			
Non-white	1,924 (3.3%)	1,391 (3.1%)	3,315 (3.2%)
White	56,167 (96.4%)	43,811 (96.5%)	99,978 (96.4%)
Missing	176 (0.3%)	191 (0.4%)	367 (0.4%)
TDI Quintile*[72]			
Least Deprived (< -2.928)	23,594 (40.5%)	19,246 (42.4%)	42,840 (41.3%)
Second Least Deprived	12,394 (21.3%)	9,651 (21.3%)	22,045 (21.3%)
Third Least Deprived	8,826 (15.1%)	6,647 (14.6%)	15,473 (14.9%)
Second Most Deprived	7,445 (12.8%)	5,323 (11.7%)	12,768 (12.3%)
Most Deprived (≥ 2.516)	5,940 (10.2%)	4,471 (9.8%)	10,411 (10.0%)
Missing	68 (0.1%)	55 (0.1%)	123 (0.1%)
Recruitment Centre			
England	52,335 (89.8%)	40,735 (89.7%)	93,070 (89.8%)
Scotland	3,803 (6.5%)	2,914 (6.4%)	6,717 (6.5%)
Wales	2,129 (3.7%)	1,744 (3.8%)	3,873 (3.7%)
Education[73]			
College/University degree	24,441 (41.9%)	20,186 (44.5%)	44,627 (43.1%)
Vocational qualification	14,753 (25.3%)	13,107 (28.9%)	27,860 (26.9%)
School degree	13,929 (23.9%)	7,736 (17.0%)	21,665 (20.9%)
No degree/qualification	4,565 (7.8%)	3,899 (8.6%)	8,464 (8.2%)
Missing	579 (1.0%)	465 (1.0%)	1,044 (1.0%)
Alcohol			
Never	3,835 (6.6%)	2,033 (4.5%)	5,868 (5.7%)
<3 times/week	29,445 (50.5%)	17,560 (38.7%)	47,005 (45.3%)
3+ times/week	24,935 (42.8%)	25,765 (56.8%)	50,700 (48.9%)
Missing	52 (0.1%)	35 (0.1%)	87 (0.1%)
Smoking			
Never	35,330 (60.6%)	23,621 (52.0%)	58,951 (56.9%)
Previous	19,283 (33.1%)	17,874 (39.4%)	37,157 (35.8%)
Current	3,505 (6.0%)	3,770 (8.3%)	7,275 (7.0%)
Missing	149 (0.3%)	128 (0.3%)	277 (0.3%)
BMI Category (kg/m²)			
<18.5	498 (0.9%)	85 (0.2%)	583 (0.6%)
18.5-24.9	26,646 (45.7%)	13,335 (29.4%)	39,981 (38.6%)
25.0-29.9	20,286 (34.8%)	22,281 (49.1%)	42,567 (41.1%)
30.0+	10,701 (18.4%)	9,585 (21.1%)	20,286 (19.6%)
Missing	136 (0.2%)	107 (0.2%)	243 (0.2%)

Table 2.3: Demographic and lifestyle characteristics of the UK Biobank Physical Activity Monitoring population. * TDI Quintiles according to UK population census in 2011[72]

Sex	Age Band	Accelerometer Cohort subpopulation Count	Estimated Prevalence Rate per 100,000 (95% CI)	Expected Accelerometer Cohort Prevalent Cases (95% CI)
All	40-49	8,862	14.2 (11.0-18.1)	1 (1-2)
Female	50-54	8,220	48.8 (34.9-66.5)	4 (3-5)
Female	55-59	9,898	95.8 (78.0-116.4)	9 (8-12)
Female	60-64	11,764	177.7 (151.6-207.1)	21 (18-24)
Female	65-69	13,956	318.2 (283.7-355.8)	44 (40-50)
Female	70-74	8,198	676.3 (618.8-737.6)	55 (51-60)
Female	75-79	1,067	925.7 (850.8-1,005.4)	10 (9-11)
Male	50-54	5,464	88.6 (69.7-111.0)	5 (4-6)
Male	55-59	6,328	159.7 (136.7-185.5)	10 (9-12)
Male	60-64	8,456	330.1 (293.8-369.7)	28 (25-31)
Male	65-69	11,788	654.1 (603.0-708.5)	77 (71-84)
Male	70-74	8,455	1,065.7 (990.0-1,145.8)	90 (84-97)
Male	75-79	1,204	1,703.3 (1,592.5-1,819.7)	21 (19-22)
All	All	103,660	-	376 (340-415)

Table 2.4: Estimation of prevalent PD cases in the UK Biobank Physical Activity Monitoring Cohort Data. Prevalence rates estimated using the Parkinson’s UK 2017 detailed report[8].

Sex	Age Band	Mean age (years)	Mean follow-up time (years)	UKBB PAM subpopulation count	Estimated population Incidence (95% CI)	UK Rate (95% CI)	Estimated PAM Diagnosis Rate (95% CI)	Expected Incident (95% CI)	Cases
Female	40-44	44.6	8.8	147	0.0 (0.0-0.0)		48.1 (35.2-65.2)	0 (0-0)	
Female	45-49	48.1	8.0	5,017	4.0 (3.1-5.2)		64.2 (46.6-87.3)	3 (2-4)	
Female	50-54	52.6	7.9	8,220	8.4 (5.9-11.7)		95.5 (72.4-125.0)	8 (6-10)	
Female	55-59	57.6	7.9	9,898	12.8 (9.8-16.6)		162.2 (131.2-199.1)	16 (13-20)	
Female	60-64	62.6	7.9	11,764	22.3 (18.1-27.2)		296.3 (252.5-345.7)	35 (30-41)	
Female	65-69	67.4	7.8	13,956	41.4 (35.5-48.0)		551.6 (485.0-624.2)	77 (68-87)	
Female	70-74	72.1	7.7	8,198	83.4 (73.7-93.9)		844.2 (754.7-940.9)	69 (62-77)	
Female	75-79	75.9	7.3	1,067	126.4 (113.5-140.3)		959.5 (859.5-1,067.8)	10 (9-11)	
Male	40-44	44.6	8.7	124	0.0 (0.0-0.0)		54.0 (41.4-70.0)	0 (0-0)	
Male	45-49	48.1	8.0	3,574	4.0 (3.1-5.2)		86.0 (67.1-109.2)	3 (2-4)	
Male	50-54	52.5	7.9	5,464	10.5 (8.0-13.6)		159.3 (129.6-193.7)	9 (7-11)	
Male	55-59	57.6	7.9	6,328	23.1 (18.9-27.9)		307.4 (263.5-356.5)	19 (17-23)	
Male	60-64	62.6	7.8	8,456	43.2 (37.2-49.9)		564.4 (502.3-632.1)	48 (42-53)	
Male	65-69	67.5	7.7	11,788	79.7 (71.3-88.8)		980.8 (888.5-1,080.1)	116 (105-127)	
Male	70-74	72.2	7.5	8,455	147.0 (133.5-161.5)		1,541.4 (1,412.1-1,679.2)	130 (119-142)	
Male	75-79	75.9	7.0	1,204	240.7 (221.2-261.4)		1,812.3 (1,659.5-1,975.0)	22 (20-24)	
All	40-79	62.3	7.8	103,660	-		-	565 (503-634)	

Table 2.5: Estimation of incident PD cases in the UK Biobank Physical Activity Monitoring (UKBB PAM) cohort data, given a mean age at baseline of 62.3 years, and 7.8 years of follow-up. Incidence rates estimated using the Parkinson's UK 2017 detailed report[8].

2.3 Parkinson's Progression Markers Initiative Verily study

The Parkinson's Progression Markers Initiative (PPMI) is a landmark research effort aimed at improving the understanding of PD, to accelerate the development of new therapies. Launched in 2010, PPMI is a global, multi-site, longitudinal study that involves collaboration between academic institutions, pharmaceutical companies, and patient advocacy groups. To date, over 4,000 participants have submitted clinical and imaging data to over 50 study sites across the world. A further set of participants have remotely or online provided more data, however, they were not used in this thesis. The PPMI study collects a variety of clinical, imaging, biologic and genetic markers to observe patterns in the progression of PD. The data collected within this study has been made open access to researchers[74, 75].

Different cohorts of participants were selectively recruited into the PPMI study to investigate the overall progression of PD. These cohorts are individuals with a clinical diagnosis for PD, healthy controls with no sign of neurological disorders, individuals with imaging scans for PD but no markers of dopaminergic deficit (SWEDD), and prodromal Parkinson's disease populations. In this study, prodromal PD was identified by the presence of specific non-motor symptoms, risk factors such as genetic mutations or family history, and biomarkers indicating increased risk. The early identification of this population allows for the monitoring of the overall progression of symptoms through the prodromal to the clinical phase of PD[62].

In 2018, a PPMI sub-study, known as PPMI-Verily, was launched. All US-based participants of the PPMI study were invited to participate in this sub-study. Participants that accepted this invitation were provided with a Verily smartwatch to wear on either wrist for up to two years. Subjects were asked to charge the watch for one hour a day, when the collected data would be uploaded to the cloud. This resulted in a median daily wear of 21 hours and an interquartile range of monitoring between 254 and 560 days. For this study, participants were not instructed nor prompted to perform specific tasks but instead wear the device during their normal daily activities[53].

In total, 356 PPMI participants returned usable Verily watch data. Their demographics and differences between disease cohorts are detailed in Table 2.6. This table also outlines the difference in demographics between the different disease cohorts of the PPMI-Verily study. Most PPMI-Verily participants identified as White (95.8%) and had at least 16 years of education (81.5%). In addition, the prodromal cohort of the PPMI-Verily study had a higher female-to-male ratio than the healthy controls and diagnosed PD cohorts.

The Verily Watch contains a variety of sensors, collecting a wide range of signal data. This includes triaxial inertial sensors, capturing both accelerometry and gyroscopy at 80Hz. The device also contains sensors for capturing electrodermal activity and environmental sensors, providing valuable insights into participants' movements and physiological signals. The PPMI group have processed this data providing summary data derived from the accelerometer sensor, including daily steps, ambulatory data, as well as access to the raw data, to derive one's own models and measures.

Characteristic	Healthy Control (N=35)	Diagnosed PD (N=161)	Prodromal PD (N=158)	SWEDD (N=2)	Overall (N=356)
Age (years)					
≤ 49	4 (11.4%)	6 (3.7%)	2 (1.3%)	0 (0%)	12 (3.4%)
50-59	4 (11.4%)	22 (13.7%)	47 (29.7%)	1 (50.0%)	74 (20.8%)
60-69	14 (40.0%)	68 (42.2%)	75 (47.5%)	0 (0%)	157 (44.1%)
70-79	8 (22.9%)	57 (35.4%)	32 (20.3%)	1 (50.0%)	98 (27.5%)
80+	5 (14.3%)	8 (5.0%)	2 (1.3%)	0 (0%)	15 (4.2%)
Sex					
Female	15 (42.9%)	65 (40.4%)	100 (63.3%)	1 (50.0%)	181 (50.8%)
Male	20 (57.1%)	96 (59.6%)	58 (36.7%)	1 (50.0%)	175 (49.2%)
Ethnicity					
Asian	0 (0%)	4 (2.5%)	0 (0%)	0 (0%)	4 (1.1%)
Black	0 (0%)	3 (1.9%)	0 (0%)	0 (0%)	3 (0.8%)
Other	0 (0%)	1 (0.6%)	1 (0.6%)	0 (0%)	2 (0.6%)
White	35 (100%)	149 (92.5%)	155 (98.1%)	2 (100%)	341 (95.8%)
Unknown	0 (0%)	4 (2.5%)	2 (1.3%)	0 (0%)	6 (1.7%)
Years of Education					
≤ 11	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
12-15	8 (22.9%)	37 (23.0%)	19 (12.0%)	2 (100%)	66 (18.5%)
16+	27 (77.1%)	124 (77.0%)	139 (88.0%)	0 (0%)	290 (81.5%)

Table 2.6: Characteristics of the PPMI Verily population

2.4 Capture-24

Capture-24 is a large, open-access, wrist-worn accelerometer dataset, captured between 2014 and 2015, with the aim of bridging the gap between objective measures of movement, and labels of activity. It consists of accelerometer data captured from over 150 healthy volunteers from the Oxfordshire area. Initial recruitment involved 166 participants, however after excluding participants with poor quality data, 151 participants remained. Participants had a mean age of 41 years, and 66% were female. These volunteers agreed to wear an AX3 Axivity accelerometer on their dominant wrist, alongside an OMG Life Autographer camera mounted on their chest for up to 24 hours. In addition, participants submitted a sleep diary, to give an indication of in-bed times[76, 77].

The protocol for accelerometer wear aligned closely with that of the UK Biobank Physical Activity Monitoring study, excluding the period of monitoring. The accelerometer device captured three axes of data at 100Hz in a dynamic range of $\pm 8g$. The device continuously captured accelerometer data, for the 24 hours.

Ego-centric chest-mounted cameras were used to capture the ground truth of activities performed by individuals, synchronised to accelerometer wear. Pictures were automatically taken by the camera every 20-40 seconds. These images produced photographic evidence for the activities and surroundings of participants. During private moments of the day, the camera could be covered, producing black images that could not be used for annotations[77].

Activity annotations for this dataset were produced using a combination of captured images and sleep diaries. Sleep diaries were used to exclusively label overnight sleep behaviour. All other annotations, however, I produced using the timestamped captured images from the camera. Annotation labels were chosen according to the compendium of physical activity, labelling a wide array of different tasks. Annotators assigned a label to all accelerometer readings occurring in the window of three consistent camera images for a given activity.

The compendium of physical activity provides a high level of granularity of annotation labels, with over 200 possible labels. To summarise these detailed

annotations into more simplified labels that have more semantic meaning, I used Walmsley’s annotation schema to break activities down into sleep, sedentary, light and moderate-vigorous (MVPA), for which data collected is represented roughly 36%, 40%, 20% and 4%, respectively[50].

The synchronised camera images, when combined with accelerometer data, enable the training of activity recognition models that can be deployed on wrist-worn accelerometer datasets that lack ground truth activity annotations. The Capture-24 data set remains the largest form of openly accessible, free-living monitoring datasets to date. This allows it to serve the perfect role as an activity training dataset within this thesis, to get insights into the activity behaviour of populations monitored using wrist-worn accelerometers[77].

2.5 OxWalk

The OxWalk dataset comprises accelerometry data collected to better understand patterns of walking and facilitate the development of walking and step detection algorithms for free-living. One of the most fundamental features of modern smartwatches is the reporting of step counts as an indication of activity levels [78]. However, there is a limited number of open-access datasets to validate these step count algorithms, particularly in a home-based, free-living environment[79].

39 healthy volunteers from the Oxfordshire area were recruited to the OxWalk study between 2019 and 2020. Participants had a mean age of 38.5 years, ranging from 19 to 81 years old, and were 49% female. Participants were instructed to wear four triaxial accelerometers concurrently: two on the dominant wrist and two clipped to the dominant-side hip at the midsagittal plane, capturing accelerometer data at a sampling rate of 25 and 100Hz. In addition to accelerometers, a foot-facing action camera was mounted on the participant’s belt line to capture video of each of the participant’s steps.

For this thesis, I focused solely on the data captured by the wrist-worn accelerometer with a sampling rate of 100Hz. This choice was made for several reasons. Firstly, the 100Hz dataset was used to train the publicly available Python package

pypi:stepcount[79] allowing this work to build upon an existing model. Secondly, this sampling frequency more closely aligns with the raw sampling frequencies of other datasets used in this thesis. Lastly, the machine learning models used later in this analysis require an input sampling frequency of 30Hz[80]. Downsampling from 100Hz minimises information loss compared to using a 25Hz signal and then upsampling.[79, 80].

The camera data was synchronised with the accelerometer recordings, allowing for the annotation of steps in the accelerometer data files. Annotators parsed through all videos, identifying periods in which either foot was not making contact with the ground. These initial annotations were then processed to give the timings for the initial heel strike for each step taken by participants. In an additional annotation column in the accelerometer data, for the rows of the nearest time index, to an identified initial heel strike, the value is set to 1, otherwise 0.

While the base set of annotations solely provides insight into the timings of each step, further processing of these annotations can produce more complex labels of activity. One useful approach taken is to identify walking, given a threshold of steps taken within a period. During this thesis, I used the definition of walking as taking at least 4 steps in 10 seconds. With this definition, I was able to relabel the OxWalk accelerometer data with labels of walking or not walking.

The OxWalk dataset allows for the supervised training of both step count and walking detection models from accelerometry data. This is essential for deployment on the UK Biobank Physical Activity Monitoring and PPMI-Verily studies, for which I did not have ground truth annotations of steps or walking. Walking is of particular interest in Parkinson’s disease and this thesis’ aims, as modifications to gait as the result of the progression of the disease is well-documented[4, 81, 82]. It may, therefore, be critical to the monitoring of early indicators of the disease to observe changes in walking patterns.

2.6 Michael J. Fox Foundation Levodopa Response study

The Michael J. Fox Foundation funded Levodopa Response (MJFF-LR) study investigates the effects of the drug, levodopa, on individuals with PD. 28 participants were recruited to either a New York or Boston clinic to wear a series of activity sensors while performing various scripted tasks. The participants recruited into the study were limited to individuals who had self-reported that they had received a diagnosis of Parkinson’s disease. Participants had a mean age of 62.5 years ranging between 46 and 80, with 32% female[83].

This study was spread over 4 days, in which on day 1, participants would perform this variety of tasks, in the clinic before a specialist. On day 2 and 3, the participant would perform these same scripted tasks but at home. On day 4, the participant would perform the same regimen of tasks in the clinic, however, while off their regular levodopa medication unlike days 1-3. The full regimen of tasks to perform is shown in Table 2.7 which typically lasted 20 minutes and was repeated 6-8 times each day with a minimal interval of 30 minutes. The activities performed include activities of daily living, as well as items of the Movement Disorder Society - Unified Parkinson’s Disease Rating Scale(UPDRS) Part III[14]. For the purpose of this thesis, these activities are further simplified to labels of walking or not walking. While performing the tasks in the clinic on days 1 and 4, participants are monitored by a specialist, rating their UPDRS Part III score[14], giving an indication of the severity of the motor impairment due to PD.

In this work, I used data collected by a GENEActiv watch, placed on the participant’s limb most affected by PD. Raw accelerometer data from this watch was collected at a sample rate of 50Hz in three axes, however, measured in meters per second squared. Separate files were provided for the raw accelerometry data, and timings for which tasks were performed.

The MJFF-LR dataset allows for the supervised training of both activity classification and walking detection models from accelerometry data, for deployment on unmonitored accelerometry datasets. This dataset specifically allows for training

Activity Description	Label
Walking in a straight line for 30s	Walking
Walking in a straight line for 30s while counting backwards	Walking
Walking upstairs	Walking
Walking downstairs	Walking
Walking through a narrow corridor	Walking
Standing	Not walking
Finger-to-nose for 15s (twice with each arm)	Not walking
Alternating hand movements for 15s (twice with each arm)	Not walking
Drawing	Not walking
Typing on a keyboard for 30s	Not walking
Opening a bottle and pouring water (three times)	Not walking
Arranging sheets of paper in a folder (twice)	Not walking
Assembling nuts and bolts for 30s	Not walking
Folding a towel three times	Not walking
Sitting	Not walking

Table 2.7: Range of activities performed by participants in MJFF-LR study[83] and corresponding walking label used in this thesis

on a known impaired PD population that may have subtly different behaviours as compared to the healthy population. This study also captures the severity of motor impairment, offering insights into patterns of model behaviour.

2.7 Data generalisability

The wrist-worn accelerometer data used in this thesis come from three distinct research-grade accelerometers: the AX3 Axivity, GENEActiv, and Verily watches. There are notable differences in the raw output accelerometer data of these devices. Many research studies have investigated the effect of these differences, with a particular focus on any systematic error that may arise as a result. Rowlands et al. (2019)[84], however, has shown that while these differences exist, their extracted outcomes are comparable, allowing for cross-device generalisability.

Data collected from consumer-grade devices, such as the Apple Watch, Fitbit,

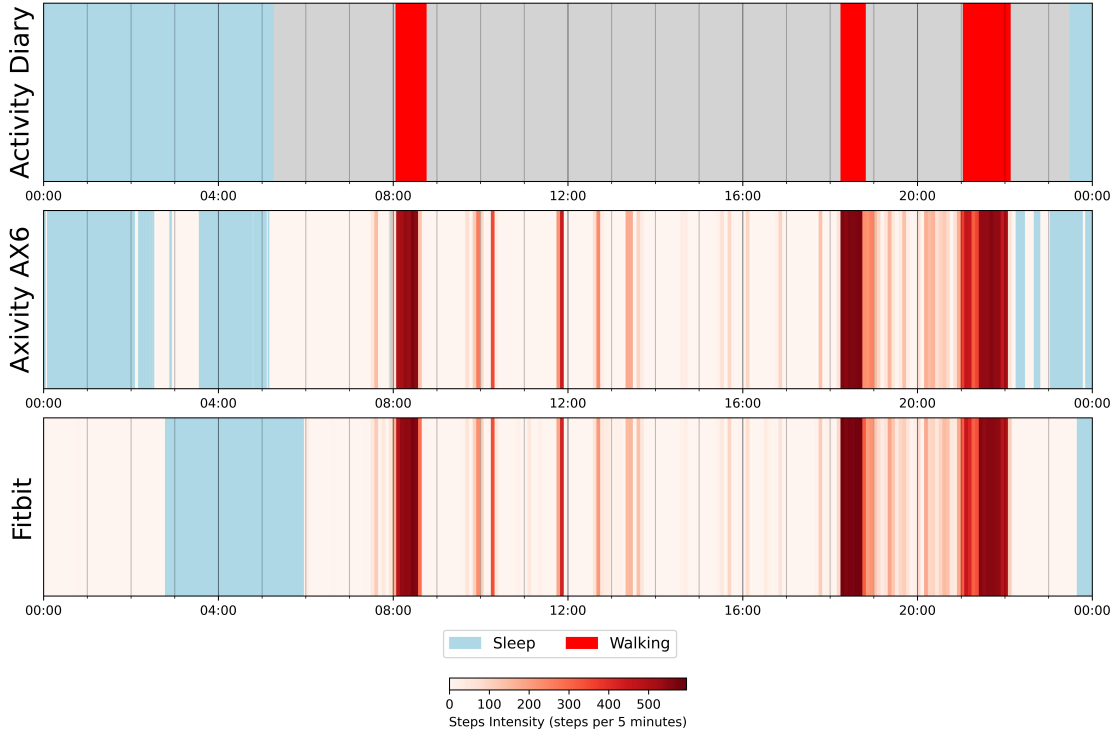


Figure 2.1: Comparison of output sleep and step counts from research-grade AX6 Activity (using `pypi:asleep` and `pypi:stepcount`) and exported Fitbit Sense 2 data.

and Garmin, have not been used in this thesis. White et al. (2024)[85] showed that these consumer-grade devices, particularly Apple Watch and Fitbit, had a high correlation with the research-grade ActiGraph wGT3X-BT accelerometer. Figure 2.1 displays data I collected from my personal Fitbit Sense 2 concurrently with a research-grade Axivity AX6, showing concordance in extracted step counts and timings. However, this comparison was limited to just a single day of monitoring, conducted on myself, and cannot offer broader generalisations, particularly to individuals with PD. Furthermore, while these research papers[84, 85] demonstrate some correlation between digital measures extracted from different devices, further work is needed to better understand how the results of this thesis may be generalised to other devices, including consumer and research-grade devices, but this is outside of the scope of this thesis.

2.8 Dataset preparation

It is essential to maintain the same preprocessing steps for the data, to ensure harmonisation between the data, and remove systematic error due to the different devices. The steps listed below were therefore performed on all datasets used in this thesis.

Firstly, data was converted to units of acceleration due to gravity (g : 9.81ms^{-2}). Next, devices are calibrated to gravity, to ensure that during stationary periods, the magnitude of acceleration is $1g$. To do so, readings are scaled by the factor for which it deviates from $1g$ [47]. The data is then passed through a lowpass 8th order Butterworth filter[86] with an upper band of 20Hz . Following this, the accelerometer signal is linearly down-sampled to 30Hz . This is to harmonise the sampling frequency across devices while meeting the Nyquist criterion for capturing the range of frequencies from human motion, as reported by Khusainov et al. (2013)[87]. This is also the expected sampling frequency for activity recognition models, used later in this work.

For the datasets lacking ground truth capture, namely UK Biobank physical activity monitoring and PPMI-Verily studies, device non-wear needed to be inferred. In past literature, non-wear was detected as periods of accelerometry for which the standard deviation of the signal remained below a threshold of 13mg for over 60 minutes [48, 49, 50, 55, 88]. To help reduce false positives, Small et al. (2024)[79] modified this approach to using a threshold of 15mg over 90 minutes, which I used in this thesis. Any data collected during these periods of detected non-wear were then removed. This non-wear detection and removal process was not applied to the datasets with ground truth capture, namely Capture-24, OxWalk, and the MJFF-Levodopa Response studies, as only labelled segments with known wear were used.

Figure 2.2 shows a 30-second sample of the raw data captured from the AX3 Axivity and GENEActiv watches and the harmonised pre-processed datasets prepared for use in this thesis.

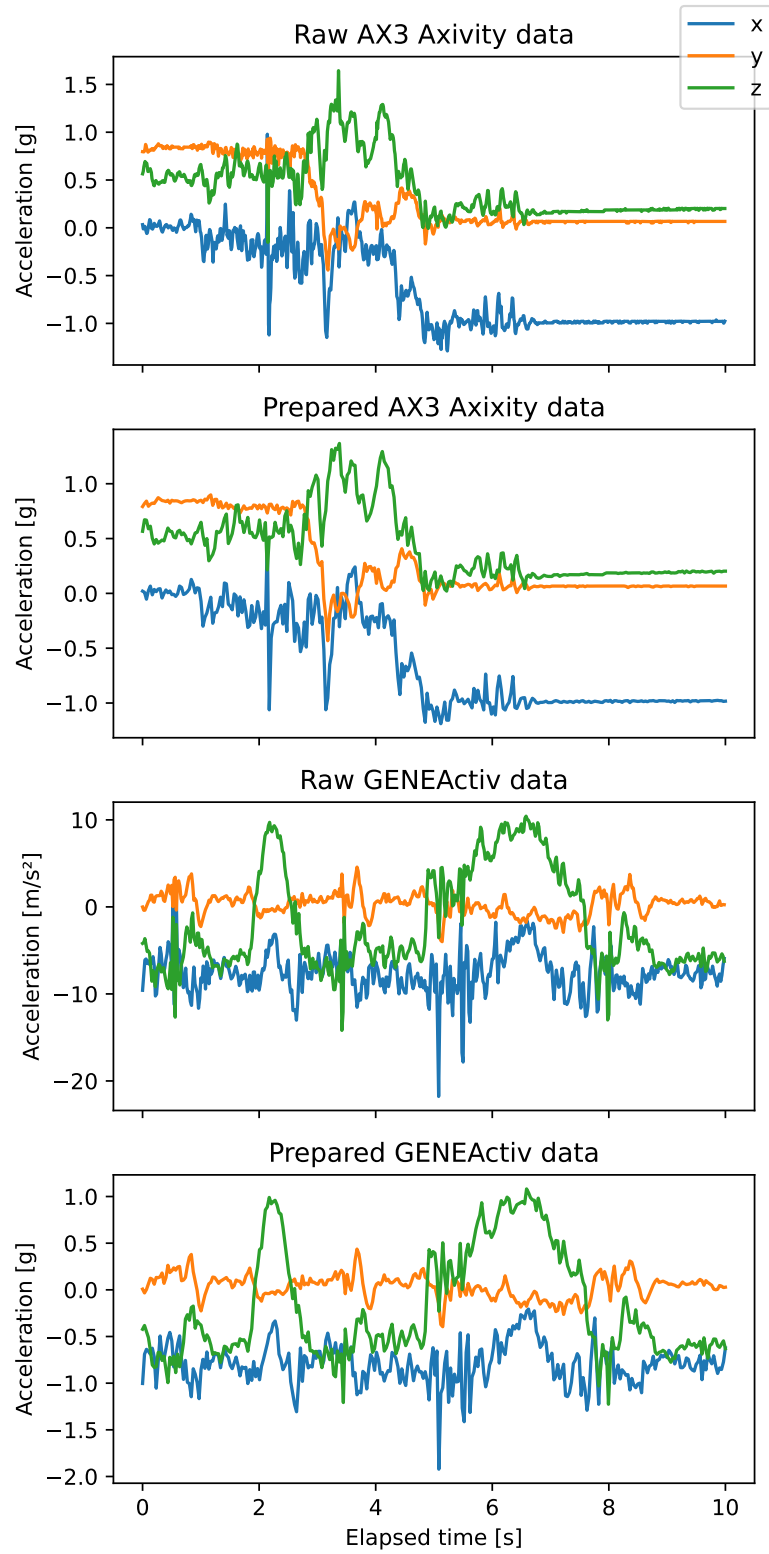


Figure 2.2: Sample plots of raw and pre-processed accelerometer data from AX3 Axivity and GENEActiv watches.

3

Improving Human Activity Recognition Models

Contents

3.1	Introduction	39
3.2	Actinet: Classification of Activity Intensity Using Self-Supervised Deep Learning	40
3.2.1	Introduction	40
3.2.2	Methods	44
3.2.3	Results	49
3.2.4	Discussion	57
3.2.5	Summary	60
3.3	Walking Recognition in Parkinson’s Disease Populations	60
3.3.1	Introduction	60
3.3.2	Methods	62
3.3.3	Results	64
3.3.4	Discussion	67
3.3.5	Summary	72

3.1 Introduction

In machine learning, Human Activity Recognition (HAR) models are used to identify periods of specific behaviours from interpreting raw signal data[89, 90]. In Chapter 1.2.1, I outlined the usefulness of such models when handling raw, wrist-worn ac-

celerometer data, to provide insights into the wearer’s activities. For purposes within this thesis, I looked to use the most state-of-the-art HAR models for characterising behaviour, which could then be associated with Parkinson’s disease (PD).

This chapter is split into two Sections for two separate and independent analyses carried out to improve existing activity recognition models. In Section 3.2, I cover the improvement of models used to categorise periods of activity intensity, namely, sleep, sedentary, light and moderate-vigorous activity. This is trained on the Capture-24 dataset[77], going from using a random forest (RF) model[52], to instead using modified version of an 18-layer ResNet-V2 model[80, 91], pre-trained using self-supervised learning (SSL). In Section 3.3, I cover the improvement of walking classifiers by training on both healthy and PD motor-impaired populations. The models trained within this section are then used in later Chapters 5 and 6.

3.2 Actinet: Classification of Activity Intensity Using Self-Supervised Deep Learning

3.2.1 Introduction

The rapid evolution of wearable technology, particularly smartwatches, has revolutionised the collection and analysis of human activity patterns, associated with health-related outcomes. All standard consumer-grade devices today have features capable of automatically detecting periods of different types of activity, including walking, running, sleeping etc. Large-scale population health studies similarly use research-grade devices, seeking to extract a breakdown of activity behaviour. In both these examples, behaviours must be inferred from the captured wearable data, relying on models capable of interpreting accelerometry into meaningful activity labels, i.e activity recognition models. The heightened use of research and consumer-grade smartwatches demands a greater need for reliable, generalisable and accurate activity recognition models.

Historically, the classification of different activity intensities was achieved using cut-point thresholds. These were predetermined values, which separate different types of activity based on intensity. Various studies have used these cut-point

thresholds for research purposes and encouraged their further use in association with health outcomes [92, 93, 94, 95, 96, 97, 98]. However, these cut-point thresholds lack generalisability, typically requiring different threshold values across various devices and studies[99].

Improved approaches to activity recognition instead used supervised machine learning models to classify activities. Various examples across different study populations using wrist-worn accelerometry to identify activities have used RF models [49, 50, 100, 101, 102, 103]. To do so, first, signal-related features are extracted from the raw accelerometry, which are chosen by researchers to be useful in separating different types of activity. These features are then input into the RF model to learn patterns of the features associated with each labelled activity behaviour. In some specific models [49, 50, 100], hidden Markov models (HMMs)[104] are then applied to the outputs of the RF prediction, to provide some smoothing to the output signal. An illustration of how these models work is shown in Figure 3.1.

The biobank accelerometer analysis tool, `pypi:accelerometer`, is a publicly available software package, implementing the categorisation of activity labels using RF and HMM trained on the Capture-24 dataset. While these models have shown great use, they are not without potential improvements. Primarily, activity recognition from the wrist remains a difficult classification task, hence, efforts remain to improve the accuracy of these algorithms. Moreover, while the classification of simpler, more prevalent tasks remains easier for the RF model, the separation of light activity from sedentary and moderate-vigorous activity shows poorer performance. Finally, the choice of hand-crafted features extracted from accelerometry data leads to a large loss of information. While this is advantageous to build simpler models, this relies on the careful selection of useful features, hypothesised to have an association to predicted activity labels. However, some of these features have shown limited separability between activities, hence, the model may benefit from a hypothesis-free, data-driven approach to feature extraction.

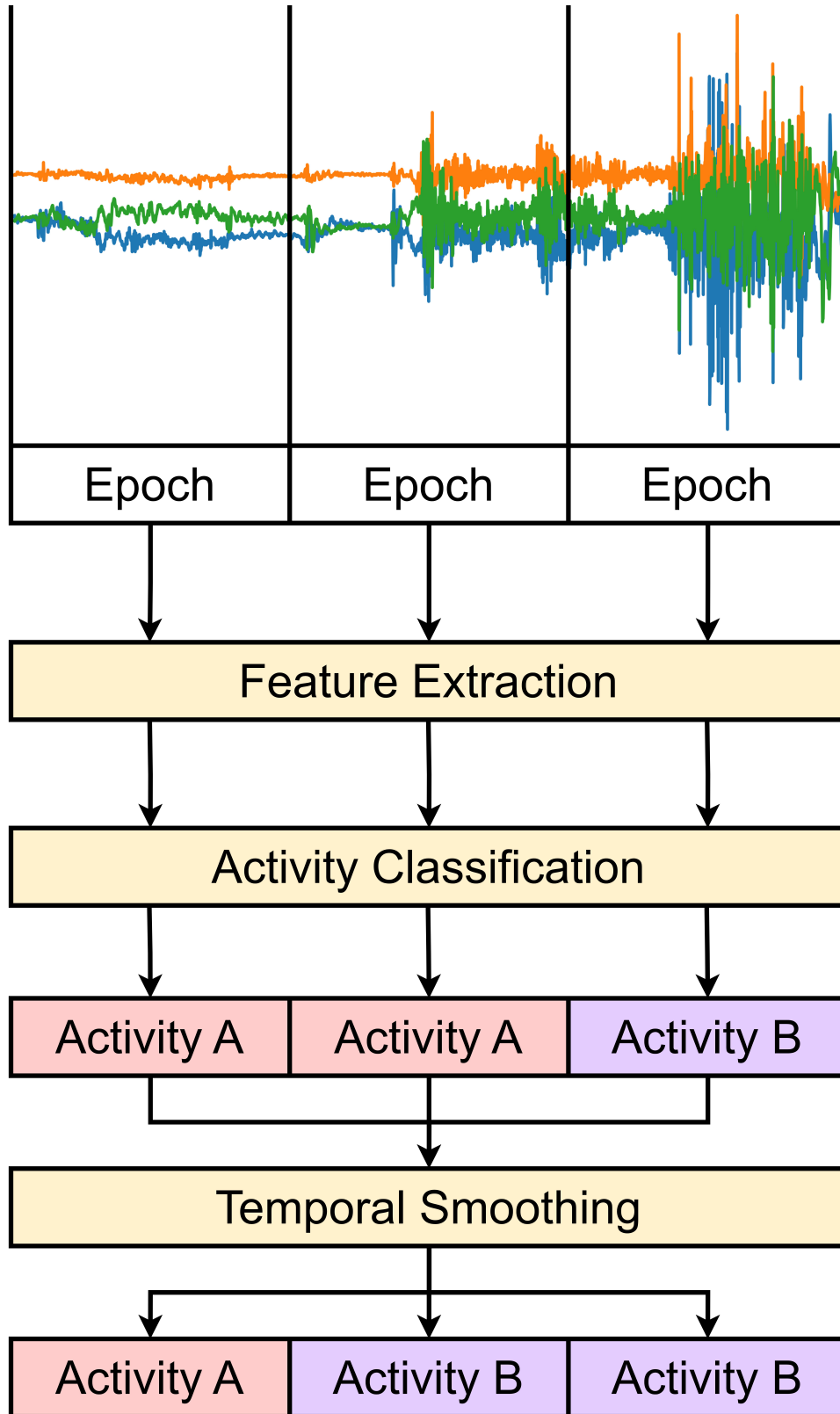


Figure 3.1: Illustration of human activity recognition models, converting raw accelerometry to labels of activity.

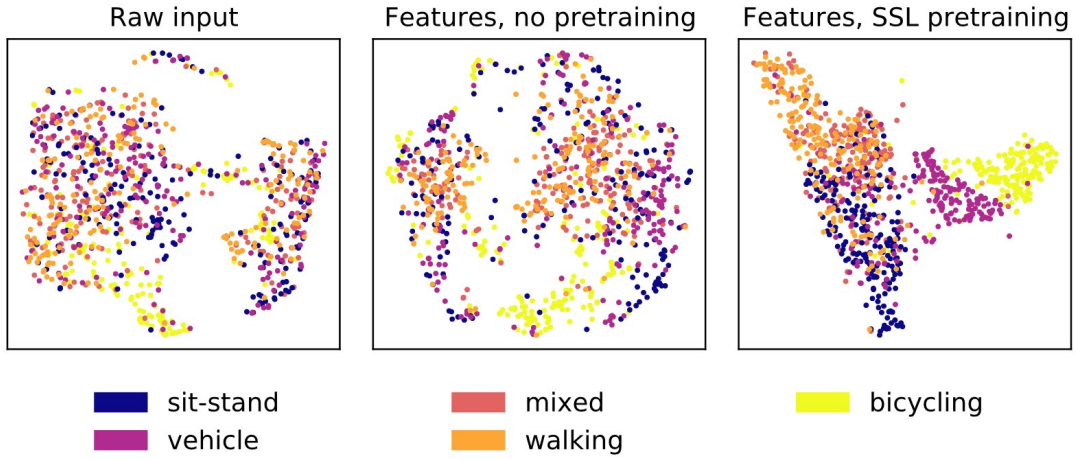


Figure 3.2: Uniform Manifold Approximation and Projections (UMAPs)[105] showing the separation of different activities using hand-crafted and deep learned features in Capture-24, taken from Yuan et al. (2024)[80]

The work by Yuan et al. (2024)[80], for which I am a co-author, leveraged the latest approaches in machine learning for the processing of accelerometer data. For this work, a modified, 18-layer ResNet-V2 model[91] was used to train a feature extractor and classifier from an accelerometry signal. To do so, this feature extractor was first SSL pre-trained using one day of data from the UK Biobank (UKBB) physical activity monitoring study population, trained to classify the difference between forward and reverse arrow of time signals. After pre-training, the deep layers of the model were frozen, while the final layer was fine-tuned to the specific classification task. The deep-learned feature extractor, as compared to the hand-crafted features used by the RF model, showed greater separability when visualised using Uniform Manifold Approximation and Projections (UMAP)[105], shown in Figure 3.2. This also showed better classification performance as compared to a RF model when trained on the Capture-24 dataset[80]. It was, therefore, believed that the use of this pre-trained, SSL ResNet-18 model with HMM smoothing may produce an improved classifier for the extraction of activity intensities from wrist-worn accelerometry.

However, there remained many unknowns to this work. Firstly, though the ResNet-18 model outperformed the RF model, this analysis did not include HMMs in the pipeline. Using just the ResNet-18 model in Yuan et al. (2024)[80], the

classification performance is worse when compared to the RF + HMM smoothing model in `pypi:accelerometer`[50]. It is therefore required to explore the performance of the SSL ResNet-18 + HMM for improved performance. In addition, it is uncertain whether this new model would have biases in its performance, for different age groups or sexes, limiting its generalisability. Finally, it remained unclear whether deploying this model would substantially change the predicted outputs of the composition of daily activity.

Therefore for this analysis, I proposed the design of `pypi:actinet`, an HAR model that would use an SSL ResNet-18 model and HMM smoothing, trained on Capture-24 to categorise activity intensities from wrist-worn accelerometry. Similar to the existing `pypi:accelerometer`, this model will then extract useful metrics related to activity composition, for the association to health outcomes. `pypi:actinet` would then be made publicly available and free for academic use as a pip package.

3.2.2 Methods

This study aimed to compare an existing activity recognition model, Accelerometer (`pypi:accelerometer`)[50], to a newly proposed model, ActiNet (`pypi:actinet`). These models were both trained on the Capture-24 dataset[76, 77] to classify labels of activity intensity: sleep, sedentary behaviour, light activity, and moderate-vigorous activity. This research investigated how these activity recognition models compare in terms of classification performance and output, with a particular focus on investigating any biases within subgroups of age and sex.

Accelerometer is an RF[52], followed by HMM[104] smoothing. The RF has 2,000 estimators, with a not minority balanced sampling strategy with replacement, to ensure training of all activity intensity labels despite imbalance within the dataset. The output classifications of the RF are input as a sequence of observations to the HMM, which uses the Viterbi algorithm[106] to calculate the most likely sequence of underlying states. These form the classification outputs of the Accelerometer model.

ActiNet, on the other hand, uses a modified version of an 18-layer ResNet-V2[91] model, followed by HMM smoothing. The deep layers of the network are SSL pre-trained, as proposed by Yuan et al. (2024)[80]. The final layer of the network was fine-tuned to the specific classification task, with an Adam optimiser[107], a learning rate of 0.0001, patience of 5 epochs before early stopping and training data augmentation of random rotations and axes swapping. The model was trained using inverse class weighting to ensure balanced training of all activity intensity labels. The output classifications of the ResNet-18 model are input as a sequence of observations to the HMM, with the most likely sequence of underlying states calculated using the Viterbi algorithm, forming the classification outputs of the ActiNet model.

The Capture-24 dataset was collected and pre-processed, with up to 24 hours of annotated wrist-worn accelerometer data from 151 participants. More details on this dataset are available in Chapter 2.4. As part of this data preparation, the raw accelerometer data was prepared for input into the respective models, following approaches used by Chan et al. (2024)[77]. For both models, the full dataset was segmented into non-overlapping 30-second windows. Each window was assigned a ground truth label based on the most frequently occurring annotation within that period. This labelling approach was still applied even if some timestamps within a window were unannotated, however, windows were excluded entirely if all timestamps lacked annotations. Additionally, any windows containing missing accelerometer data, due to NaN values in the signal or the data not spanning the full 30 seconds, were also excluded. The windowed annotation was further simplified to a label of activity intensity using the Walmsley mapping[50]. For `pypi:accelerometer`, signal features were extracted for the full population, listed in Table 3.1. For `pypi:actinet`, the accelerometer signal was linearly down-sampled from 100Hz to 30Hz to meet the expected input shape of the pre-trained ResNet-18 model.

To investigate how `pypi:actinet` performed compared to `pypi:accelerometer`, we first used the training approaches taken by Chan et al. (2024)[77], in which both models were trained on the first 100 participants (P001-P100). For `pypi:accelerometer`, the RF model was trained on all 100 participants' data. The out-of-bag estimates

for the RF model, calculated as the average error of predictions from trees not containing certain observations in their bootstrap samples[108, 109], and the ground truth labels were then used to train the HMM. Training an HMM requires the calculation of the prior, transition, and emission matrices, which are necessary for the Viterbi algorithm[106]. The prior was calculated as the distribution of ground truth labels, the transition matrix as the probability of moving between states based on ground truth labels, and the emission matrix as the confusion between ground truth labels and the RF out-of-bag estimates.

For `pypi:actinet`, on the other hand, a train validation split was made on the first 100 participants, with 20% (20 participants) randomly selected to be held out. The final layer of the SSL ResNet-18 was fine-tuned to the activity recognition tasks using the remaining 80 participants with the held-out participants used for early stopping of the model. In addition, once the model had finished training, it was deployed on the 20 held-out participants' data, with the output logits for the classification model and ground truth labels used to train the HMM. The prior and transition matrices are calculated similarly to `pypi:accelerometer`, but only using 20 participants' data. The emission matrix was calculated using the confusion between the ground truth labels and the output logits made on the 20 held-out participants' data.

Both models were then deployed on the remaining 51 test participants' (P101-P151) data. To compare the performance of these models, I extracted metrics for output classifications made by the model compared to the ground truth labels using accuracy, balanced accuracy, Cohen's kappa and macro F1 score, which was necessary due to class imbalance in activity intensity labels. Note that performances are reported using mean $\pm$ standard deviation in the per-participant activity classification performance. To investigate potential biases in classification performance or outputs between the two models, particularly within subgroups of age and sex in Capture-24, the `pypi:actinet` and `pypi:accelerometer` models were then trained using 5-fold stratified group cross-validation[110]. This approach leveraged the full dataset for training and evaluation while ensuring that all valid

data was used for testing in only one fold. Group cross-validation, where the ‘group’ refers to participant ID, prevented data from the same participant being used in both training and testing. Stratification by activity intensity label ensured a fair representation of all activity intensity labels in both training and testing, giving the model a fair opportunity to learn in each fold. The participant IDs used for testing in each fold are reported for the reproducibility of these validation splits.

A `pypi:accelerometer` and `pypi:actinet` model were trained for each fold, using 80% of participant data for training, and holding out 20% for testing. For `pypi:accelerometer`, all training data was used to train the RF classifier, with out-of-bag estimates applied to train the HMM as previously outlined. For `pypi:actinet`, a single, nested, 80:20 group shuffle split was applied within the training data (80% of the full population) for each fold. As a result, data from 64% of participants was used to train the ResNet-18 model, while 16% was held out for validation, used for early stopping and HMM training as outlined previously. The model was then evaluated on the 20% of participants in the held-out test set. Performance across all participants when they were in the held-out test set during the nested cross-validation was aggregated across folds for evaluation.

To compare the model performance in different subgroups within Capture-24, participants were divided according to sex (female, male) and age group (18-29, 30-37, 38 – 52 and 53+). Model performance within these subpopulations was compared visually using box plots and confusion matrices. In addition, Bland-Altman plots[111] were used to compare the bias in the classification outputs between the two models.

Finally, I created the deployed version of the ActiNet model, deployed to `pypi:actinet` v0.1.0. Following the approach taken by `pypi:accelerometer`, the model is trained on the full Capture-24 population. To do so, first, a single train validation split is made of 20% of the participants’ data. This is used for early stopping of the ResNet-18 model and training of the HMMs as outlined previously. This model was then deployed on the full UKBB physical activity monitoring study population alongside the `pypi:accelerometer` v7.1.1 model for comparison.

Feature Name	Description	Units
Moment Features		
xMean, yMean, zMean, mean	Mean of x, y, z, v	g
xStd, yStd, zStd, sd	Standard deviation of x, y, z, v	g
xRange, yRange, zRange	Range of x, y, z	g
xyCov, xzCov, yzCov	Covariance of xy, xz, yz	g^2
min, max	Minimum and maximum of v	g
25thp, median, 75thp	25th, 50th, 75th percentile of v	g
coefvariation	Coefficient of variation of v	
MAD	Mean absolute deviation of v	g
MPD	Mean pairwise difference of v	g
skew, kurt	Skewness, Kurtosis of v	
enmoTrunc, enmoAbs	Truncated, Absolute ENMO	g
Correlation Features		
autocorr	Auto-correlation of v	
corrxy, corrxz, corryz	Correlation between xy, xz, yz	
Angle Features		
avgyaw	Average yaw angle	rad
sdroll, sdpitch, sdyaw	Standard deviation of roll, pitch, yaw	rad
rollg, pitchg, yawg	Roll, Pitch, Yaw angle in gravity frame	rad
avgArmAngel	Average arm angle	rad
avgArmAngelAbsDiff	Average absolute difference in arm angle	rad
Spectral Features of v		
fmax	Maximum frequency	Hz
pmax	Maximum power	g^2/s
fmaxband	Frequency band of maximum power	Hz
pmaxband	Power of maximum frequency band	g^2/s
entropy	Spectral entropy	nats
fft1, fft2, ..., fft10	Power spectral density for frequencies 1Hz, 2Hz, ..., 10Hz	g^2/s
f1, f2	1st and 2nd dominant frequencies	Hz
p1, p2	Power spectral densities of respective dominant frequencies	g^2/s
f625	Dominant frequency between 0.6 and 2.5Hz	Hz
p625	Power at dominant frequency between 0.6 and 2.5 Hz	g^2/s
totalPower	Total power	g^2/s

Table 3.1: `pypi:accelerometer` feature names and descriptions. ENMO is the Euclidean Norm Minus One[48]. rad is radians. g is acceleration due to gravity. v is the Euclidean norm of the signal in x, y and z.

Both `pypi:accelerometer` and `pypi:actinet` were deployed on the UKBB Physical Activity Monitoring population[55], explored in further detail in Section 2.2. Necessary exclusions needed to be made for less reliable data. These exclusions included any participants with accelerometer data for which either model:

1. Failed to produce outputs. This was typically due to empty or corrupted files or issues running the model on the data file.
2. Failed to be calibrated. This can occur if there are insufficient stationary samples, poor distribution of samples, the final calibration error remains too high, or the optimisation process does not converge within the maximum iterations.[47]
3. Failed to detect a device wear time of at least 3 days.
4. Failed to detect device wear at each hour of the day.
5. Detected unrealistically high overall acceleration values.

Using the remaining population available for this study, I compared the outputs of both models. This was done by first comparing the overall observed rate of activity intensities over the 24-hour cycle. Following this, for each activity intensity, Bland-Altman plots[111] were produced, comparing outputs of `pypi:accelerometer` and `pypi:actinet`.

This was further investigated for any patterns in subgroups in the UKBB of sex (female, male) and age (40-50, 50-60, 60-70 and 70-80).

3.2.3 Results

The Accelerometer and ActiNet models were trained and validated on the Capture-24 dataset. From this dataset, a total of 318,686 30-second annotated frames collected from 151 participants were used. This corresponds to an average of 17.6 hours of data per person. From ground truth annotations, roughly 40% of windows were labelled as sedentary behaviour, 36% sleep, 20% light activity and 4% moderate-vigorous activity.

First, I internally validated the model by training on the first 100 participants (P001-P100) and evaluating on the remaining 51 participants (P101-P151), corresponding to 213,375 and 105,311 30-second windows, respectively. The prior, transition and emission matrices trained for the HMM of both models are provided in Appendix Section A.1. The trained transition matrices are similar across the two models. However, the emission probabilities along the principal diagonal, representing the likelihood of correctly observing each true activity state, are higher for sleep, sedentary, and light activity in ActiNet’s HMM model, compared to Accelerometer’s.

Table 3.2 and Figure 3.3 display the performance metrics and confusion matrices for the `pypi:accelerometer` and `pypi:actinet` models. Example deployments of these models on three test participants, compared to ground truth labels of activity intensity, are displayed in Figure A.1. `pypi:actinet` outperforms the `pypi:accelerometer` model according to all evaluation metrics used. The ActiNet model achieved a mean ($\pm$ standard deviation) macro F1 of 0.81 ± 0.10 , and balanced accuracy of 0.83 ± 0.09 as compared to 0.79 ± 0.11 and 0.82 ± 0.09 for the Accelerometer model, respectively. However, the ActiNet model had slightly lower accuracy in the classification of moderate-vigorous activity, as compared to `pypi:accelerometer`, when trained and tested using this 100:51 split.

To further investigate the difference in performance between the two models, both were trained and evaluated using 5-fold stratified group cross-validation. The comparison of the performance of these models is displayed in Table 3.3 and Figure 3.4. The ActiNet model consistently outperformed the Accelerometer model across all evaluated metrics used. Specifically, ActiNet had a higher mean ($\pm$ standard deviation) macro F1 score of 0.83 ± 0.11 , compared to 0.77 ± 0.11 for the Accelerometer model. Similarly, it had a superior balanced accuracy of 0.86 ± 0.09 , compared to 0.81 ± 0.10 , respectively. To compare the output classifications made by both models, Bland-Altman plots, presented in Figure 3.5 were used.

In addition to comparing the summarised overall difference in performance, it is also useful to observe the per-participant difference in performance between the

Model	Accuracy	Balanced Accuracy	Cohen's kappa	Macro F1
Accelerometer	0.874 ± 0.077	0.815 ± 0.088	0.808 ± 0.111	0.786 ± 0.105
ActiNet	0.895 ± 0.062	0.833 ± 0.087	0.840 ± 0.090	0.806 ± 0.095

Table 3.2: Internal validation of `pypi:accelerometer` and `pypi:actinet` trained on the first 100 participants of the Capture-24[76, 77] dataset and tested on the remaining 51 participants. Scores are reported as mean $\pm$ standard deviation.

Model	Accuracy	Balanced Accuracy	Cohen's kappa	Macro F1
Accelerometer	0.881 ± 0.071	0.807 ± 0.098	0.815 ± 0.105	0.773 ± 0.112
ActiNet	0.917 ± 0.050	0.863 ± 0.091	0.870 ± 0.075	0.827 ± 0.105

Table 3.3: Internal validation of `pypi:accelerometer` and `pypi:actinet` using 5-fold stratified group cross-validation on the Capture-24[76, 77] dataset. Scores are reported as mean $\pm$ standard deviation.

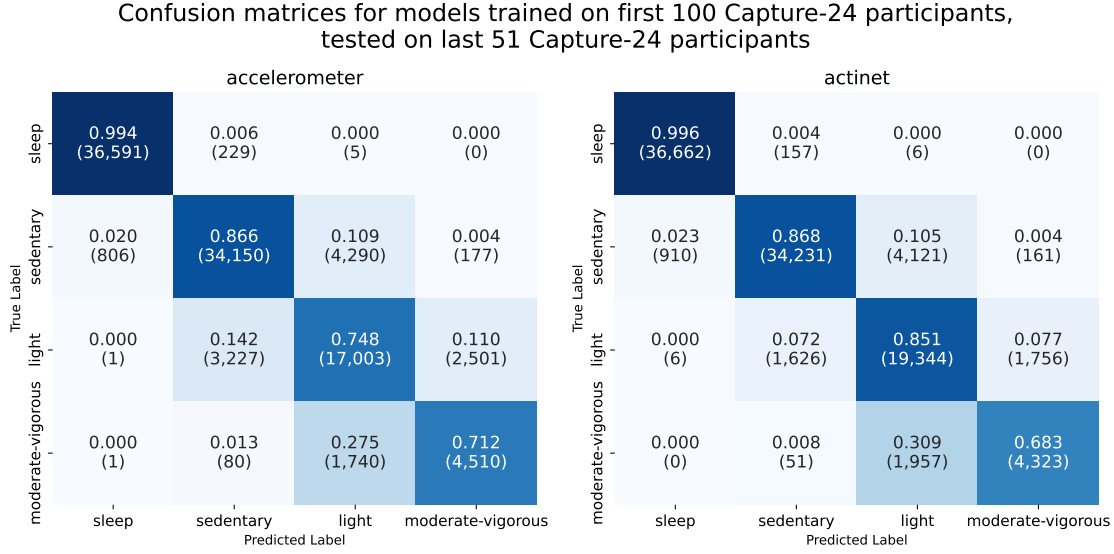


Figure 3.3: Confusion matrix for performance of activity recognition models classifying labels of activity intensity trained on first 100 participants of Capture-24 (P001-P100), evaluated on final 51 participants of Capture-24 (P101-P151). Count of 30-second windows provided in parentheses.

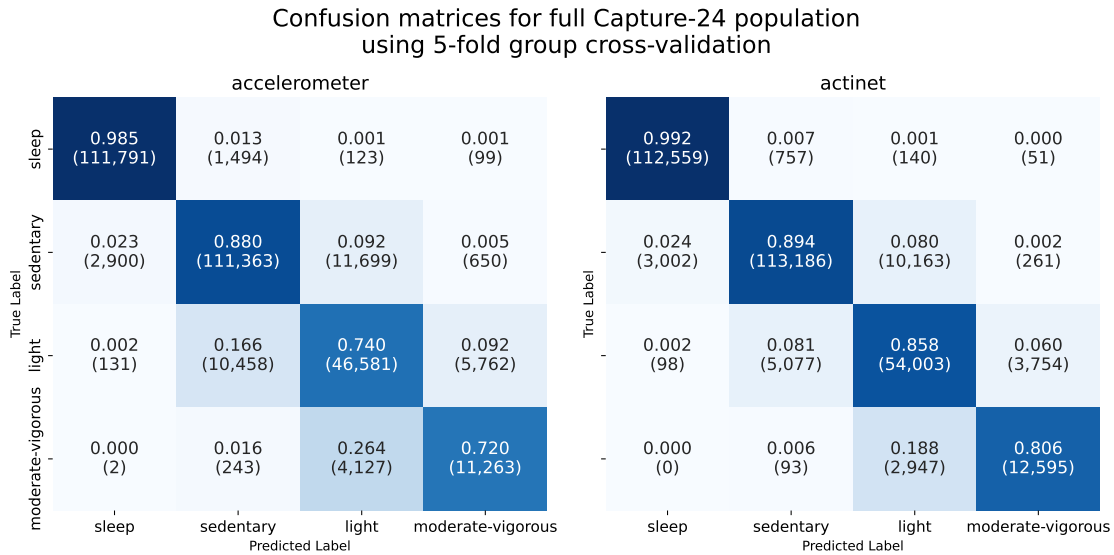


Figure 3.4: Confusion matrix for performance of activity recognition models classifying labels of activity intensity using 5-fold stratified group cross-validation. Count of 30-second windows provided in parentheses.

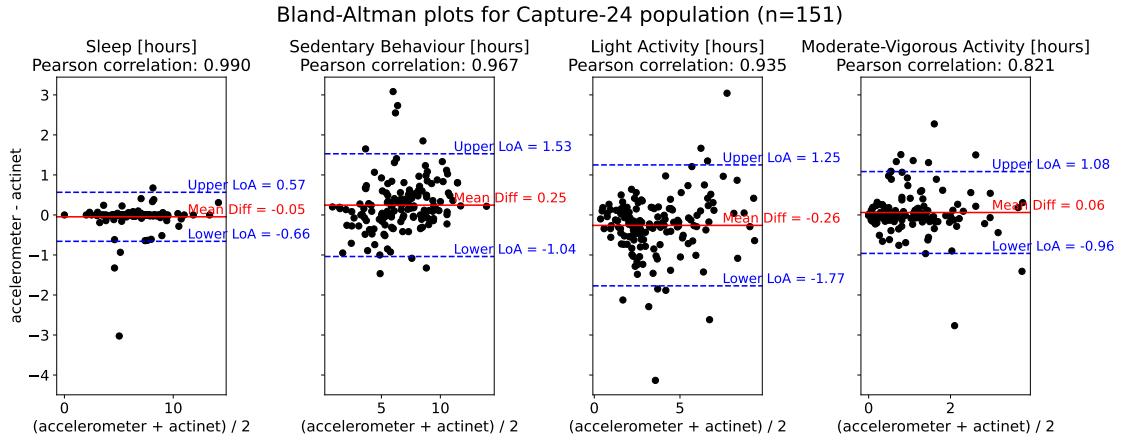


Figure 3.5: Bland-Altman plot for comparing classifications made by activity recognition models for labels of activity intensity trained using 5-fold stratified group cross validation on the Capture-24 dataset.

two models. This is illustrated in Figure A.2. I performed paired t-tests for each evaluation metric to determine whether the difference in performance between the two models significantly deviated from 0. In all metrics, p-values far less than 0.05 were found from these tests, indicating that evaluation metrics were significantly higher in `pypi:actinet` than `pypi:accelerometer`.

The performance improvement of `pypi:actinet` relative to `pypi:accelerometer` was then investigated for generalisability, across subpopulations of interest. Figure 3.6 displays the performance of both models within populations separated by age or sex.

These model architectures, trained on the full Capture-24 population, were then saved. Using these packages, the models can be deployed to extract the labels of activity intensities at a time point in monitoring. An example deployment on a sample participant is shown in Figure A.8.

These models were then deployed on the UKBB Physical Activity Monitoring population. Figure 3.7 shows the exclusion flow diagram for participants to be used in this study. After all exclusions, 96,283 participants data was used within this study. Of these participants, roughly 56% were female. Study participants ranged in age between 43 and 79 at the time of device wear, with 19% of participants aged 70 and above.

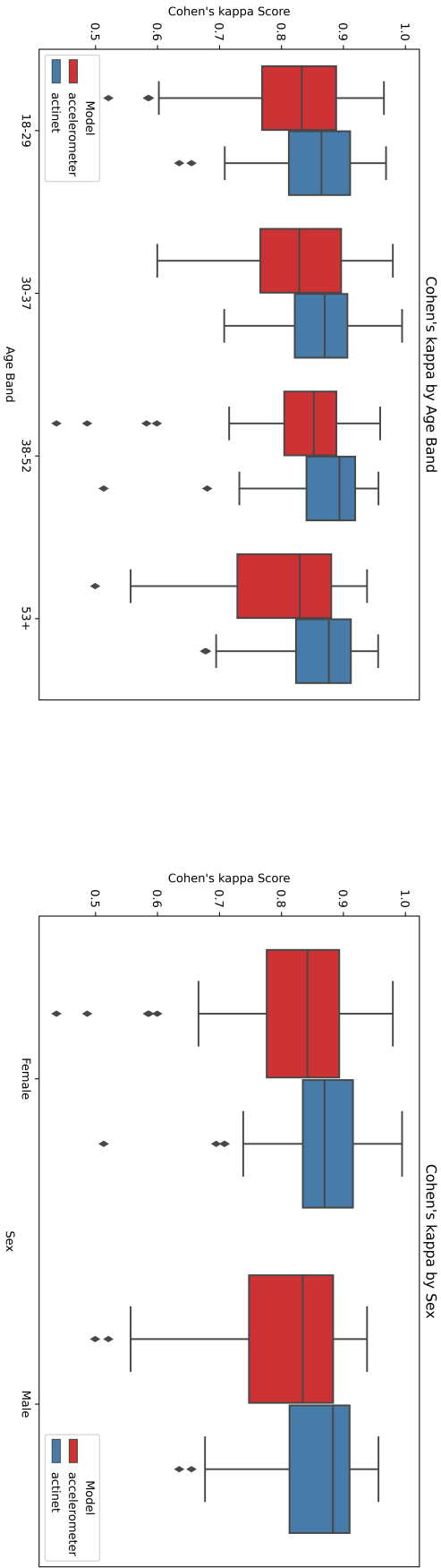


Figure 3.6: Cohen's kappa score for activity intensity classification in subpopulations of Capture-24.

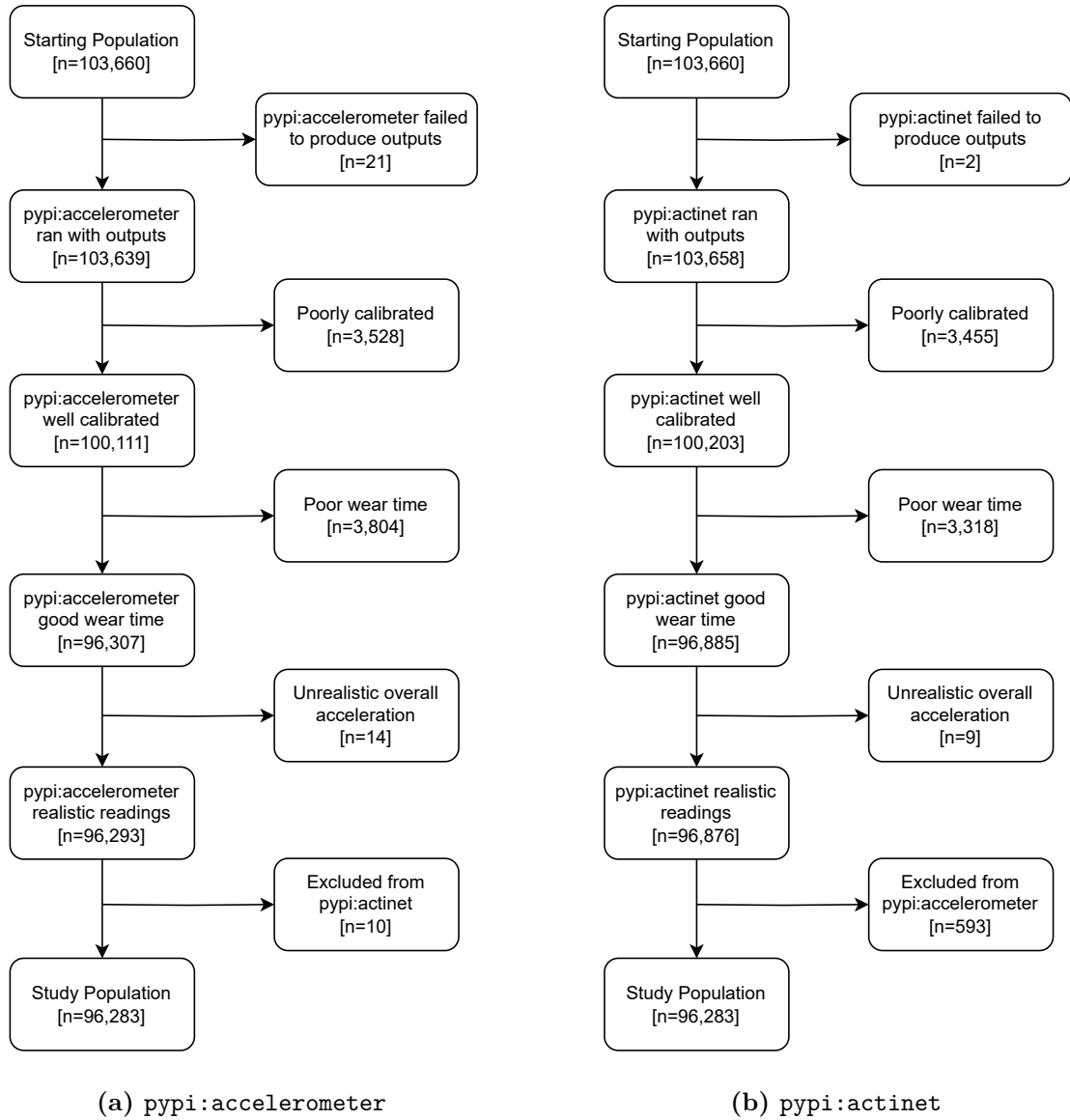


Figure 3.7: Exclusion flow diagram for activity recognition models' deployment in UK Biobank Physical Activity Monitoring study.

Figure 3.8 shows the observed rate of activity intensities classified by both HAR models when deployed on the UKBB physical activity monitoring population. To further compare the differences in the deployed output classifications made by the different models, I use Bland-Altman plots, shown in Figure 3.9.

Differences between the models were most noticeable in the prediction of light activity. The Pearson correlation coefficient for overall light activity predicted per person was 0.632 between the two models, indicating a strong relationship[112]. On

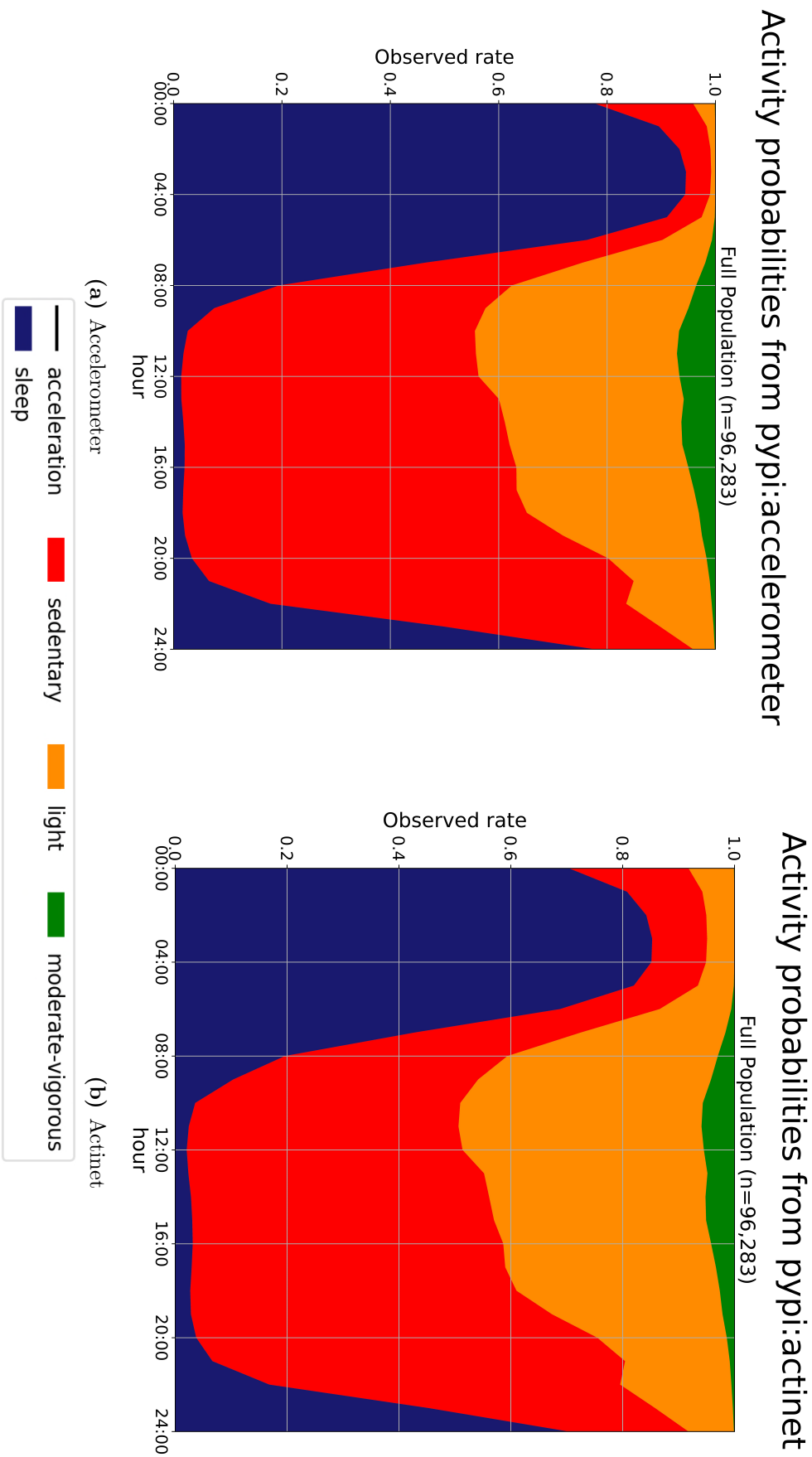


Figure 3.8: Observed rate of activities from predictions made by deployment of pypi:accelerometer and pypi:actinet in UK Biobank Physical Activity Monitoring population.

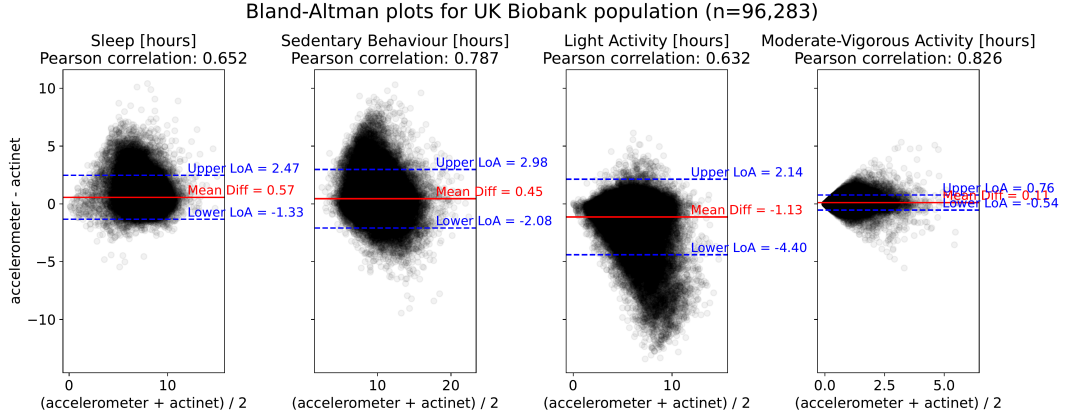


Figure 3.9: Bland-Altman plots comparing model prediction outputs of `pypi:accelerometer` and `pypi:actinet` deployed on the UK Biobank physical activity monitoring population.

average, `pypi:actinet` predicts an extra 5% of the day was spent in light activity in the study participants as compared to `pypi:accelerometer`, which corresponds to roughly an extra hour. Moderate-vigorous activity, however, shows a very strong correlation of 0.826, with a mean bias between accelerometer and actinet of 0.00[112].

Attached in the Appendix are similar plots showing the difference in outputs of `pypi:accelerometer` and `pypi:actinet` models in subpopulations based on age groups and sex. These plots showed the same patterns in differences to those seen in the full population.

3.2.4 Discussion

This study investigated the potential of an SSL ResNet-18 model combined with HMM smoothing to enhance the classification performance of a HAR model for detecting activity intensity labels from wrist-worn accelerometry. The results were promising, with the SSL ResNet-18 + HMM outperforming the RF + HMM across all metrics evaluated, as determined using both a single train-test split, and 5-fold stratified group cross-validation. This performance improvement was statistically significant according to paired t-tests. Furthermore, the enhancement in performance was consistent across different subgroups of the validation cohort, including both sexes and all age groups of interest. The model maintained consistent performance across these subgroups. The deployment of the new

`pypi:actinet` model demonstrated strong to very strong correlation with the existing accelerometer model.

The SSL ResNet-18 model exhibited superior classification performance compared to the balanced RF model. For the RF model, hand-selected features were extracted from the raw accelerometry data, which were presumed to be effective in distinguishing different activity intensities. However, these features may not fully capture critical differences between types of activity. Yuan et al. (2024) [80] demonstrated that self-supervised learned features extracted from accelerometry more effectively differentiate similar activities of daily living compared to hand-crafted features. This approach to feature extraction, which is hypothesis-free regarding difference accelerometry patterns associated with activity intensity labels, enables the use of crucial features related to activity behaviour.

Hidden Markov models continue to serve as an effective method for smoothing output predictions. Without this smoothing, the 30-second windows are treated as independent and identically distributed, leading to significant fluctuations in predicted labels. Instead, HMMs are an effective tool to account for the temporal component of activity behaviour, specifically the probability of transitioning from one activity to another every 30 seconds, thereby improving model performance.

The `pypi:actinet` model, when evaluated through cross-validation, demonstrated consistent performance across age and sex-based subpopulations within the Capture-24 dataset. In contrast, `pypi:accelerometer` showed a lower median macro F1 score in populations aged 53 and above compared to those aged 38 to 52, as seen in Figure 3.6. The ability of the model to generalise to older populations is crucial for its deployment in the UKBB Physical Activity Monitoring study, where the majority of participants are older adults aged 53 and above. This generalisability is also vital for research into PD, which is more prevalent in older populations.

The exclusion criteria for valid participants in this study differed slightly between the `pypi:accelerometer` and `pypi:actinet` models. These two models have distinct pipelines for pre-processing and deployment, which explains some of these differences. One key distinction is that `pypi:actinet` used `pypi:actipy` to read

and pre-process the raw accelerometry files. In doing so, `pypi:actipy` employs a different approach to detect non-wear periods, using a rolling window of 10 seconds with a patience of 90 minutes, during which the standard deviation of acceleration should not fall below 15 mg. This is a less stringent definition of non-wear compared to `pypi:accelerometer`, allowing a greater number of frames to be considered as wear time and reducing the likelihood of failing the insufficient wear time check. As a result, using `pypi:actinet` allows for more data to be available for analysis compared to `pypi:accelerometer`.

The deployment of the two models in the UKBB population revealed similar patterns in the overall rate of label predictions throughout the 24-hour cycle. This consistency increases confidence in the reliability of the results, as indicated by moderate to high correlation coefficients in predicted labels. These correlations were also consistent across all subpopulations of age and sex.

However, some noticeable trends in the differences between the predictions made by the two models were observed. Firstly, `pypi:actinet` predicted a noticeable increase in sedentary and light activity instead of sleep during night hours. The Capture-24 dataset contains sleep annotations obtained from sleep diaries, which may indicate in-bed time rather than actual sleep. The different models may handle the training of these imperfect labels differently, possibly leading to the `pypi:actinet` model being more sensitive to movement during in-bed times, instead classifying this as sedentary or light activity. Furthermore, the change in emission probabilities, due to the superior performance of the SSL ResNet-18 compared to the RF during training of the HMM, may allow for the detection of short windows of behaviour. These short periods of movement may now be detected more frequently by `pypi:actinet`, whereas they would have been smoothed out by the `pypi:accelerometer` model, leading to differences in the volume of detected activity intensity labels between the two models.

This research effort has several strengths. Capture-24 remains the largest annotated dataset of wrist-worn accelerometry, and training on this dataset provides

high reliability. The hypothesis-free nature of the model, with no hand-crafted features extracted, addresses concerns regarding the model’s generalisability.

However, there are limitations to this study. One significant limitation is the lack of external validation, primarily due to the limited availability of suitable datasets. Additionally, the training of HMMs in this study used a smaller dataset than that used for `pypi:accelerometer`, relying on the validation data on which early stopping was performed. This may reduce the reliability of the HMMs in comparison to those used in the `pypi:accelerometer` model. Alternative approaches could include the use of Long Short-Term Memory (LSTM) networks instead of HMMs, or implementing nested cross-validation to train the HMMs.

3.2.5 Summary

This research has investigated using `pypi:actinet` over `pypi:accelerometer` for activity recognition in the UK Biobank physical activity monitoring study. The results show `pypi:actinet` to be the superior model in performance and generalisability. This therefore encourages the use of `pypi:actinet` with this thesis as well as other similar research studies.

3.3 Walking Recognition in Parkinson’s Disease Populations

3.3.1 Introduction

Walking recognition models are a type of HAR model, used to identify periods of walking. As walking is the primary mode of human movement, many insights can be garnered from its monitoring, relying on its accurate detection. Walking consists of a series of gait cycles, which serve as its basic and essential unit. A gait cycle can be described as the combination of motor activities occurring between two consecutive ground contacts (typically heel strikes) of the same foot. This cycle is divided into two phases: 1) the stance phase, where the foot remains in contact with the ground; and 2) the swing phase, where the limb moves forward in a pendulum-like motion[113, 114]. Gait cycles are inherently variable, differing

both within an individual and among different individuals due to factors such as age, sex, and mood, among others[114].

The collection of gait cycle data is commonly done using wearable sensors, like smartwatches, which can capture the nuanced patterns in walking behaviour. These devices have been used extensively in health studies, such as the UK Biobank[55] and NHANES[115], due to their ability to monitor participants in their natural environments over extended periods. To train walking recognition models capable of generalising gait patterns, data can be collected from participants wearing smart watches, while monitored to record when they were or were not walking. Such study data is typically collected from relatively young, healthy volunteers, willing to take part in such studies. When generalising walking in these populations, walking recognition models have quite high accuracy, relying on rhythmic, characteristic heel strikes and arm swings[116, 117].

These generalised patterns observed in walking in healthy populations have noticeable differences when observed in individuals with PD. PD is a progressive neurological disorder characterised by both non-motor and motor symptoms, which include tremors, rigidity, reduced arm swing during walking and bradykinesia (slowness of walking)[9, 13]. The severity of motor impairment experienced by individuals with PD is scaled by the International Parkinson and Movement Disorder Society - Unified Parkinson's Disease Rating Scale (MDS-UPDRS)[14] Part III, in which patients are made to carry out a range of motor tasks, and scored by an expert.

The purpose of this study was to investigate the performance of a walking recognition model trained on a young healthy population when evaluated on a PD population. Additionally, I explored how the model's performance was influenced by increasing motor impairment severity within the PD group, as measured by the MDS-UPDRS Part III. Furthermore, I examined the effectiveness of a model trained on both healthy and PD populations when tested across both groups. This approach aimed to answer the research question: Is it possible to develop a single walking classifier that reliably detects walking in both PD and healthy populations?

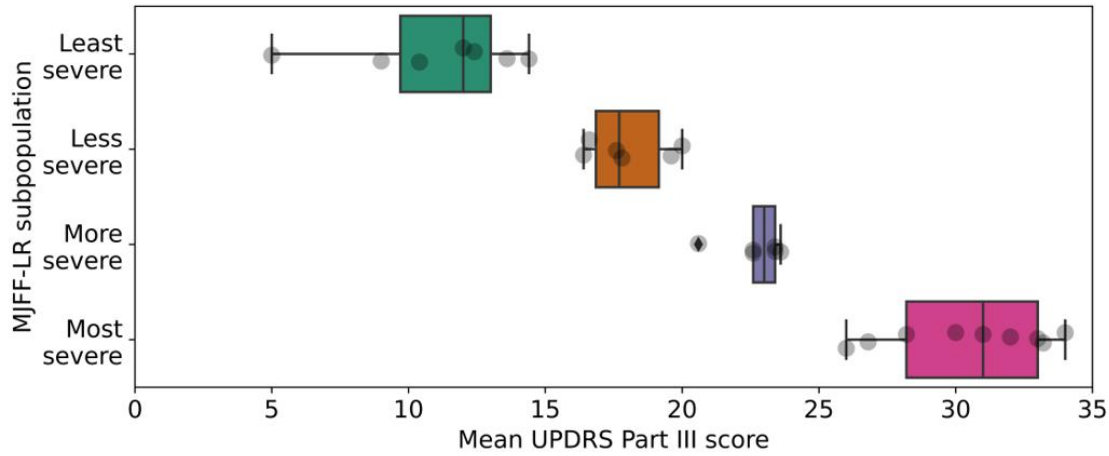


Figure 3.10: Grouping of MJFF Levodopa Response participants by severity of PD indicated by mean UPDRS Part III score.

3.3.2 Methods

This study relies on existing annotated wrist-worn accelerometer data collected in two separate studies. The data from these studies are collected and prepared for the training and evaluation of the walking recognition models. These studies are: the OxWalk study[79], and the MJFF Levodopa Response (MJFF-LR) study[83].

The OxWalk dataset, included 39 healthy individuals who wore an AX3 Axivity sensor operating at 100Hz. Each participant was monitored for approximately one hour in free-living conditions. The mean age of participants in this dataset was 38.5 years. Further information on this dataset is provided in Section 2.5[79].

The MJFF-LR dataset consisted of 28 individuals with a self-reported PD diagnosis. These participants wore a GENEActiv sensor, among other devices, operating at 50Hz and were monitored for approximately two hours over four days. The data collection took place both in and out of the lab, with participants being monitored both on and off their medication. The mean age of participants in this dataset was 62.5 years. To investigate patterns with increased severity of PD in this population, study participants were grouped based on their mean MDS-UPDRS Part III scores. These groupings are found in Figure 3.10. Further information on this dataset is provided in Section 2.3[83]. A summary comparison of both datasets can be found in Section 2.1.

Accelerometry data collected from both studies are put into the pre-processing pipeline, to ensure harmonisation between the two datasets. First, the data for each participant is read and converted into units of gravity ($g = 9.81\text{ms}^{-2}$). Then, the data is calibrated, to ensure that during idle periods, the magnitude of acceleration is $1g$. Next, the frequency of the signal was downsampled to 30Hz using linear downsampling. Finally, the data is divided into 10-second windows of behaviour.

Obtaining the ground truth walking label for the 10-second window depended on the study. For the OxWalk dataset, ground truth annotations are provided for each step. Using this data, walking was defined as walking at least 4 steps in 10 seconds, following the convention used in Small et al. (2024)[79]. For MJFF-LR, on the other hand, ground truth annotations are provided for the activity code performed at a given time window, which is then converted into a label of walking/not walking according to the Table 2.7. Window annotations are calculated as the most frequent activity label in a given window. Any 10-second windows for which no label was the most frequent label were removed from the study. This walking data includes tasks performed both in-clinic and at home, and when on and off levodopa medication.

The walking classifier used in this study was a SSL ResNet-18 model[80, 91]. We use the model with pre-trained weights from SSL training on the UKBB, which included 100,000 person-days of accelerometry data, to detect arrow of time[80]. The deep layers of the model were frozen, however, the final, dense layer of the model with 2 outputs was then fine-tuned specifically for the walking classification task using the collected datasets. Models were trained using an Adam optimiser[107] for a maximum of 100 epochs, and patience for early stopping of 10 epochs. The learning rate, ranging between $1e-6$ and $1e-3$, and the batch size, either 100, 500 or 1000, were optimised for this classification task using Bayesian optimisation[118] to produce the best F1 score. An inverse weighted cross entropy loss function[119] was used during training, with augmentations applied by randomly rotating/switching axes.

The SSL ResNet-18 walking classifier was first trained only on the OxWalk population. Once training was completed, the optimal weights for the model were then loaded, and the model was deployed on the full set of MJFF-LR data. Though

the OxWalk and MJFF-LR studies used different wrist-worn accelerometers, the data from these devices were treated as interchangeable for the purposes of this study.

For the first part of this analysis, I sought to investigate how a walking recognition model trained only on the OxWalk population would perform on the MJFF-LR population. Twenty percent of the OxWalk participants were randomly selected validation participants used for early stopping for the training of the model on the remaining participants' data. Once this model was trained, it was evaluated against the MJFF-LR data, with F1 score reported per participant. Participants within MJFF-LR were grouped based on severity of motor impairment, offering the ability to compare how this related to the performance of the healthy trained walking classifier.

For the second part of this analysis, I sought to discover how the generalisability of walking classifier model was affected by the training data. To do so, I evaluated how the model when trained on either OxWalk, MJFF-LR or both datasets performed when evaluated on OxWalk, MJFF-LR or both. Unlike in the previous walking classification analysis, in some cases for this analysis, there was an overlap between study data used for training and evaluation. In these cases, to avoid leakages between training and evaluation, a 5-fold stratified group cross-validation was used. For each fold, data is split into training, validation and testing, with the proportions 64%, 16% and 20% of participants, respectively. The performance of the model trained in each fold was evaluated for the respective test participants. The participant scores are then aggregated over the 5 folds. I then compared the performance of the 3 trained models within the OxWalk and MJFF-LR populations. F1 scores produced by these 3 models are then compared for significant differences in performance, using paired t-tests, with significance considered when p-values fall below 0.05.

3.3.3 Results

Wrist-worn accelerometer data was used from the 67 participants across the two studies. This data was divided into 10-second windows with a corresponding annotation of walking/not-walking. In total, up to 32,099 windows were used in this analysis, corresponding to roughly 89 hours of data. This is roughly 38 hours of

data taken from OxWalk study, and 51 hours of data from MJFF-LR. The OxWalk dataset contains a total of 11.5 hours of labelled walking 10-second windows. The MJFF-LR dataset, however, contains fewer labelled walking windows compared to OxWalk, totalling 10.5 hours. Greater details on the breakdown of walking frames detected within these populations are outlined in Table 3.4.

Within the search space specified, the optimal parameters were a batch size of 1000 and a learning rate of 6.20×10^{-4} . The performance of this model, deployed on participants with differing severity of PD is displayed in Figure 3.11.

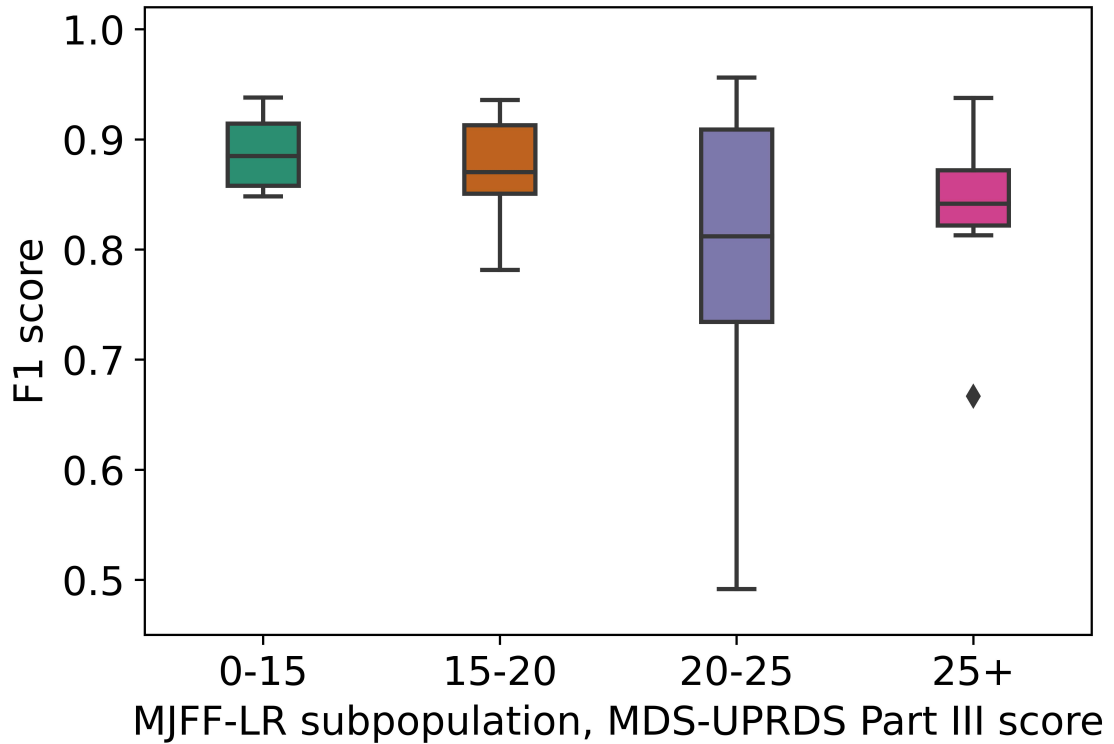


Figure 3.11: Performance of walking classification models trained on the OxWalk population evaluated on MJFF Levodopa Response study participants, grouped by mean Unified Parkinson’s Disease Rating Scale scores Part III for increased severity of motor impairment due to Parkinson’s disease.

The healthy only trained walking classifier trended towards better performance in the subpopulation of the MJFF-LR study that had lowest UPDRS Part III scores, under 15. This population had a mean ($\pm$ standard deviation) F1 score for walking classification of 0.89 ± 0.04 as compared to 0.84 ± 0.08 in the subpopulation with the highest severity, with mean UPDRS Part III scores over 25. Performance was

	OxWalk	MJFF-LR	Both
Number of participants	39	28	67
Age (mean, range) [years]	38.5 (19 - 81)	62.5 (46 - 80)	48.5 (19 - 81)
Sex	19 Female, 20 Male	9 Female, 19 Male	28 Female, 39 Male
Total walking frames duration [hours]	11.5	10.5	22.0
Total non-walking frames duration [hours]	26.3	40.8	67.1
Total frames duration [hours]	37.8	51.4	89.2
Average duration of walking frames per participant [minutes] (range)	17.7 (1.2 - 55.0)	22.5 (8.5 - 34.0)	19.7 (1.2 - 55.0)
Average percentage walking incidence per participant [%] (range)	31.0 (1.9 - 97.0)	26.0 (14.1 - 47.5)	28.9 (1.9 - 97.0)

Table 3.4: Summary of walking and non-walking frames in OxWalk and MJFF Levodopa Response datasets.

worst in a participant in the more severe category of severity (mean UPDRS Part III score between 20 and 25), with an F1 score 0.49.

Following this, a series of walking classifiers were trained and evaluated on each combination of study data used in this work. The performance of these different models, given the combination of training and test participant data is displayed in Table 3.5. This is also displayed in Figure 3.12.

Overall, the performance of the walking classifiers trained was good, with an F1 score of over 0.85 in all classifiers when evaluated on both datasets. Unlike the model trained only on the OxWalk dataset, classifiers trained on MJFF-LR and both datasets did not see less reliable performance, indicated by lower F1 scores in subpopulations of the MJFF-LR study with a higher severity of PD.

Furthermore, the walking classifier model trained on both datasets had the best performance compared to the OxWalk only and MJFF-LR only trained models when deployed on both datasets, with an average F1 score of 0.91 ± 0.09 compared to 0.86 ± 0.10 and 0.90 ± 0.12 , respectively. A paired t-test was used to determine the significance ($p < 0.05$) of the difference in performance between the models, as shown in Table 3.6. The model trained on both datasets showed significant improvements over training on one dataset and testing on the other. In addition, the performance did not significantly change when evaluated on the OxWalk nor higher severity subpopulations of the MJFF-LR study, when comparing the model trained on both datasets and only the respective dataset.

3.3.4 Discussion

This research investigated the performance of walking recognition models when trained and tested on a combination of healthy and diagnosed PD populations. A walking recognition model trained solely on data collected from healthy participants exhibited worse classification performance in individuals with higher motor impairment due to PD. In contrast, training the model on data from both healthy and diagnosed PD participants resulted in a model with significantly improved and more generalisable performance across both populations, as compared to a model

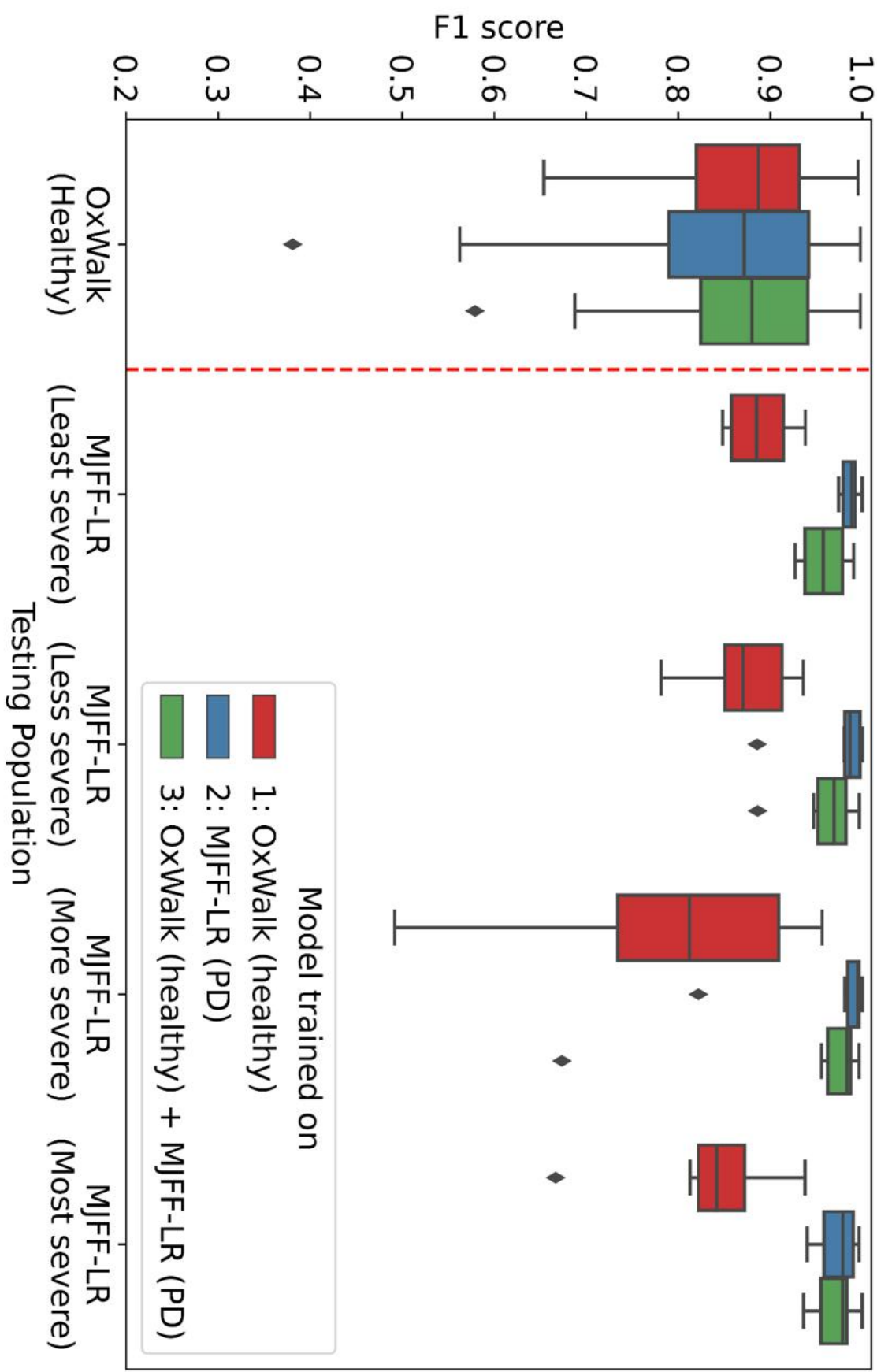


Figure 3.12: Performance of walking classification models given an increased severity of Parkinson’s disease.

Model tested on: $\rightarrow$	OxWalk	MJFF-LR				Both Datasets
Model trained on: $\downarrow$		Least severe (0-15)	Less severe (15-20)	More severe (20-25)	Most severe (25+)	All
1: OxWalk	0.868 ± 0.094	0.888 ± 0.035	0.872 ± 0.056	0.787 ± 0.170	0.839 ± 0.076	0.847 ± 0.097
2: MJFF-LR	0.845 ± 0.131	0.987 ± 0.010	0.973 ± 0.044	0.965 ± 0.070	0.974 ± 0.022	0.975 ± 0.039
3: OxWalk + MJFF-LR	0.872 ± 0.093	0.958 ± 0.025	0.959 ± 0.040	0.930 ± 0.127	0.971 ± 0.021	0.957 ± 0.061
						0.859 ± 0.095
						0.899 ± 0.121
						0.907 ± 0.091

Table 3.5: F1 score for the classification of walking using different models training populations, tested on OxWalk and MJFF-LR datasets. Performances are reported using mean $\pm$ standard deviation per participant. The MJFF-LR test population are sub-divided according to their mean Unified Parkinson’s Disease Rating Scale Part III score.

Test population: $\rightarrow$		OxWalk	MJFF-LR					Both Datasets
Models to compare $\downarrow$			Least severe	Less severe	More severe	Most severe	All	
1: OxWalk	2: MJFF-LR	0.121	6.592e-05****	1.114e-03**	1.382e-02*	3.449e-04****	6.458e-10****	5.289e-03**
1: OxWalk	3: OxWalk + MJFF-LR	0.596	1.341e-03**	2.866e-03**	1.307e-02*	5.382e-04****	3.644e-09****	2.214e-06****
2: MJFF-LR	3: OxWalk + MJFF-LR	2.891e-02*	1.543e-02*	0.087	0.191	0.730	5.667e-03**	0.305

Table 3.6: p-values from the statistical comparison of models across different train and test populations. p-values are reported from a paired t-test. * denotes p-values <0.05 , ** <0.01 , *** <0.001 , indicating significance.

trained only on healthy populations. Notably, this model did not exhibit a trend of declining performance with increasing severity of PD.

PD is associated with numerous motor changes that may confound walking detection algorithms trained exclusively on data from young, healthy populations. For instance, the high-frequency resting tremors characteristic of PD could be misclassified as walking by the OxWalk-trained model, as these tremors may resemble running in the dataset. Moreover, individuals in the early stages of PD often exhibit reduced arm swing, which might not be adequately captured by the OxWalk-trained algorithm, leading to further inaccuracies in walking classification. These issues become more pronounced with higher levels of motor impairment, as indicated by the MDS-UPDRS Part III score [9, 120].

Incorporating data from both healthy and PD participants into the training process expands the range of walking patterns the model can learn, thereby enhancing its ability to accurately detect walking across diverse populations. Interestingly, the inclusion of PD participant data also led to a moderate improvement in the model's performance when classifying walking in healthy individuals, from a mean F1 of 0.868 to 0.872. However, this improvement was not statistically significant ($p\text{-value} \approx 0.6 > 0.05$).

This study has several strengths. Data was collected from a diverse range of sources, and efforts were made to harmonise different devices and participant characteristics. Additionally, the use of publicly available datasets and GitHub code enhances the study's transparency and reproducibility. The hypothesis-free models employed in this study performed well overall, demonstrating the potential for such models in walking classification tasks. These strengths allowed for the production of highly accurate walking classification models.

However, the study also has limitations. The datasets included participants of varying ages and sexes, which could have influenced the model's performance in addition to PD severity. This was not explored in this study due to the limited representation of some of these minority categories. Additionally, the presence of other impairments in PD participants may have affected walking patterns,

complicating the classification task, such as walking aids. The overall incidence of walking in the training data was higher than what is typically expected in daily living, suggesting that the models may require careful handling when applied to 24-hour monitoring data. Furthermore, the MJFF-LF study data were collected during monitored tasks, often in clinical settings, which may not fully reflect the range of activities and walking behaviours observed in free-living, unprompted scenarios. Finally, the study would have benefited from further external validation, though the availability of suitable public datasets was limited at the time of the research.

3.3.5 Summary

This study underscores the potential of wrist-worn accelerometers for the detection of walking behaviour in PD populations. The results demonstrate that walking classifiers trained on data from both healthy and PD populations can offer better generalisability and robustness across different severities of PD, shown by the significant improvement in performance of this model over the healthy-only trained model, when deployed on individuals with PD. These findings suggest that fine-tuning activity models with diverse data sources can significantly improve their applicability in real-world settings.

4

Investigating the association between daily steps counts and incident Parkinson's disease

Contents

4.1	Introduction	74
4.2	Methods	75
4.2.1	Study Cohort	75
4.2.2	Accelerometer-derived physical activity processing	75
4.2.3	Ascertainment of Parkinson's disease	76
4.2.4	Statistical Analysis	76
4.3	Results	80
4.4	Discussion	83
4.5	Summary	86

The materials for this chapter have been adapted from work made available online under the following citation: **Aidan Acquah**, Dr. Andrew Creagh, Dr. Valentin Hamy, Alaina Shreves, Charilaos Zisou, Dr. Charlie Harper, Dr. Stefan Van Duijvenboden, Prof. Chrystalina Antoniadou, Prof. Derrick Bennett, Prof. David Clifton, Prof. Aiden Doherty, "Daily steps are a predictor of, but perhaps not a modifiable risk factor for Parkinson's disease: findings from the UK Biobank"[121]. I conducted all the experiments myself and produced all the results presented in this chapter.

4.1 Introduction

Various studies that were explored in Chapter 1 of this thesis highlighted the potential use of wearables to identify modifiable risk factors for Parkinson’s disease (PD). This Chapter focuses on whether median daily steps, a proxy for physical activity, may be a modifiable risk factor for PD, observed in the UK Biobank Physical Activity Monitoring population.

In existing literature, low self-reported physical activity has been associated with a higher risk of incident PD[29, 122]. Despite these observations, uncertainties remain, stemming from a reliance on self-report measurements of physical activity, which are crude and less reliable, creating greater uncertainties in observed associations[123, 124]. In addition, PD’s progressive and prolonged development poses challenges in distinguishing between incident and prevalent cases. This ambiguity can introduce reverse-causation bias, where the observed associations may be influenced by the change in behaviour of individuals who already have underlying PD, rather than vice-versa. One common approach to mitigate potential reverse causation is to exclude the first few years of follow-up data. In existing PD studies, there has been inconsistency around the number of years removed across studies, ranging between two to ten years[29, 59]. However reduced activity levels were still significantly associated with incident PD after this exclusion[29, 59].

Daily step counts serve as a valuable proxy for measuring physical activity levels. One of the primary advantages of using average daily steps over other physical activity metrics is their simplicity, allowing for clear communication and understanding among the general population. This ease of interpretation helps translate research findings into effective public health messaging[125]. Existing research into the association between physical activity and incident PD have typically used metabolic equivalent of task (MET) scores to estimate physical activity[29]. However, the increasing adoption of wearable devices, capable of estimating step counts, has encouraged research on their accuracy and reliability, particularly in populations diagnosed with PD[126, 127]. The translatable nature of daily step

counts encourages the expansion of this research into incident PD, allowing for results to offer insight into public health recommendations for potential disease prevention.

I, therefore, aimed to investigate the association between accelerometer-measured daily step counts and incident Parkinson's disease, and how this association changed with latter windows of follow-up.

4.2 Methods

4.2.1 Study Cohort

The UK Biobank is a prospective cohort study of 502,536 adults in the United Kingdom who were recruited between 2006 and 2010. The consented participants completed a questionnaire, interview, and provided physical measurements, as well as blood, urine, and saliva samples[64, 68]. A subgroup of participants consented to wear an Axivity AX3 wrist-worn accelerometer for up to 7 days between 2013-2015[55]. All participants provided written informed consent and the study was approved by the National Information Governance Board for Health and Social Care and the National Health Service North West Multicentre Research Ethics Committee (06/MRE08/65).

4.2.2 Accelerometer-derived physical activity processing

Accelerometer data was processed to derive step counts using the OxWearables `pypi:stepcount` package. This model is a hybrid self-supervised machine learning model that was trained on ground truth free-living data and counts peaks in detected walking windows of behaviour[79]. An illustration of this hybrid model is found below in Figure 4.1.

Steps were predicted for each participant's full period of collected data, with missing periods due to non-wear imputed by averaging the step counts in the corresponding times across valid days, as in previous studies[55, 79, 88]. The median of the imputed daily step counts, across the period of monitoring, was then calculated as the primary exposure in this study.

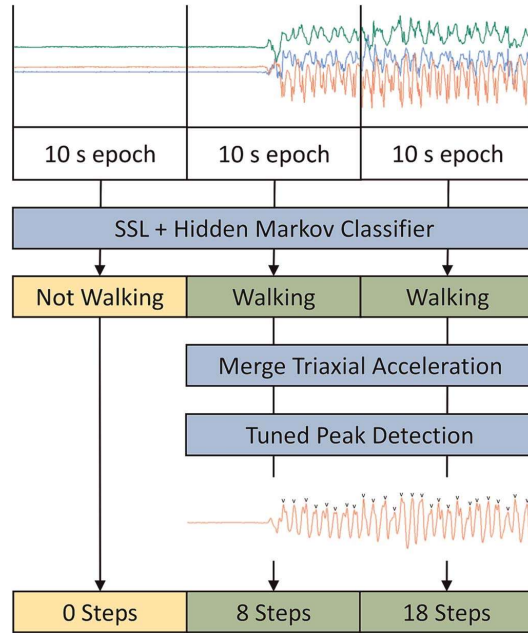


Figure 4.1: Illustration of hybrid stepcount algorithm taken from Small et al. (2024)[79].

4.2.3 Ascertainment of Parkinson’s disease

The diagnosis of PD was obtained from a combination of the UK Biobank’s linked hospital data, inpatient hospital admissions data, and death data. Diagnoses were coded according to the International Classification of Diseases (version 10) coding system, with outcomes recorded as the first instance of a G20 diagnostic code[71]. This follows the approach recommended by the UK Biobank[71] and used by other studies in studying PD in the physical activity monitoring cohort[58, 59, 60]. Incident PD however was defined at their first diagnosis of PD. Participants were censored at the first date of PD diagnosis, date of death, or at the end of their follow-up period (31 October 2022 for England, 31 August 2022 for Scotland, and 31 May 2022 for Wales), whichever came first[66].

4.2.4 Statistical Analysis

I processed the accelerometer data from 103,614 participants. I excluded participants with device calibration or data reading errors ($>1\%$ of values outside $\pm 8g$ range), inadequate wear time (<3 days, or non-wear at a particular time of day for all days of monitoring), unreasonably high average acceleration (>100 mg), and stepcount

model execution failure[50, 58, 79, 128]. I also excluded all 150 prevalent PD and 22 other Parkinsonism cases, diagnosed before accelerometer wear, and participants with missing healthcare linkages or covariate data. The final analysis included 94,696 participants (Figure 4.2).

A Cox proportional hazard regression model[129] was used to estimate adjusted hazard ratios (HRs) with 95% confidence intervals (CIs) for the association between daily step counts and incident PD. The association was first evaluated using quintiles of median daily step count, allowing for interpreting trends in the association and identifying any outliers. CIs were calculated using group-specific log HR variances to enable comparisons across all groups, presented as floating absolute risks[130].

I then evaluated how the risk of PD was associated with each 1,000 steps per day higher. I sequentially applied the full set of adjustments to the multivariable model, by progressively incorporating the full set of covariates. The impact of these sequential adjustments was assessed using a chi-squared test to see how the significance of the median daily steps term changed as each covariate is added. Specifically, for the per-1000 median daily steps with one degree of freedom, I compared the chi-squared values before and after adjustment. A likelihood ratio test comparing quintiles of daily steps as categorical and ordinal for the full multivariable model was performed to test that the association between daily steps and incident PD did not significantly depart from linearity[131]. This allows for the reporting of hazard ratios per 1,000 steps rather than quintiles of step count. At each step of this sequential adjustment, confounding was assessed by observing the chi-squared from a likelihood ratio test of the association.

The models used age as the time scale and were first minimally adjusted for season of wear (spring, summer, autumn, winter) and sex (female, male). Multivariable models further expanded the minimally adjusted by also adjusting for ethnicity (white, non-white), Townsend Deprivation Index (TDI, stratified by quintiles in UK population[72]), geographic region of assessment centre (London, Midlands, Yorkshire, Northeast, Northwest, Southeast, Southwest England, Scotland, Wales), educational attainment (degree, diploma, A/AS levels, professional qualification,

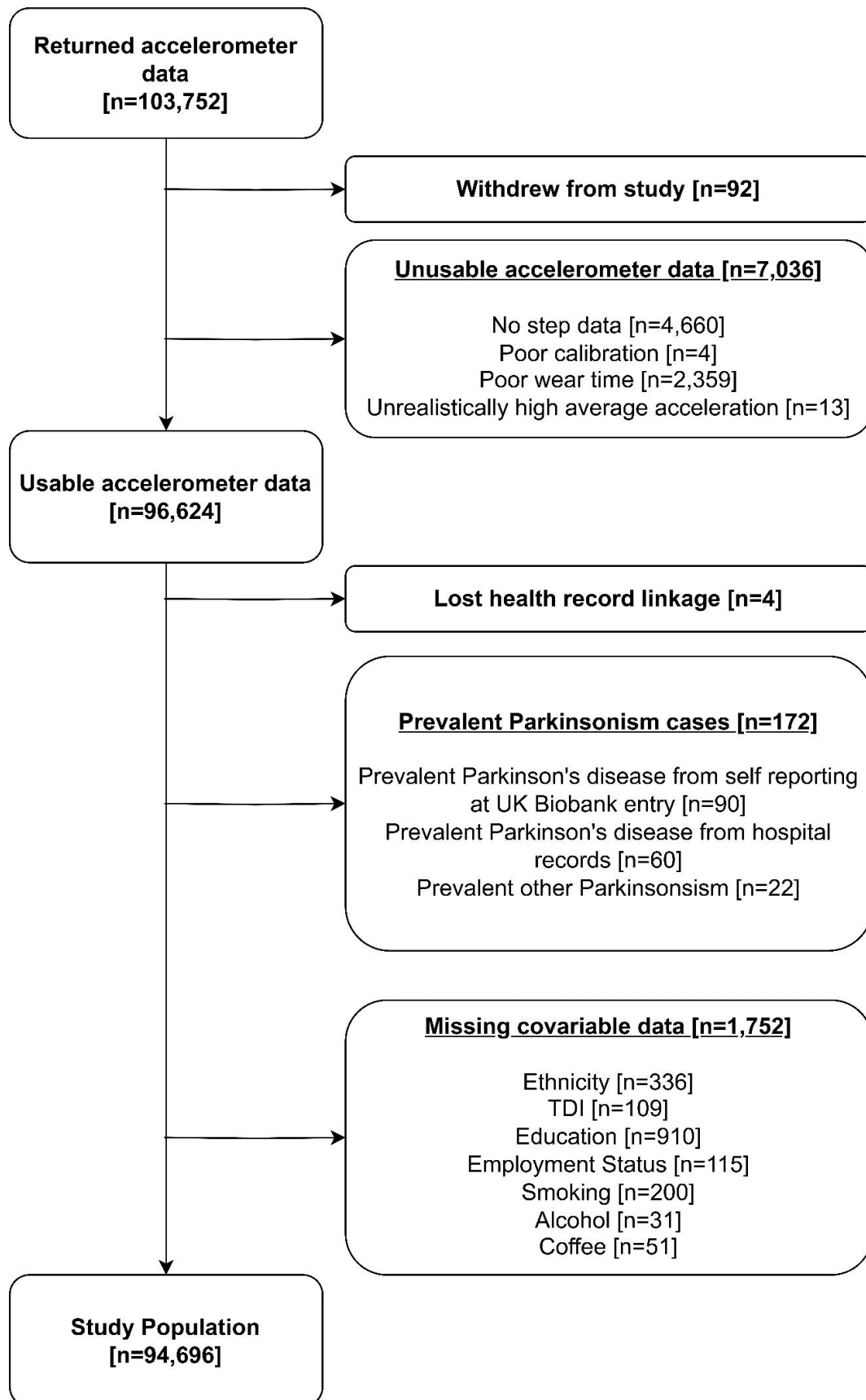


Figure 4.2: Exclusion flow diagram for steps PD association in UK Biobank.

GCSE/O levels, no degree/qualification), employment status (employed, not employed), smoking status (never, previous, current) and alcohol consumption (never, <3 times per week, ≥ 3 times per week) and coffee intake (non-drinker, ≤ 2 cups a day, >2 cups a day). All covariate data were provided by participants at the UK Biobank baseline assessment visit, with details provided in supplementary material. The proportional hazards assumption was tested using Schoenfeld residuals[132] and no violations were observed.

I conducted several additional analyses using the models assessing the risk for an increase in 1,000 daily steps. First, I expanded the multivariable-adjusted models by additionally adjusting for body mass index (BMI) at baseline, depression, type 2 diabetes, constipation, bladder dysfunction derived through a combination of self-report and health records, and sleep duration derived from accelerometer data. I then explored whether the observed associations differed within subpopulations of interest, including by age groups, sex, BMI (<25, 25–30, 30+ kg/m²)[133] and history of depression. Additionally, I performed an analysis examining the associations after excluding participants with any history of neurological disorder.

To address concerns regarding reverse causation bias and examine how the association may have changed over successive follow-up periods, I calculated the HR within specific follow-up periods (0-2 years, 2-4 years, 4-6 years, and 6+ years). Using Lexis expansion[134, 135], I expand the original study population into 4 subpopulations. Within each of the four follow-up periods, the respective subpopulation is chosen from the full study population censored before and after the corresponding follow-up period, as seen in Floud et al. (2020)[136]. To censor before the follow-up period, participants were removed from the study subpopulation if they had a follow-up time less than the left bound of the period. This was typically either due to death, a diagnosis of PD, or a loss to follow-up that occurred earlier than the left bound of the respective follow-up period. To censor after the follow-up period, the follow-up time is capped to the right bound of the respective period. In addition, any deaths, diagnoses for PD, or losses to follow-up occurring after this right bound are ignored within this subpopulation[134, 135].

A sensitivity analysis was then run, to ensure that the observed associations were not influenced by the limitations of step count algorithm. This was done by repeating the main analysis using a step count algorithm trained on both PD and healthy participants. To do this, the walking detector in the hybrid stepcount model illustrated in Figure 4.1 was replaced with the walking detector trained on the MJFF Levodopa Response and OxWalk datasets, as described in Section 3.3. Using the step counts estimated from this new algorithm, the full analysis pipeline is re-run, and the results are compared to the main analysis.

Results for this analysis were reported according to Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines, attached in Appendix C[137].

4.3 Results

Among the 94,696 participants, 56% were female, 97% identified as White. Most were in the least to second least deprived TDI quintile (< -1.509), compared to the TDI distribution of the general UK population in 2011[72]. Those with step counts in the highest quintile (12,369+ daily steps) were younger and had lower BMI compared to those in the lowest step count group ($< 6,276$ daily steps). Further details of the breakdown of the baseline characteristics of the study population can be found in Table 4.1

Participants had a median follow-up time of 7.9 years (IQR: 7.4, 8.4), during which 407 incident PD cases occurred, with a median time to diagnosis of 5.2 years (IQR: 3.1, 6.8). I observed a strong and significant inverse linear association between median daily step quintiles and incident PD using data from the entire duration of follow-up. Those who walked over 12,369 steps daily had a 59% lower risk of PD (HR 0.41; 95% CI: 0.31-0.54), compared to those walking less than 6,276 steps daily (HR 1.00; 95% CI: 0.84-1.19), after multivariable adjustment. There was no evidence of departure from linearity ($p=0.65$).

For the multivariable model, a 1000-step increase in daily step counts was associated with an 8% lower risk of PD (HR 0.92; 95% CI 0.89-0.94). The sequential

Median daily step count	<6,276	6,276-8,145	8,146-9,966	9,967-12,368	12,369+	All
No. of Participants	18,933	18,939	18,941	18,942	18,941	94,696
Demographic factors						
Age, mean (SD), (years)	63.3 (7.96)	62.6 (7.89)	62.3 (7.85)	62.1 (7.74)	61.4 (7.65)	62.3 (7.84)
Female, No. (%)	10,792 (57.0%)	10,916 (57.6%)	10,780 (56.9%)	10,727 (56.6%)	10,158 (53.6%)	53,373 (56.4%)
White ethnicity, No. (%)	18,288 (96.6%)	18,321 (96.7%)	18,377 (97.0%)	18,416 (97.2%)	18,406 (97.2%)	91,808 (97.0%)
Socio-economic factors, No. (%)						
Most deprived TDI quintile	2,242 (11.8%)	1,711 (9.0%)	1,720 (9.1%)	1,759 (9.3%)	1,881 (9.9%)	9,313 (9.8%)
No degree/qualifications	2,157 (11.4%)	1,643 (8.7%)	1,489 (7.9%)	1,305 (6.9%)	1,270 (6.7%)	7,864 (8.3%)
Not employed	8,418 (44.5%)	7,271 (38.4%)	7,133 (37.7%)	6,793 (35.9%)	6,498 (34.3%)	36,113 (38.1%)
Recruitment Centre						
England	16,979 (89.7%)	16,997 (89.7%)	17,079 (90.2%)	16,961 (89.5%)	16,950 (89.5%)	84,966 (89.7%)
Scotland	1,147 (6.1%)	1,220 (6.4%)	1,170 (6.2%)	1,270 (6.7%)	1,357 (7.2%)	6,164 (6.5%)
Wales	807 (4.3%)	722 (3.8%)	692 (3.7%)	711 (3.8%)	634 (3.3%)	3,566 (3.8%)
Behavioural factors						
Alcohol 3+ times/week, No. (%)	7,732 (40.8%)	8,918 (47.1%)	9,506 (50.2%)	9,988 (52.7%)	10,234 (54.0%)	46,378 (49.0%)
Never smoker, No. (%)	9,971 (52.7%)	10,827 (57.2%)	11,025 (58.2%)	11,214 (59.2%)	11,057 (58.4%)	54,094 (57.1%)
Non coffee drinker, No. (%)	4,168 (22.0%)	3,751 (19.8%)	3,611 (19.1%)	3,555 (18.8%)	3,516 (18.6%)	18,601 (19.6%)
Sleep Duration, mean (SD), (hours)	9.28 (1.53)	8.97 (1.22)	8.82 (1.13)	8.69 (1.09)	8.50 (1.04)	8.85 (1.24)
BMI, mean (SD), (kg/m²)	28.6 (5.58)	27.1 (4.45)	26.5 (4.16)	25.9 (3.88)	25.5 (3.68)	26.7 (4.53)
Prior disease, No. (%)						
Type 2 Diabetes	1,533 (8.1%)	857 (4.5%)	611 (3.2%)	535 (2.8%)	474 (2.5%)	4,010 (4.2%)
Depression	7,376 (39.0%)	6,534 (34.5%)	6,217 (32.8%)	5,928 (31.3%)	5,487 (29.0%)	31,542 (33.3%)
Constipation	1,020 (5.4%)	707 (3.7%)	652 (3.4%)	627 (3.3%)	548 (2.9%)	3,554 (3.8%)
Bladder Dysfunction	177 (0.9%)	122 (0.6%)	115 (0.6%)	89 (0.5%)	80 (0.4%)	583 (0.6%)
Season of device wear, No. (%)						
Spring	4,113 (21.7%)	4,160 (22.0%)	4,432 (23.4%)	4,413 (23.3%)	4,604 (24.3%)	21,722 (22.9%)
Summer	4,334 (22.9%)	4,536 (24.0%)	4,989 (26.3%)	5,378 (28.4%)	5,897 (31.1%)	25,134 (26.5%)
Autumn	5,675 (30.0%)	5,830 (30.8%)	5,663 (29.9%)	5,459 (28.8%)	5,305 (28.0%)	27,932 (29.5%)
Winter	4,811 (25.4%)	4,413 (23.3%)	3,857 (20.4%)	3,692 (19.5%)	3,135 (16.6%)	19,908 (21.0%)

Table 4.1: UK Biobank Physical activity monitoring study population characteristics. Most deprived TDI quintile from 2011 census (≥ 2.516)[72]. TDI=Townsend deprivation index, BMI=body mass index, SD=standard deviation.

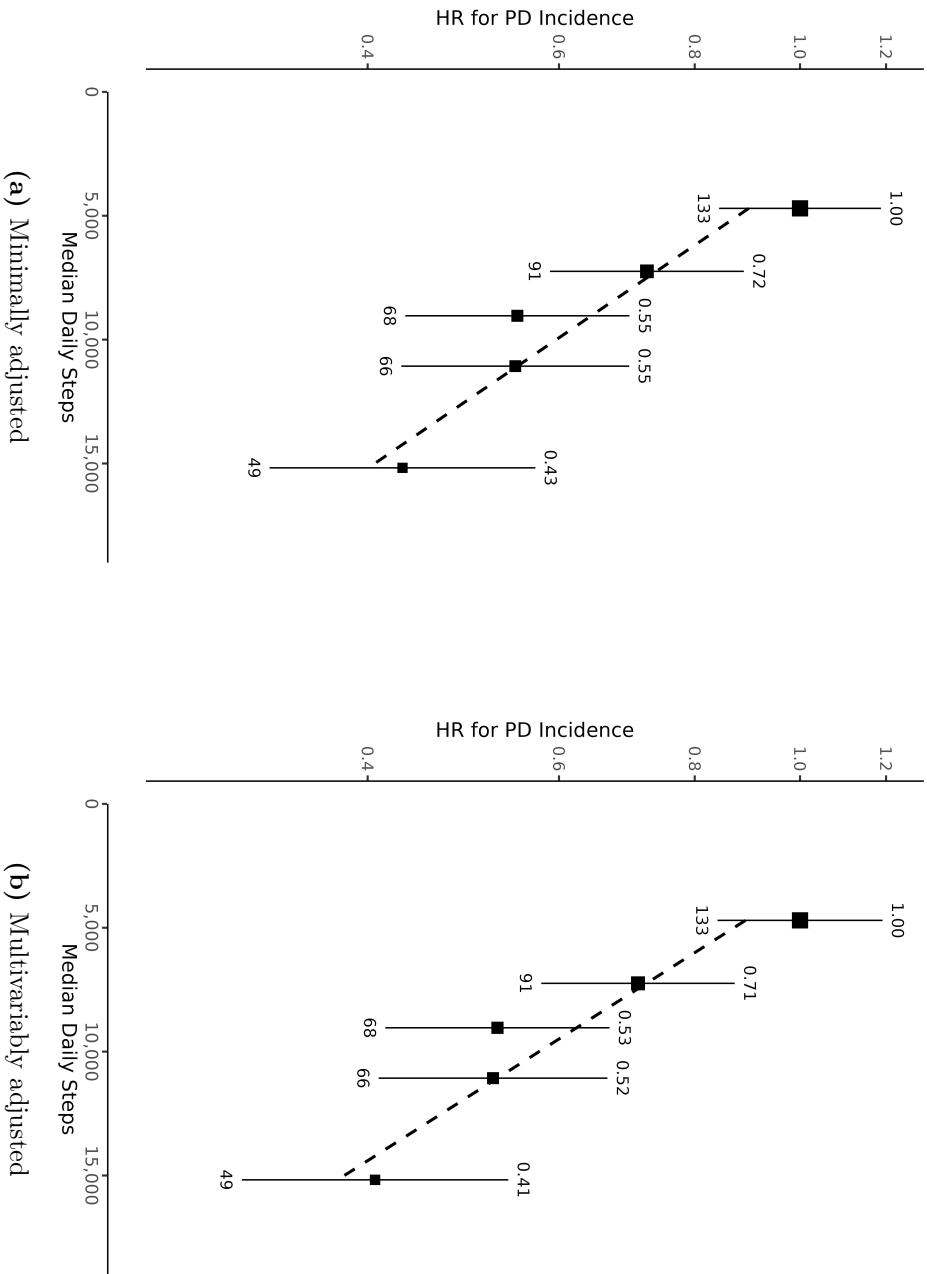


Figure 4.3: Association between quintiles of median daily step count and incident Parkinson's disease. Hazard ratios (HR) and 95% confidence intervals were calculated using age as a timescale. HR is above and the number of events is plotted below each data point. a) Minimally adjustment model, accounting only for season of wear and sex. b) Multivariably adjusted model, adjusted for season of wear, sex, ethnicity, Townsend deprivation index, geographic region, education, employment, alcohol intake, smoking status and coffee intake. The HR (95% CI) per 1,000 increase in daily steps for the multivariably adjusted model is 0.92 (0.89 – 0.94). For each category, the area of the square is inversely proportional to the log HR, which also determines the 95% CI.

adjustment of covariates from the univariable to multivariable model showed minimal changes to the strength of the association, and the chi-squared increased by 33% (shown in the supplementary material). The further adjustment of additional covariates minimally altered the observed associations. However, adjustment for sleep duration did attenuate the strength of association, with a reduction in the chi-squared by 64%, as shown in Figure B.4. Furthermore, I observed no significant difference across subgroups, including age groups, sex, BMI status and history of depression. After removing 9,734 individuals with prevalent neurological cases, the observed association was similar to that observed in the main study sample (HR 0.92; 95% CI 0.89-0.95).

To address potential reverse causality, I assessed how the association between step counts and PD risk may have changed over follow-up periods of <2, 2-4, 4-6 and 6+ (Figure 4.4). I observed the strongest association within the first two years of follow-up (HR 0.83; 95% CI 0.70-0.90), during which there were 55 incident cases. However, the association was attenuated for later windows of follow-up, trending towards the null. This is despite having higher relative numbers of cases in latter periods of follow-up, reducing uncertainty in estimates. For the follow-up period of over 6 years, the HR was no longer statistically significant, with a hazard of 0.96 (95% CI: 0.92 – 1.01).

These results remained consistent in the sensitivity analysis, with the step count algorithm trained on both PD and healthy participants. The outputs for this algorithm showed high correlation with the healthy-only trained step count algorithm, with a Pearson correlation coefficient of 0.966. The use of this algorithm did not significantly alter the observed associations. These results are presented in Appendix B.

4.4 Discussion

Higher accelerometer-derived physical activity levels, indicated by median daily step counts, were associated with a lower risk of incident PD when using data from all periods of follow-up. This association remained statistically significant after

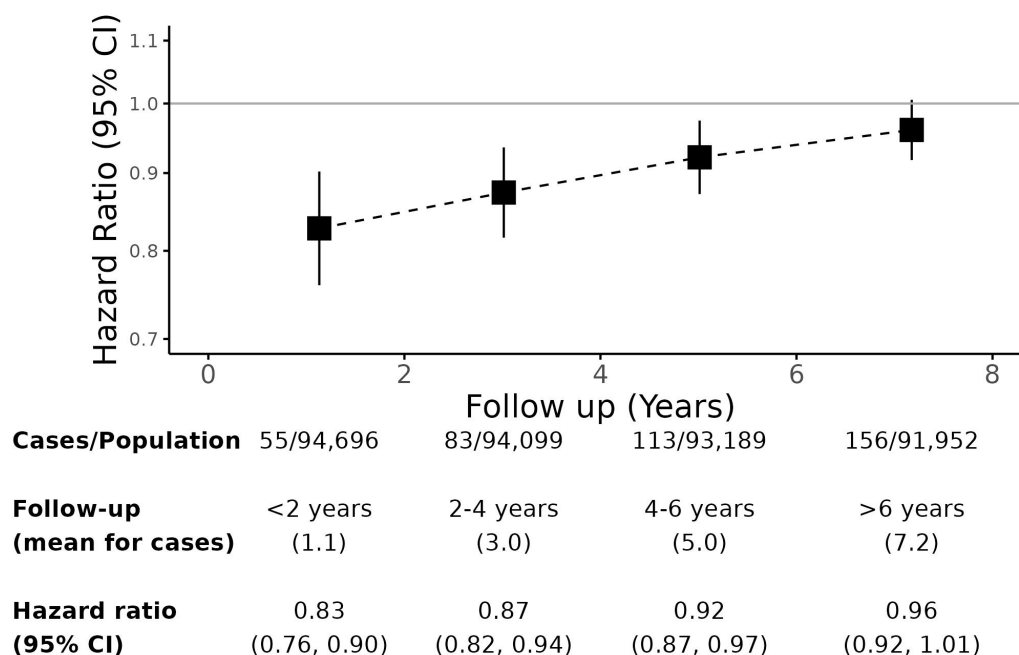


Figure 4.4: Hazard ratio for incident Parkinson's disease by period of follow-up per 1000-step increase in daily steps. Models use age as timescale, and are adjusted for season of wear, sex, Townsend deprivation index, geographic region of assessment centre, educational attainment, employment, smoking, alcohol and coffee consumption. CI=confidence interval.

multivariable adjustment, investigating potential factors, and among subpopulations of interest. However, this association was null when considering later periods of follow-up, indicating that reverse causation bias is likely. Overall, these findings do not provide evidence that daily step counts may be a modifiable risk factor for PD.

These findings are in contrast to some previous literature in the field that supports an association between physical activity levels and incident PD, even after the consideration of reverse causation[29, 59, 60, 138]. In a meta-analysis of this association with a median (range) follow-up of 12 (6.1-22.0) years, the relative risk for incident PD comparing the highest and lowest categories of total physical activity was 0.79 (95% CI 0.68-0.91), and remained significant when the first four to ten years of follow-up was removed from the constituent studies (0.76; 95% CI 0.64-0.91)[29]. In this study, removing years of follow-up reduced the strength of the observed association. Similar attenuations between physical activity and diseases of ageing have been reported in other studies. These findings are more consistent with

results from a study on physical activity and dementia that was conducted in the Whitehall Study and Million Women Study[136, 139]. These results however support existing literature in finding lower physical activity levels to be a strong predictor for late-stage prodromal PD/early-stage clinical PD among those who already have a PD diagnosis[28, 138]. One study reported that this reduction in activity may occur up to 4 years before a clinical diagnosis[138]. A separate study assessing minutes of daily physical activity[28] found a reduction of 29% in comparing clinic PD patients to controls, in minutes of daily physical activity. There are several hypotheses for these activity changes, including changes in overall disability, poorer walking performance, and fear of falls/anxiety/depression[28]. It should be noted, however, that this reduction in activity is not fully explained by known factors, with this study encouraging further research into explaining this behaviour[28].

This study has several strengths. First, objective measures of activity derived from accelerometer data are used, which offers greater reliability in the measurement of this exposure compared to questionnaire-based studies. Second, the use of step count as a measure of daily physical activity offers the opportunity for better communication as step counts are easily translatable for public health messaging. Next, this work has built on previous studies investigating PD using the UK Biobank physical activity monitoring study[58, 59, 60], however with longer follow-up, while also addressing some of these studies' limitations. Of these limitations, the association was explored to understand how it may change over follow-up periods to better describe the potential impact of reverse causation.

There are also some limitations to this study. Firstly, this analysis was purely observational; hence one cannot exclude the possibility of unmeasured or residual confounding in the observed associations, despite considering a comprehensive set of confounders. Secondly, the UK Biobank provides limited health data linkage for participants, which is used to identify participants with PD. I use the approach recommended by UK Biobank to ascertain PD, combining hospital inpatient, death register and self-reported data[71]. However, this does not provide information regarding the severity or subtype of PD, which may offer further insights into

this association. Furthermore, the use of primary care data could bolster the PD cases, but this data is currently limited to just 45% of participants and only up to 2017. Lastly, this study is limited by the number of cases observed within the period of follow-up. With 407 incident cases of PD from all periods of follow-up, and just 156 diagnosed at least 6 years after participating in the study, this study was somewhat limited in power to observe small effects of increased activity. It was also limited by a maximum follow-up time of 9.4 years, which remains within the estimated prodromal window of PD of over 10 years. Future studies in this population can benefit from higher case numbers from a longer follow-up period, giving further insights into this association.

4.5 Summary

Results from this prospective study add to the literature on physical activity and incident PD, suggesting that higher daily step counts were associated with a lower risk of incident PD, but only for short follow-up periods. Therefore, the observed association may be explained by reverse causation bias. This work supports the continued labelling of low physical activity as a marker for Parkinson's disease, rather than as a modifiable risk factor leading to Parkinson's disease.

5

Use of active walking windows to categorise Parkinson's disease

Contents

5.1	Introduction	88
5.2	Methods	89
5.2.1	Study population	90
5.2.2	Pre-processing	90
5.2.3	Human activity recognition	91
5.2.4	Parkinson's disease classification using active walking windows	94
5.2.5	Model Evaluation and Explainability	94
5.3	Results	96
5.3.1	Study population	96
5.3.2	Human activity recognition	97
5.3.3	PD Classification	98
5.3.4	Explainability	101
5.4	Discussion	103
5.4.1	Comparison with published literature	103
5.4.2	Strengths and Limitations	105
5.5	Summary	106

5.1 Introduction

Parkinson's disease (PD) is a neurodegenerative condition characterised by both motor and non-motor symptoms, including resting tremors and bradykinesia, which impair daily activities such as walking[9, 140]. Although treatments like levodopa provide symptom relief by replenishing dopamine, they often lead to complications such as dyskinesia and lose effectiveness as the disease progresses, leaving the underlying pathology unaddressed [9, 16, 140]. The growing prevalence of PD and lack of curative treatments has led some researchers to describe it as an impending epidemic, requiring increased research to understand and prevent it [141, 142]. In response, initiatives like the Progressive Parkinson's Markers Initiative (PPMI) focus on improving knowledge of PD's aetiology and progression, aiming to facilitate the development of disease-modifying therapies through the study of participants with diagnosed, prodromal, or no signs of PD [75].

Wearable technologies and digital biomarkers offer promising advancements in PD. Smartwatches equipped with sensors, such as accelerometers, enable long-term, objective data collection, revealing patterns in motor function and disease progression that are difficult to capture with traditional diagnostic methods like the MDS-UPDRS scale[14, 143]. Deep learning techniques, surpassing earlier statistical and machine learning approaches like Euclidean Norm Minus One (ENMO) and random forests, are increasingly used to analyse these data and extract meaningful digital biomarkers, showing potential in disease association[48, 49, 50, 53, 144]. Explainable machine learning (XML) methods further enhance our understanding of the patterns discovered by deep learning, providing insights into PD's complex motor behaviours.

This research aimed to investigate the classification of labels of PD status through the analysis of predicted walking behaviour obtained from free-living accelerometry data. This analysis built on the approach presented by Atri et al. (2022)[53], a proof-of-concept study in which wearable sensors were used to distinguish seven PD patients from four healthy controls by identifying "walk-like" events through activity recognition techniques and training neural networks to classify PD status[53]. This study extended their work by incorporating all eligible participants. Furthermore,

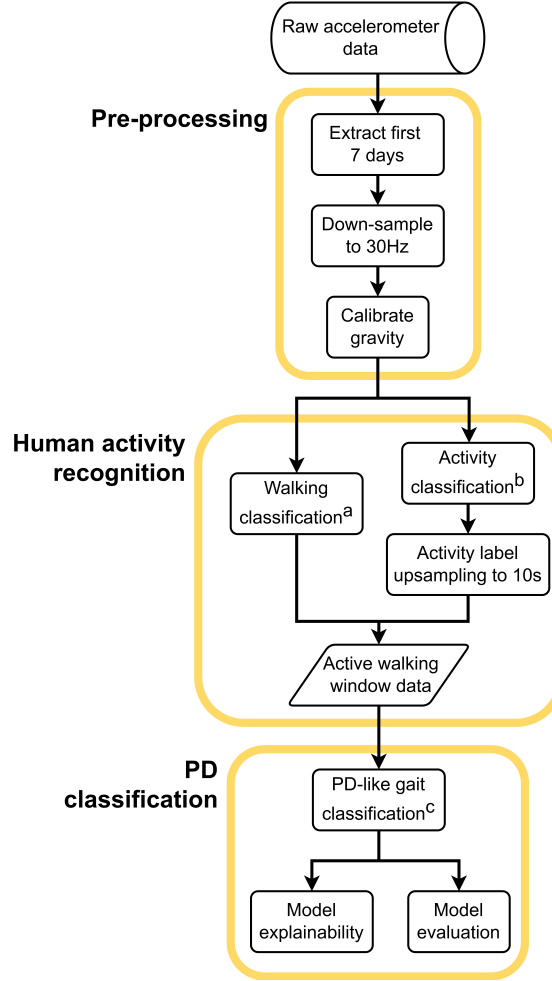


Figure 5.1: The method flowchart for categorising Parkinson's disease status using active walking windows in the PPMI-Verily population. a) Walking classification explored in Section 3.3 using 10-second self-supervised ResNet-18 trained on both OxWalk and MJFF-LR datasets to produce binary label of walking/not walking. b) Activity classification using `pypi:actinet` explored in Section 3.2; a 30-second self-supervised ResNet-18 and hidden Markov model smoothing trained on Capture-24 to produce labels of sleep, sedentary, light or moderate-vigorous activity. c) PD-gait classification using machine learning models taking 10-second windows of accelerometry to predict labels of: diagnosed PD, prodromal PD or healthy control.

through the use of explainable AI techniques, greater insights were gathered from patterns in PD-like walking when compared to healthy controls.

5.2 Methods

The methodology pipeline used in this study is visually depicted in Figure 5.1, providing a comprehensive overview. This is further outlined in the Sections below.

5.2.1 Study population

At the time of this research, the PPMI dataset contained comprehensive information regarding 2,227 participants involved in the study. Among these participants, 354 individuals engaged in the Verily sub-study, during which they used the Verily watch for passive monitoring. These study subjects wore the Verily watch while continuing with their regular activities for periods of up to 23 hours each day, spanning from a few days to two years, with an interquartile range of monitoring between 254 and 560 days[53]. The average and median durations of daily wear were approximately 18 and 21 hours, respectively[53]. Directives were given to charge the device using the Verily StudyHub and transfer data to the cloud for at least one hour daily. The data gathered from the Verily Watch included measurements from triaxial inertial sensors (accelerometer and gyroscope) captured at a frequency of 80Hz[53, 74].

Our analysis primarily focused on participants classified into three distinct categories: healthy control, prodromal PD, or diagnosed PD. The criteria for categorisation were carefully delineated following the guidelines provided in Marek et al. (2018)[75]. To ensure the robustness of this analysis, I excluded participants who were labelled as scans without evidence of dopaminergic deficit or belonged to the Early Imaging cohort. These exclusions were made due to the ambiguity surrounding the PD status of such individuals[145]. Furthermore, participants were excluded if their wear time amounted to fewer than 3 days or if any discrepancies or errors were identified in the timestamps of the stored data. Notably, timestamp errors emerged as a result of attempts made to estimate the unique timestamps associated with each data point attributed to an individual.

5.2.2 Pre-processing

Pre-processing steps were used for achieving data consistency and comparability across different participants, as the raw data obtained from various individuals exhibited discrepancies in duration and calibration. Firstly, only the initial seven days of monitoring data were used for this analysis. This time frame was chosen to ensure similar monitoring periods across all eligible participants and is consistent

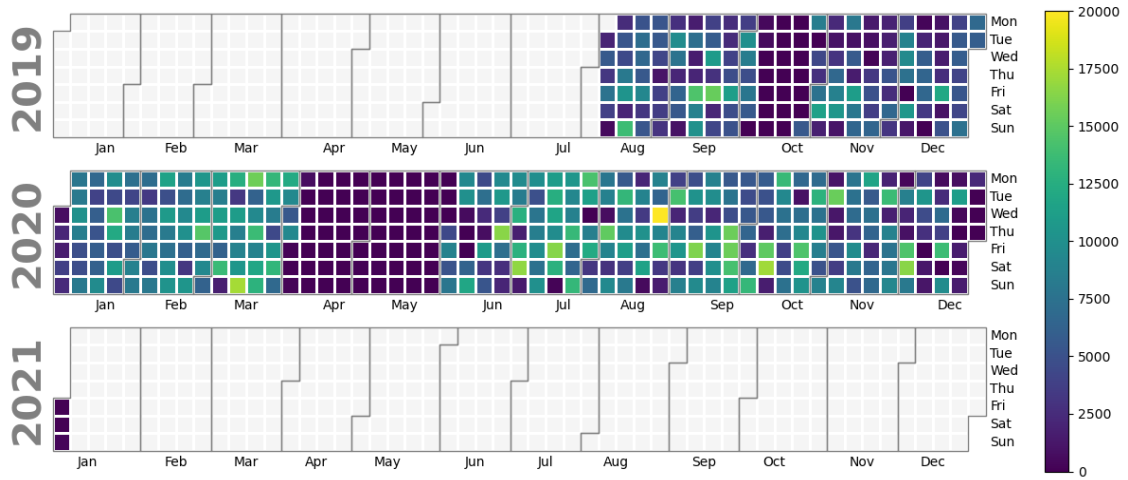


Figure 5.2: Daily step count for example PPMI-Verily participant.

with that of similar studies[146]. In addition, the full monitoring period for many participants coincided with the COVID-19 pandemic, during which activity patterns may have been influenced by lockdown measures. A calendar for daily step count for an example participant is shown in Figure 5.2. Due to limited information on the specific timing and impact of these lockdowns on participants' motor phenotypes, the analysis focused solely on the data collected before the pandemic.

5.2.3 Human activity recognition

Activity recognition involves models that classify the activities of individuals based on the data collected from sensors[89, 90]. This analysis aimed to classify PD status based solely on patterns of active walking. As participants were not monitored while wearing the device, there is no ground truth label for actions performed by participants. Instead, robust machine learning classifiers were required to identify periods of walking and activity intensity from the accelerometer data. The outputs of these independent classifiers were then merged to identify active walking windows. Only these windows were then used to classify PD status. A detailed illustration of this classification pipeline is provided in Figure 5.3.

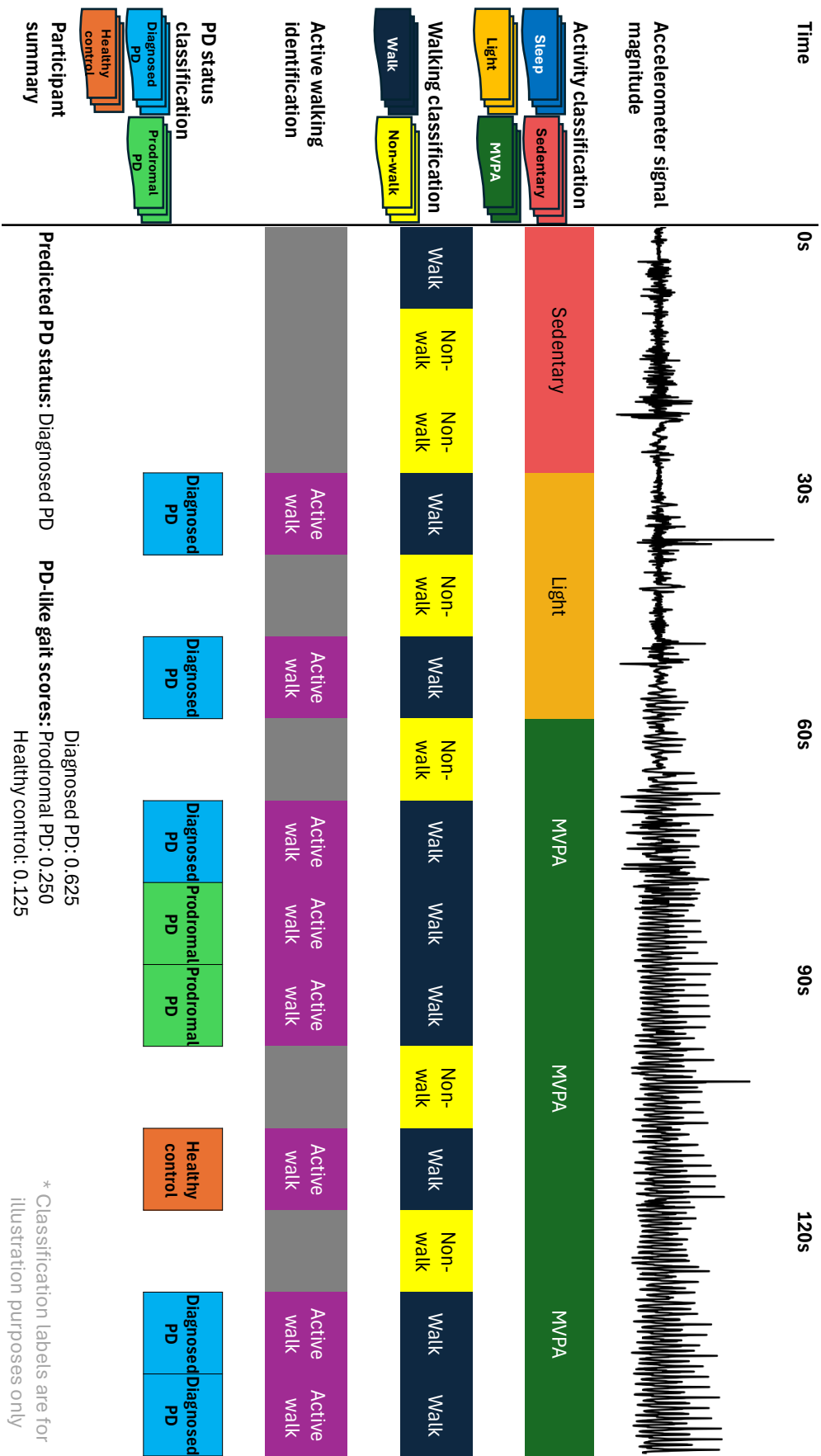


Figure 5.3: Illustration of the use of activity intensity and walking classifiers to identify 10-second windows of active walking. These active walking windows are then used to classify PD status, producing a summary participant classification of the most frequent PD status label, as well as PD-like gait score.

Activity classification

The activity classifier model, `pypi:actinet`, comprised a modified ResNet-18[91] model followed by hidden Markov model (HMM) smoothing, outlined in Section 3.2. I used the deployed version of this model, v0.1.0, which was trained on 80% of the Capture-24 dataset[76, 77], with 20% was held for early-stopping and training of the HMM. The mean ($\pm$ standard deviation) per-participant performance of this model using 5-fold stratified group cross-validation is a macro F1 score of 0.83 ± 0.11 .

To produce 10-second windows with activity intensity labels, the 30-second windowed predictions are resampled to 10-second windows by dividing into three windows of equal length.

Walking classification

The walking classifier model was a modified ResNet-18[91] model trained on the OxWalk[79] and MJFF Levodopa Response study datasets[83], outlined in Section 3.3. This walking classifier was trained on data from both healthy and diagnosed PD populations due to potential differences in walking in these populations[9, 140]. The raw accelerometer signal was downsampled to 30Hz, and divided into 10-second windows, with the binary label of either walking or not-walking. The output label of the ResNet-18 model did not have any additional smoothing. The model was trained on 80% of the participants from both datasets, with 20% reserved for early-stopping of training. Using 5-fold stratified group cross-validation, the classifier has a mean ($\pm$ standard deviation) F1 score of 0.91 ± 0.09 .

Active walking identification

To identify active walking windows, the classification labels produced by the activity and walking classifiers were combined. In this research, active walking windows were defined as 10-second windows of accelerometry that were classified as light or moderate-vigorous activity by the activity classifier and as walking by the walking classifier. Following this identification of active walking windows, all other windows of data were filtered out and not used for PD classification.

Model Name	Balanced	Other parameters
Logistic Regression	True	
Random Forest	False	n_estimators: 1000, max_features: *sqrt
Balanced Random Forest	True	n_estimators: 1000, sampling_strategy: not minority, max_features: *sqrt
XGBoost	False	max_depth: 13, n_estimators: 250
ResNet-18	True	Pretrained: False, learning_rate: 0.0001, patience: 10
SSL ResNet-18	True	Pretrained: True, learning_rate: 0.0001, patience: 10

Table 5.1: List of models used in this analysis. *sqrt indicates each tree randomly selects a subset of features equal to the square root of the total number of features at each split. Both ResNet-18 models use an Adam optimiser[107], and are balanced using inverse weighting cross entropy loss[119].

5.2.4 Parkinson’s disease classification using active walking windows

Each active walking window input to the PD classifier model was given a ground-truth label according to the PD status of the participant.

To classify PD status, a range of machine learning (ML) models were proposed, listed in Table 5.1. All these models, except the "Random Forest", handle the imbalance within the dataset, which would naturally bias predictions towards the majority class. For all classifiers except the ResNet-18 models, hand-crafted, signal-related features were extracted from the 2-norm of the acceleration signal, which was then input to the ML model. The list of features extracted from the signal for analysis is shown in Table 5.2. The ResNet-18 models, on the other hand, are either pre-trained or not pre-trained on the self-supervised arrow of time task, as seen in Yuan et al. (2024)[80]. The purpose of evaluating all these models is to allow for the comparison of each of these models on this classification task.

5.2.5 Model Evaluation and Explainability

PD classification performance was evaluated using 10-fold stratified group cross-validation, with reporting of standard performance metrics. Here, groups represent

Name	Description (Units)
Moment features	
avg	Mean (g)
std	Standard deviation (g)
skew	Skewness
kurt	Kurtosis
Quantile Features	
min	Minimum (g)
q25	Lower quartile (g)
med	Median (g)
q75	Upper quartile (g)
max	Maximum (g)
Autocorrelation features	
acf_1st_max	Maximum autocorrelation
acf_1st_max_loc	Location of 1st autocorrelation maximum (s)
acf_1st_min	Minimum autocorrelation
acf_1st_min_loc	Location of 1st autocorrelation minimum (s)
acf_zeros	Number of autocorrelation zero-crossings
Spectral features	
entropy	Spectral entropy calculated from the power spectrum (nats)
power	Sum of the power spectrum (g^2/s)
f1, f2, f3	Dominant frequencies in the power spectrum (Hz)
p1, p2, p3	Power associated with respective dominant frequencies (g^2/s)
FFT features	
fft0, fft1, fft2, ...	Power values at specific frequencies obtained using Welch's method (g^2/s)
Peak features	
npeaks	Number of peaks in the signal per second (1/s)
peaks_avg_promin	Average prominence of peaks (g)
peaks_min_promin	Minimum prominence of peaks (g)
peaks_max_promin	Maximum prominence of peaks (g)

Table 5.2: Description of various features and their units

active walking windows taken from the same participants, such that data from a given participant appears either in the training or test set only, never in both, to prevent information leakage between training and validation folds. Stratification by PD status class label was used to ensure representative examples of PD statuses across training and validation folds. The performance metrics used are accuracy, macro F1, Cohen’s kappa, balanced accuracy and area under the receiver operating characteristic curve (AUROC) scores. With these metrics, it is possible to compare the ML models trained within this study, as well as to other similar studies.

The ML models developed are trained to produce a PD status label and score per 10-second window of active walking. These scores and labels could be used to calculate the per-window performance across the 10 folds. However, I was also interested in how the model performs per participant. To aggregate model performance to per-participant, I used the most frequent label to identify the predicted label for the participant. Scores for participant label prediction were calculated using the average of each of the window scores.

Model explainability is key to this analysis to derive insight into model behaviour. From the sample 10-second windows of active walking, clusters of behaviour were sought after using clustering, through a Gaussian mixture model (GMM), aided by a t-distributed stochastic neighbour embedding (t-SNE) plot for visualisations[147]. The association between these clusters of behaviour and predicted labels of the ML classifier provides some insight into the model decision process. In addition, the univariable association of a subset of extracted signal features found in Table 5.2 to PD classifier scores were observed, investigating patterns in model behaviour with expected signal properties.

5.3 Results

5.3.1 Study population

The cohort for this analysis comprised a total of 314 participants. Among them, 133 individuals were diagnosed with PD, 147 participants were identified as prodromal, and 34 participants were classified as healthy controls. To aid in visualising the

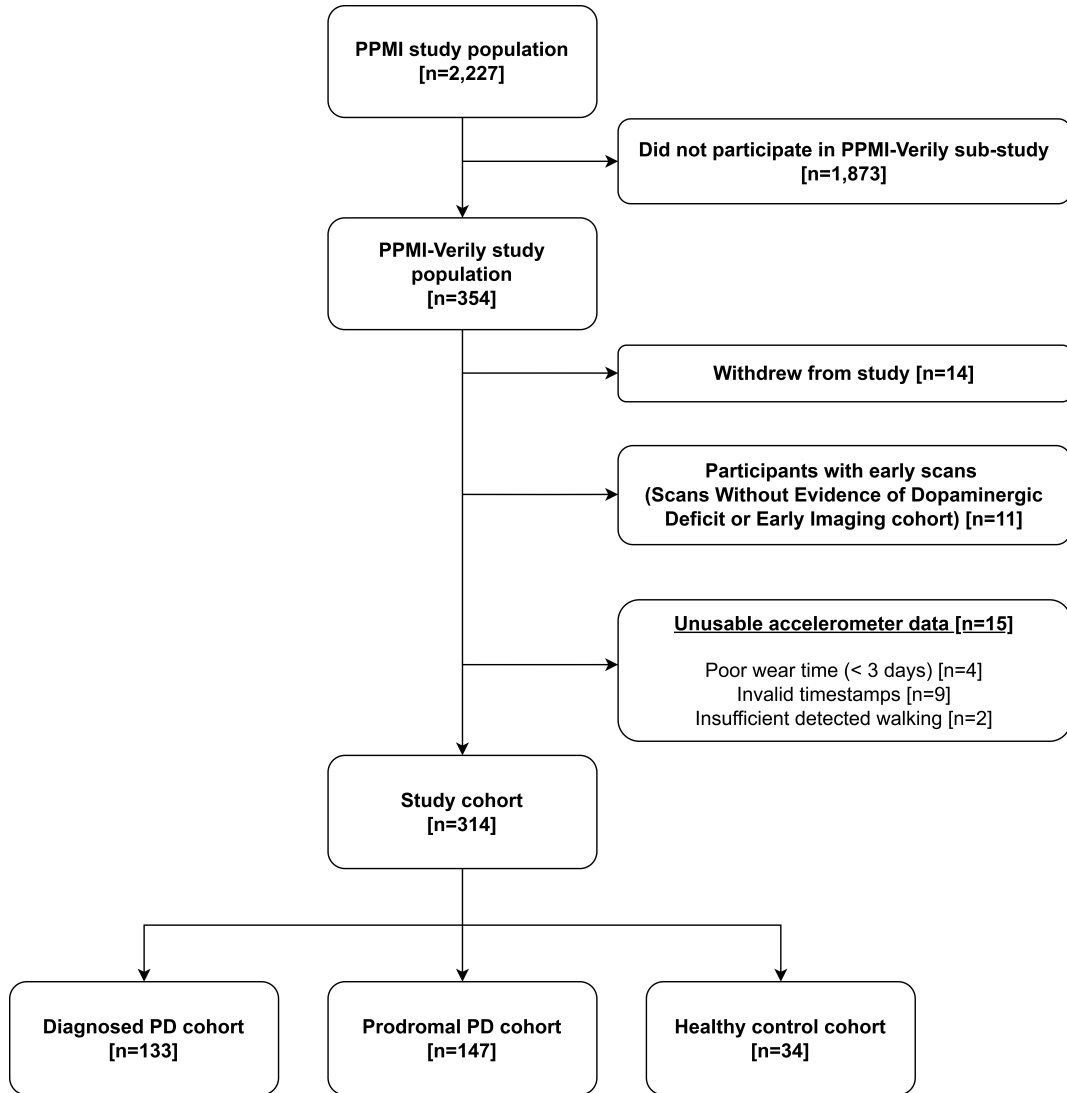


Figure 5.4: Exclusion flowchart for categorising Parkinson's disease status using active walking windows in the PPMI-Verily population

exclusion process and participant flow, Figure 5.4 shows the exclusion flow diagram for participants within this study.

5.3.2 Human activity recognition

Figures 5.6 a, b and c display the output of the activity and walking classification models and the corresponding active walking identification in an example participant. Predicted profiles from participants matched the expected behaviour of activity and walking profiles observed from participants in the UK Biobank with regard to the incidence of activity levels and walking[50, 79]. Figure 5.5 displays the overall

distribution of the number of 10-second active walking frames detected from the first 7 days of monitoring for participants. This shows that active walking was detected in all participants, hence, data from all 314 participants can be used in this analysis.

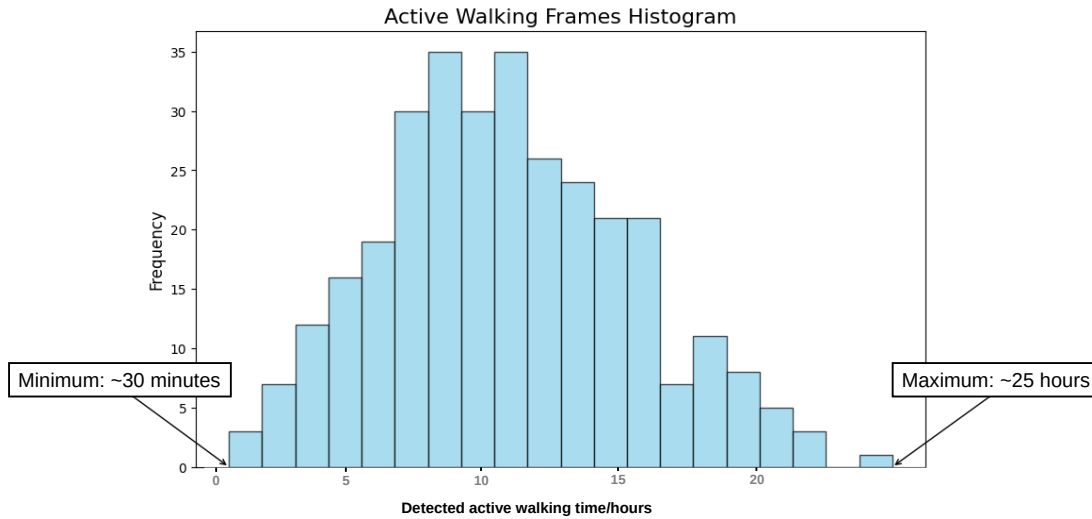


Figure 5.5: Histogram for the count of detected 10-second active walking in the PPMI-Verily population.

5.3.3 PD Classification

The performance of the ML models developed to classify PD status was evaluated per window and participant in Table 5.3. Noticeably, both per window and per participant, the SSL ResNet-18 model outperforms all other models evaluated using all metrics except accuracy, for which the random forest performs best. It should be noted, however, that this dataset is quite unbalanced; hence, macro F1, Cohen’s kappa and balanced accuracy are more reliable performance metrics in assessing model performance.

The confusion matrix for the SSL ResNet-18 model is displayed in Figure 5.7. This model was able to best classify diagnosed PD; however, it struggled to differentiate prodromal PD from healthy controls.

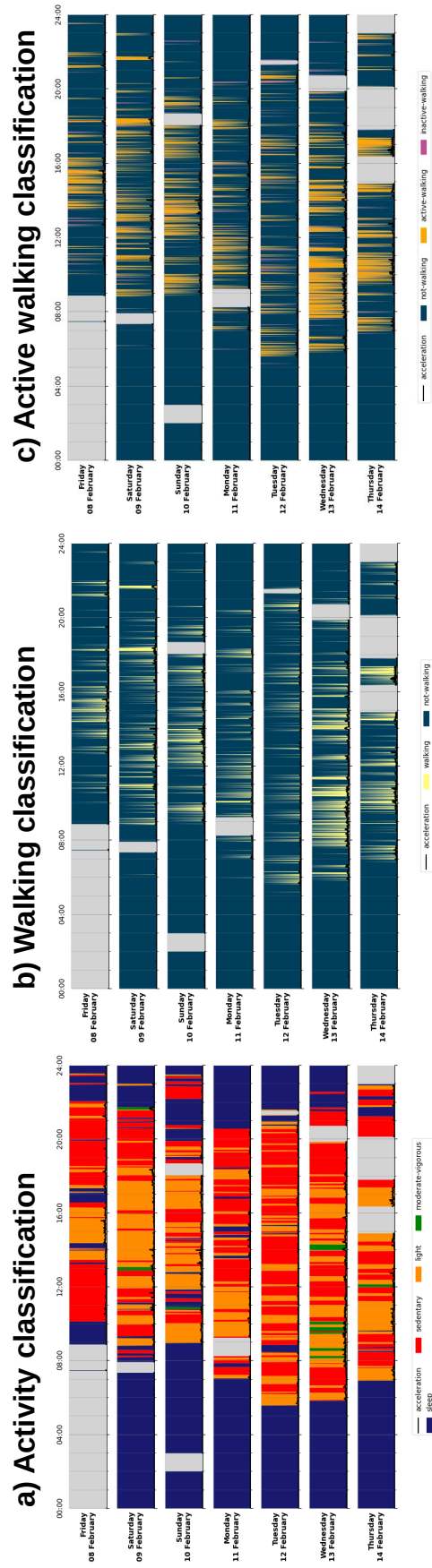


Figure 5.6: Detected activity plot for sample participant. a) Activity intensity classified using `pypi:actinet` to detect labels of sleep, sedentary, light or moderate-vigorous activity. b) Walking or not-walking classified using SSL ResNet-18 model trained on OxWalk and MJFF Levodopa Reponse study datasets. c) Active walking classification defined as period or detected walking by b) and light or moderate-vigorous activity detected by a).

Model	Accuracy	Macro F1	AUROC	Cohen's kappa	Balanced Accuracy
a) Per window model performance across 10 folds					
Logistic Regression	0.423 $\pm$ 0.026	0.390 $\pm$ 0.021	0.612 $\pm$ 0.023	0.128 $\pm$ 0.030	0.414 $\pm$ 0.022
Random Forest	0.563 $\pm$ 0.020	0.396 $\pm$ 0.015	0.630 $\pm$ 0.021	0.206 $\pm$ 0.036	0.423 $\pm$ 0.016
Balanced Random Forest	0.395 $\pm$ 0.021	0.380 $\pm$ 0.018	0.629 $\pm$ 0.023	0.133 $\pm$ 0.025	0.425 $\pm$ 0.022
XGBoost	0.563 $\pm$ 0.020	0.398 $\pm$ 0.016	0.632 $\pm$ 0.022	0.207 $\pm$ 0.037	0.424 $\pm$ 0.016
ResNet-18	0.468 $\pm$ 0.032	0.428 $\pm$ 0.022	0.654 $\pm$ 0.030	0.190 $\pm$ 0.033	0.452 $\pm$ 0.025
SSL ResNet-18	0.505 $\pm$ 0.020	0.465 $\pm$ 0.017	0.694 $\pm$ 0.024	0.247 $\pm$ 0.025	0.485 $\pm$ 0.023
b) Per participant model performance across 10 folds					
Logistic Regression	0.557 $\pm$ 0.054	0.475 $\pm$ 0.057	0.786 $\pm$ 0.038	0.334 $\pm$ 0.062	0.550 $\pm$ 0.092
Random Forest	0.662 $\pm$ 0.056	0.454 $\pm$ 0.048	0.812 $\pm$ 0.033	0.378 $\pm$ 0.104	0.488 $\pm$ 0.042
Balanced Random Forest	0.427 $\pm$ 0.045	0.376 $\pm$ 0.050	0.807 $\pm$ 0.038	0.242 $\pm$ 0.050	0.529 $\pm$ 0.060
XGBoost	0.668 $\pm$ 0.056	0.462 $\pm$ 0.047	0.813 $\pm$ 0.034	0.390 $\pm$ 0.104	0.493 $\pm$ 0.043
ResNet-18	0.632 $\pm$ 0.100	0.521 $\pm$ 0.072	0.840 $\pm$ 0.052	0.430 $\pm$ 0.124	0.579 $\pm$ 0.083
SSL ResNet-18	0.685 $\pm$ 0.070	0.586 $\pm$ 0.082	0.856 $\pm$ 0.056	0.501 $\pm$ 0.099	0.615 $\pm$ 0.108

Table 5.3: Model performance metrics across 10 folds for per window and per participant analysis to classify PD status using active walking windows.

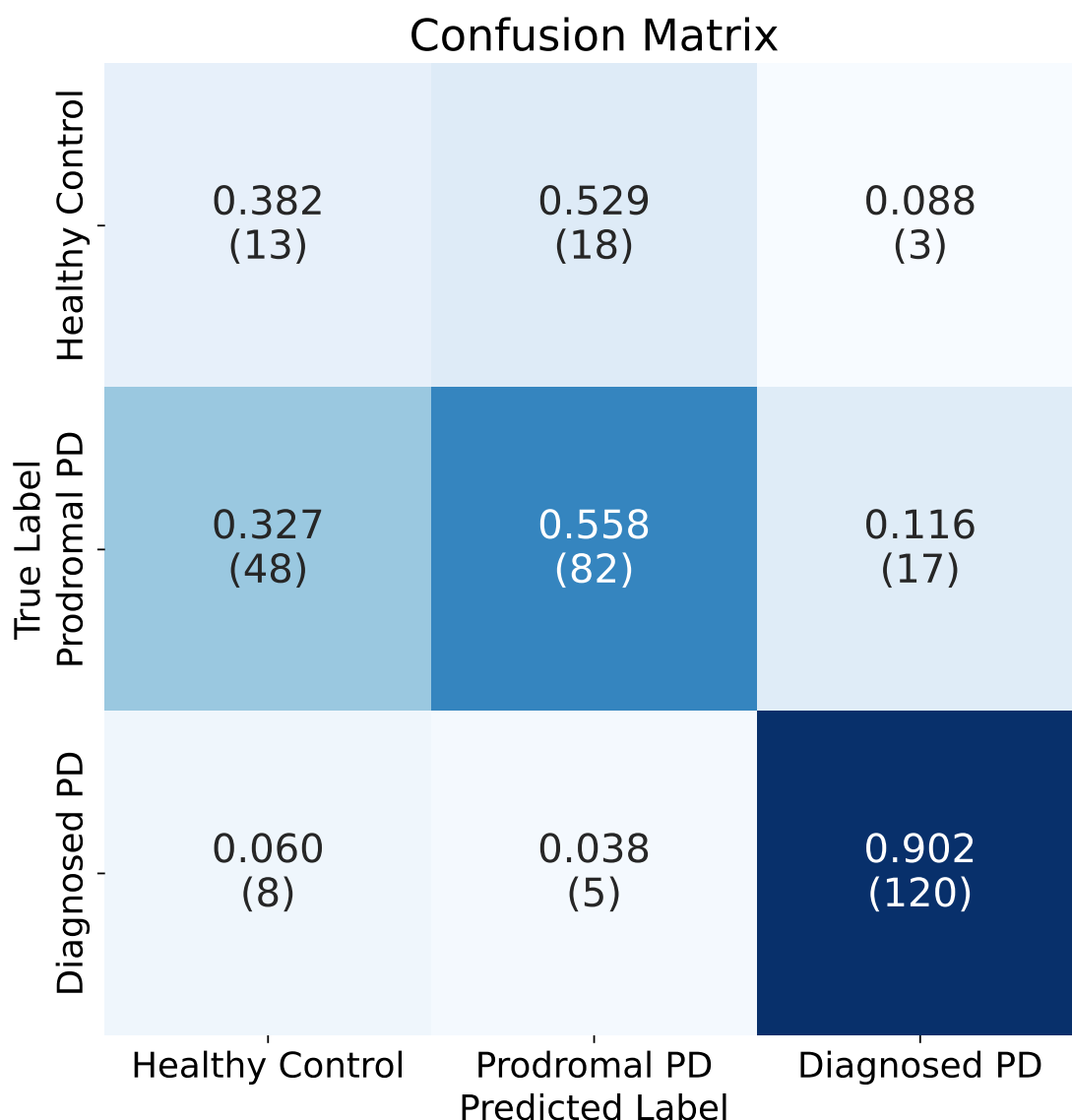


Figure 5.7: Confusion matrix for SSL ResNet-18 model to classify PD status in the PPMI-Verily population. Participant count indicated in parentheses.

5.3.4 Explainability

To explain the model decision process, the extracted features listed in Table 5.2 were first leveraged. Although models trained using these features performed worse compared to the SSL ResNet-18, the features held valuable, interpretable information about each frame, which could be associated with model classification behaviour. However, the univariate correlations between prediction scores for each PD classification label and the listed features did not reveal any insightful

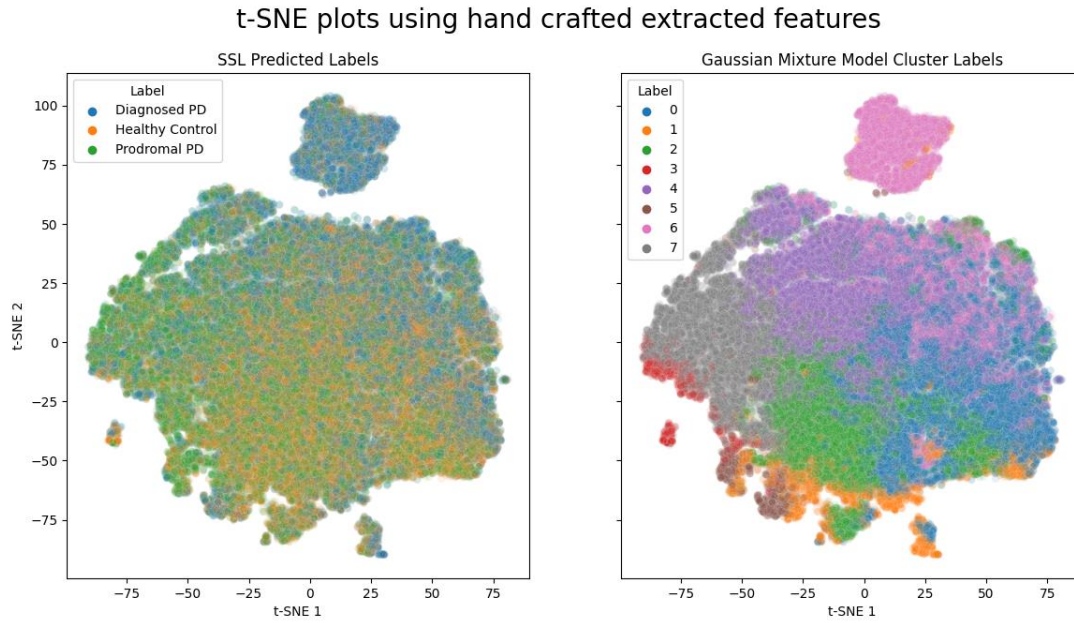


Figure 5.8: Visualisations of the spread of 10s detected active walking, differentiated using PD label of the source of the window, and Gaussian Mixture model clusters.

patterns regarding model behaviour.

To further interpret the basis of model predictions, associations between predicted labels and unsupervised learned clusters were examined. For 100,000 balanced subsamples of active walking windows, a GMM was used to produce eight clusters, representing distinct types of walking behaviour captured from the training dataset based on the extracted features. Figure 5.8 displays a side-by-side visualisation, using a t-SNE plot of these subsamples, showing GMM cluster labels and the PD status of each participant's source walking frame. The cross-tabulation of clusters and PD status is shown in Table 5.4. An association was found between diagnosed PD and Cluster 6, prodromal PD and Cluster 7, and healthy controls and Cluster 0.

GMM cluster	0	1	2	3	4	5	6	7
Predicted PD status								
Diagnosed PD	0.327	0.262	0.189	0.308	0.35	0.298	0.641	0.241
Healthy Control	0.385	0.386	0.392	0.328	0.312	0.283	0.210	0.226
Prodromal PD	0.287	0.352	0.418	0.364	0.338	0.42	0.149	0.533

Table 5.4: Relative cross tabulation of GMM cluster and PD predicted status - Ratio of each predicted PD status per cluster

To observe some useful samples of clusters, the top 10 frames closest to the

centroid of these listed clusters of interest were visualised, displayed in Figure 5.9. This gives a representative sample of typical accelerometry behaviour that the SSL model associated with each PD status. Walking frames in Cluster 6, which were linked to model-classified diagnosed Parkinson's disease (PD), typically exhibited low signal amplitude. In contrast, walking frames in Cluster 7, associated with model-classified prodromal PD, displayed more structured movement patterns compared to those in Cluster 0, which were linked to model-classified healthy controls.

5.4 Discussion

This study investigated the ability to classify PD status by analysing active walking behaviour from free-living data collected with a wrist-worn accelerometer. Findings demonstrate that a deep learning model, specifically a self-supervised ResNet-18, effectively identifies periods of active walking and classifies PD status, outperforming traditional machine learning models across all evaluation metrics. This model's superior performance highlights its capacity to capture relevant walking patterns, advancing the potential for deep learning in digital biomarker extraction by uncovering insights into disease-specific movement profiles.

5.4.1 Comparison with published literature

In the context of existing studies, this research expands on PD classification by distinguishing among three PD statuses: diagnosed PD, prodromal PD, and healthy controls. Most similar studies only compare diagnosed PD cases and healthy control groups. Furthermore, this study uses free-living accelerometry data, which introduces variability by capturing more diverse, real-world walking patterns, as opposed to data from structured walking tasks. Despite these challenges, the model achieved an AUROC comparable to that reported by Zhang et al. (2020) [148] ($\text{AUROC} = 0.856 \pm 0.056$ vs. 0.8558 in Zhang's study). However, the model falls short of the near-perfect accuracy seen in the work of Atri et al. (2022) [53], which only classified PD vs. non-PD among a limited subset of 13 participants. These findings support the hypothesis that the ResNet-18 model can effectively handle

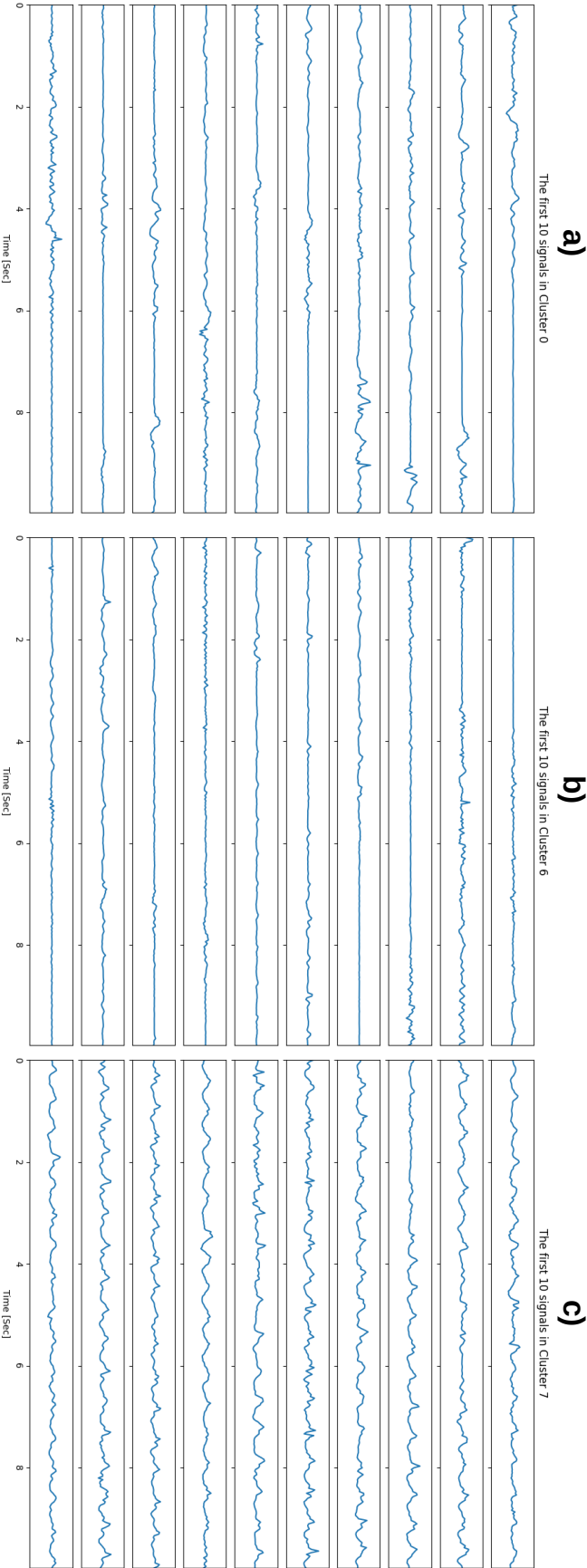


Figure 5.9: Active walking samples closest to centroid of: a) GMM label 0 associated with Healthy Controls, b) GMM label 6 associated with Diagnosed PD, c) GMM label 7 associated with Prodromal PD

real-world walking data, contributing to PD classification research by integrating data from a broader and more variable set of walking behaviours.

Beyond standard model evaluation, this study explored model interpretability by analysing correlations between model predictions and feature patterns. Limited insights could not be extracted from univariate correlations between signal features and model classifications. Instead, I made use of unsupervised clustering techniques to observe patterns in clusters associated with model-classified PD status. Notably, a cluster of walking windows with low amplitude was associated with the diagnosed PD classification, aligning with known clinical indicators such as reduced arm swing strength in PD [81, 120]. Additionally, a novel finding was an association between prodromal PD and clusters with higher regularity in walking patterns. This association, not widely documented in prior literature, may reflect the limited motor symptom expression typical in prodromal PD, suggesting a potential early marker for disease detection that warrants further investigation.

5.4.2 Strengths and Limitations

A key strength of this study was its use of robust, deep learning models to classify raw accelerometry data. This allowed for a hypothesis-free data-driven approach to feature extraction, rather than relying on selected window features for model training. The use of self-supervised learning leveraged the latest machine learning approaches to classification, which further enhanced model performance. This study used a variety of large, diverse datasets to train the classification models, enhancing their generalisability. The approaches taken to explain model operation offered valuable insights into the learned patterns of the machine learning models, providing a useful understanding of the trained decision-making process.

However, several limitations should be addressed. First, the dataset used was restricted to a subset of participants from the PPMI-Verily study, with a relatively small number of healthy controls compared to the diagnosed and prodromal PD groups. Expanding the sample size to have more balanced samples may enhance the robustness of these findings. Additionally, the analysis was based on data

collected over a single 7-day period. Repeated monitoring over a longer term would offer an indication of the stability of the digital measures extracted. Furthermore, the PD status was determined by aggregating the most frequent label predicted across all active walking windows. Future studies could benefit from more nuanced modelling approaches, such as employing more sophisticated approaches to handling the per-window classifications. Improved phenotyping and subtyping based on specific PD-related motor symptoms, as well as timed labelling of PD-like behaviour, may also contribute to improved PD status classification models.

5.5 Summary

This study demonstrated the potential of using deep learning models and digital biomarkers extracted from wearable sensors to accurately classify Parkinson’s disease status based on predicted walking behaviour. The self-supervised ResNet-18 model was able to classify PD status based on walking patterns, and the explainability analysis provided insights into learned patterns and associations. These findings suggest that digital biomarkers may offer some potential in association with PD, offering a more objective and comprehensive approach compared to traditional methods. As this field continues to evolve, future research should focus on expanding the dataset, improving phenotyping, and conducting longitudinal studies to validate and enhance the models’ performance over time. Within the PPMI-Verily dataset, the next steps will involve exploring disease trajectories and observing fluctuations in disease severity over time.

6

Use of wrist-worn accelerometers to improve Parkinson's disease risk models in the UK Biobank

Contents

6.1	Introduction	107
6.2	Methods	109
6.2.1	Study design and outcomes	109
6.2.2	Predictor variables	110
6.2.3	Model derivation and validation	112
6.3	Results	113
6.3.1	Study population and Baseline Characteristics	113
6.3.2	Incidence of Parkinson's disease	116
6.3.3	Predictor variables	116
6.3.4	Validation	116
6.4	Discussion	118
6.5	Summary	121

6.1 Introduction

The progressive degeneration of dopaminergic neurons within the substantia nigra pars compacta (SNpc), characteristic of Parkinson's disease (PD), leads to significant changes to motor and non-motor behaviour[2, 9]. The pathophysiology of PD

has been well-documented in existing literature, underscoring the need for early detection and intervention strategies in the absence of a cure[9, 10, 11]. Alongside well-established risk factors such as age and sex, the identification of early-stage biomarkers is critical to enhance predictive models for PD. By detecting PD in its nascent phases, it becomes possible to identify individuals at elevated risk, allowing for earlier diagnosis, monitoring and intervention[11]. One promising avenue for early detection lies in the subtle motor alterations that often precede clinical PD diagnosis. Emerging evidence suggests that these motor changes can be tracked through digital biomarkers derived from wearable technology, such as wrist-worn accelerometers [51, 149]. This study, therefore, investigated the potential of wrist-worn accelerometry data to improve six-year PD risk prediction models, leveraging the extensive data available through the UK Biobank.

The identification of early biomarkers could have significant clinical and research implications. Early detection facilitates the timely preparation of affected individuals and their families, enabling better disease management. It also provides an opportunity to identify and closely monitor individuals at higher risk, which is crucial for understanding disease progression and testing new therapeutic interventions. This is particularly relevant as research into disease-modifying drugs, such as Glucagon-like peptide-1 receptor (GLP-1R) agonists, suggests the possibility of slowing PD progression by mitigating neurodegenerative processes[150]. Consequently, early-stage biomarkers can also serve as critical endpoints in clinical trials, monitoring disease progression and therapeutic efficacy.

Known risk factors for PD, detailed in Section 1.1.1, include demographic, behavioural, and genetic factors. Models like the Predict-PD, developed by Noyce et al. (2014)[151], incorporate these factors, while later iterations have included polygenic risk scores, improving the model's predictive power[24]. However, existing models often lack early predictive measures that capture pre-clinical and prodromal phase changes. Digital biomarkers, therefore, represent a promising approach to fill this gap[51, 149].

Research has shown that the prodromal phase of PD is often marked by early motor changes, observable years before clinical diagnosis[9, 10, 11]. Wrist-worn accelerometers, widely available in consumer smartwatches, can capture these subtle motor fluctuations, creating a potential source for digital biomarkers. Chapter 4 demonstrated that lower daily step counts, as measured through median step count metrics, are predictive of incident PD. These findings are consistent with similar studies by Schalkamp et al. (2023)[58], Liu et al. (2023)[59], and Zhao et al. (2023)[60], which employ diverse accelerometer-derived metrics to capture activity levels as PD risk indicators. In Chapter 5, I further explored how machine learning techniques could classify gait characteristics, distinguishing between various stages of PD neurodegeneration, with particularly strong predictive performance in diagnosed PD cases.

This research aimed to assess how digital measures derived from wrist-worn accelerometry data might enhance predictive accuracy in risk models for incident PD. Specifically, this study compared traditional PD risk models, which use demographic and genetic factors found in the Predict-PD model[151, 24], with models incorporating additional digital metrics from wrist-worn accelerometers. Through these comparisons, the study evaluates the extent to which accelerometer-derived metrics improve predictive performance and identifies which specific metrics yield the most significant enhancements in model performance.

6.2 Methods

6.2.1 Study design and outcomes

This study was conducted on the UK Biobank Physical Activity Monitoring population, outlined in Section 2.2. Participants were excluded from this study if they had unusable accelerometer data or had a diagnosis of PD at baseline. Unusable accelerometer data was defined as having less than 3 days of valid data, not having 24-hour coverage of wear over the up to 7 days, having poorly calibrated accelerometer readings[47], having unrealistically high ENMO ($>100\text{mg}$), or machine learning models failing to run.

The diagnosis of PD was obtained from a combination of the UK Biobank’s linked hospital data, inpatient hospital admissions data, and death data. Diagnoses were coded according to the International Classification of Diseases (version 10) coding system, with outcomes recorded as the first instance of a G20 diagnostic code[71]. This follows the approach recommended by the UK Biobank[71] and used by other studies in studying PD in the physical activity monitoring cohort[58, 59, 60]. Incident PD, however, was defined at their first diagnosis of PD. Participants were censored at the first date of PD diagnosis, date of death, or at the end of their follow-up period (31 October 2022 for England, 31 August 2022 for Scotland, and 31 May 2022 for Wales), whichever came first[66].

6.2.2 Predictor variables

This analysis used a variety of predictors, extracted from the UK Biobank data, and the wrist-worn accelerometry data. The table of how each of these predictors were extracted is shown in Appendix C.

The combination of these features was used to create 7 different models. Each of the models builds on the baseline demographic features used in the PREDICT-PD model[151], excluding erectile dysfunction, listed in the demographic model. This is displayed in Table 6.1.

Accelerometer-derived predictors

As outlined in Table 6.1, the models used in this analysis include a variety of accelerometer-derived metrics. These metrics were extracted from the wrist-worn accelerometry data, collected over the period of monitoring. Overall acceleration was calculated as the average of the truncated Euclidean Norm Minus One (ENMO) values, calculated over the entire monitoring period[48, 58]. Sedentary, light and moderate-vigorous activity were calculated as the respective average of the proportion of the day predicted by the `pypi:actinet` package developed in Section 3.2. This aligned with models deployed in UK Biobank showing the distribution of activity intensity may be predictive of PD[59, 60]. The PD-like gait score was

Model Number	Model Name	Related Research	Features List
1	Demographic factors model	[151]	Age, sex, smoking, family history of PD, coffee, alcohol, constipation, depression
2	Polygenic risk model	[24]	Demographic factors*, polygenic risk score for PD
3	Overall activity model	[24, 58]	Demographic factors*, polygenic risk score for PD, overall activity
4	Activity composition model	[24, 59], Section 3.2	Demographic factors*, polygenic risk score for PD, average light activity, average sedentary activity, average moderate-vigorous activity
5	All activities model	[24, 60], Section 3.2	Demographic factors*, polygenic risk score for PD, overall activity, average light activity, average sedentary activity, average moderate-vigorous activity
6	PD-like gait score model	[24], Chapter 5	Demographic factors*, polygenic risk score for PD, PD-like gait score
7	All factors model	Chapter 5, Section 3.2, [24, 60]	Demographic factors*, polygenic risk score for PD, PD-like gait score, overall activity, average light activity, average sedentary activity, average moderate-vigorous activity

Table 6.1: Summary of predictive models used in this analysis. The features list is a summary of the features used in the model, with the accelerometer-derived measures in bold. All models use the same time-to-event outcome and are Cox regression models with the baseline predictors listed in the demographic factors model. *Demographic factors refer to the set of predictors listed for the demographic factors model.

calculated as the output of the machine learning model developed in Chapter 5, comparing the active walking windows of participants to the diagnosed PD population in the PPMI-Verily study. The correlation matrix of accelerometer-derived measures was produced to ensure that each of these measures provided unique insight into participant behaviour.

6.2.3 Model derivation and validation

The risk models used in this analysis are Cox regression models[129], with the outcome being the time to incident PD. Multiple imputations with chained equations (MICE)[152] were used to replace all missing values in predictors. Imputed values were calculated using all other risk factors, as well as the PD incidence outcome, assuming data is missing at random. This was done five times which is typically sufficient[153], and this balanced the trade-off between computational time and estimation accuracy. Five respective Cox models were fit for each imputation and then combined using Rubin’s rules[154].

All continuous predictors were mean-centred. Categorical predictors however were set to reference to either value corresponding to least consumption (smoker, coffee, alcohol), false (family history of PD, depression, constipation) or had a higher population count (female sex).

The Cox proportional hazards models estimated coefficients for each risk factor for the entire population[129]. To evaluate the 6-year risk, all participants are right censored at 6-year follow-up time. Therefore all incident cases occurring after 6 years were treated as controls. This was to ensure a relatively equal amount of follow-up data from all participants in the study, as follow-up time for those that were not lost to follow-up ranged between 6.4 and 9.4 years.

To assess and adjust for optimism in model performance due to training and validating on the same population, a bootstrapping approach was employed. The seven Cox regression models were trained and validated on eligible participants from the UK Biobank physical activity monitoring cohort. Training and validating on the same dataset can lead to over-optimism in performance metrics, often due to

over-fitting. To mitigate this, optimism was estimated by drawing 1,000 bootstrap samples, each of the same size as the original dataset, with replacement. For each bootstrap iteration, the model was re-derived and then validated against the original dataset. The optimism for each metric was calculated as the average difference in model performance between the bootstrapped samples and the original dataset. This average difference, representing the shrinkage due to optimism, was then subtracted from the apparent performance metrics to yield a corrected estimate of the model’s predictive ability. This shrinkage was also used to calculate the final model coefficients.

To compare the performance of each model, a variety of evaluation metrics were used. The discrimination of the models was evaluated using Harrell’s Concordance index[155], also referred to as C-index, and the continuous net reclassification improvement (NRI)[156]. The explained variance was evaluated using Nagelkerke’s R^2 statistic[157]. The calibration of the models was evaluated using the calibration slope[158] and Integrated Calibration Index (ICI)[159] and visualised using calibration curves.

A summary illustration of the methods used in this analysis is shown in Figure C.1. The reporting of the results is in accordance with the TRIPOD-AI guidelines[160], displayed in Table C.3 in the Appendix.

6.3 Results

6.3.1 Study population and Baseline Characteristics

This study was conducted on eligible participants from UK Biobank physical activity population. Figure 6.1 shows the exclusion flow diagram for participants within the study. In total, 96,405 were used in this risk analysis. This was after excluding participants with unusable accelerometer data or a diagnosis of PD at baseline. Participants with missing demographic/genetic data were, however, not removed. The only risk factors with missing data used in this study, along with their percentage of missingness in the total study population, were: alcohol consumption (0.1%), smoking status (0.3%), coffee intake (0.1%), and polygenic risk score (2.4%).

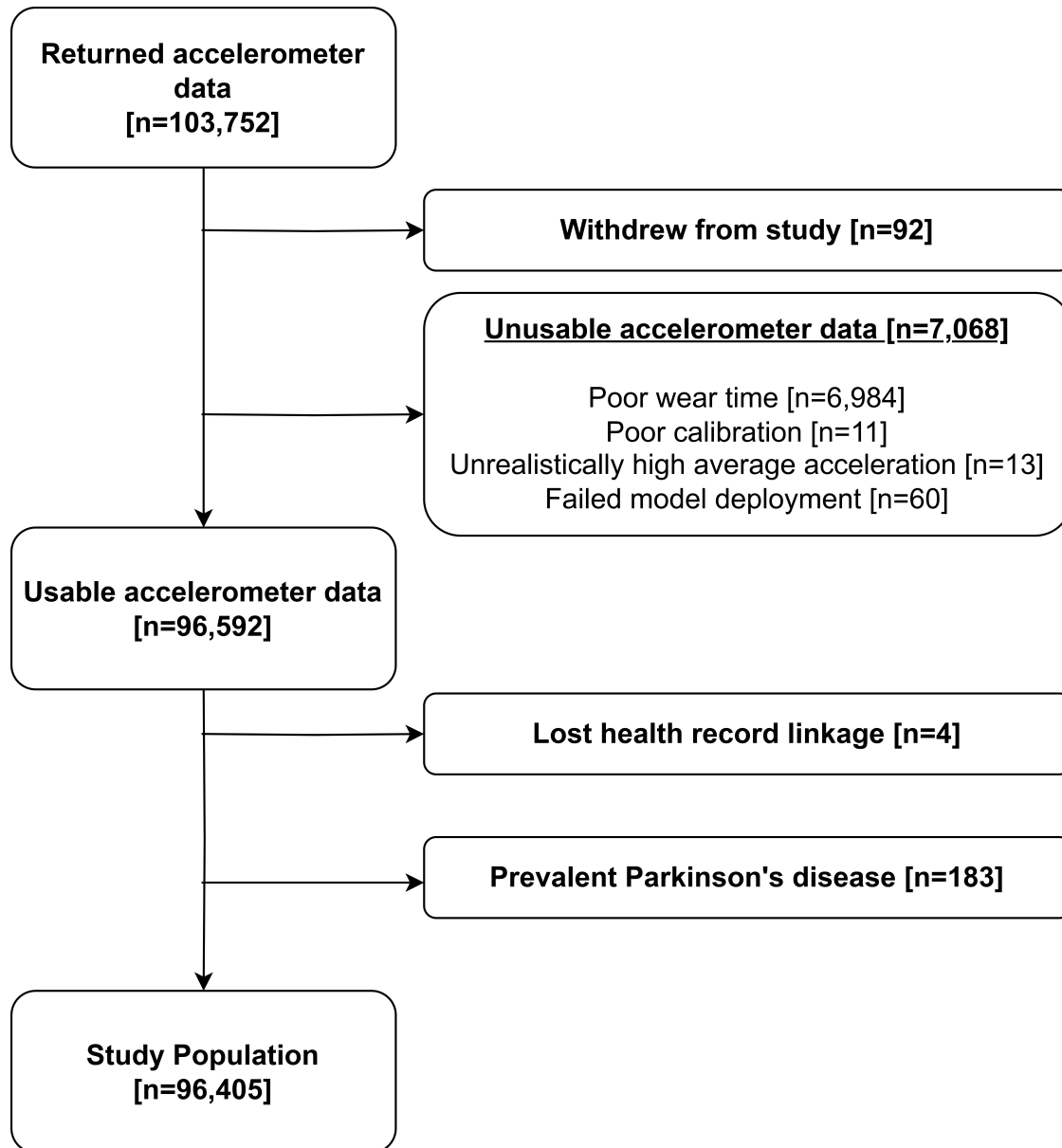


Figure 6.1: Exclusion flow diagram for PD risk analysis in UK Biobank physical activity monitoring population.

The baseline characteristics of this population are displayed in Table 6.2. The study population were aged between 43 and 78 at the time of device wear and 56.3% were female. The majority of the population drank alcohol once a week or more (73.9%), drank coffee more than once a day (51.8%) and had never smoked (56.9%).

Category	Subcategory	Count/Value
Total population		96,405
Demographic factors		
Age (mean (SD)) [years]		62.36 (7.84)
Sex (%)	Female	54,301 (56.3)
	Male	42,104 (43.7)
Alcohol (%)	Less than once a week	25,079 (26.0)
	Once a week or more	71,245 (73.9)
	Missing	81 (0.1)
Smoking (%)	Current	6,669 (6.9)
	Never	54,899 (56.9)
	Previous	34,579 (35.9)
	Missing	258 (0.3)
Coffee (%)	More than once a day	49,928 (51.8)
	One a day or less	46,375 (48.1)
	Missing	102 (0.1)
Health-related factors (%)		
Depression		32,127 (33.3)
Constipation		3,627 (3.8)
Genetic factors		
Family history of PD (%)		4,153 (4.3)
Polygenic risk of PD (mean (SD))		-0.13 (1.02)
	Missing (%)	2,336 (2.4)
Accelerometer-derived factors		
Overall activity [mg]		28.02 (8.23)
Average sedentary time [hours per day]		9.94 (2.07)
Average light activity time [hours per day]		6.20 (2.17)
Average moderate-vigorous activity time [minutes per day]		34.66 (32.55)
PD-like gait score		0.18 (0.10)

Table 6.2: Summary of population demographics and health-related factors for PD risk analysis in the UK Biobank physical activity monitoring population. Missing counts and percentages are provided for all factors that contain missing data. SD=standard deviation, mg=milli-gravity units.

6.3.2 Incidence of Parkinson's disease

The Kaplan-Meier plot illustrating the probability of PD incidence is presented in Figure C.2. There was notable variability in the follow-up duration among participants, with a median follow-up time of 7.9 years. However, the follow-up periods ranged from 6.4 to 9.4 years for participants who neither passed away nor were lost to follow-up. To ensure a comprehensive and equitable assessment of participants' risk, the risk models were censored to evaluate the 6-year risk, aligning with the minimum follow-up duration available for all participants, rounded to the nearest year. The 6-year incidence of PD was determined to be 0.27% (95% CI: 0.24–0.31%), corresponding to 259 participants.

6.3.3 Predictor variables

The hazard ratio and C-index for the univariate Cox model using all follow-up for each of the predictors is shown in Table 6.3.

Accelerometer-derived predictors

Special caution was taken to observe the correlation between accelerometer-derived predictors. To do so, the correlation matrix was calculated and is shown in Figure C.3. A high correlation (absolute value of Pearson correlation coefficient > 0.7) was only seen between average sedentary and light activity behaviour.

6.3.4 Validation

The models' performance was assessed using multiple evaluation metrics, with results adjusted for shrinkage to mitigate overestimation due to training and validation on the same dataset. A comprehensive summary of these metrics is provided in Table 6.4.

Discrimination

Model discrimination was primarily evaluated using Harrell's Concordance index (C-index)[155]. Figure 6.2 presents the comparative C-index values for each model.

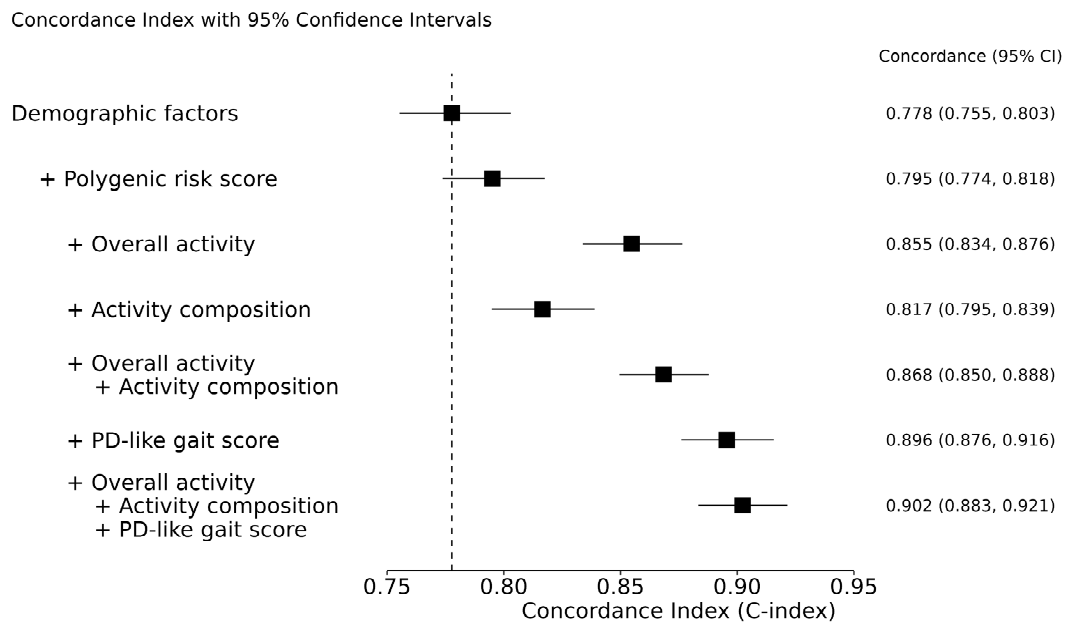


Figure 6.2: Comparison of the optimism adjusted C statistic for each of the PD risk models.

Each model demonstrated a higher C-index than the baseline model, which relied solely on demographic features. Notably, models incorporating accelerometer-derived metrics exhibited improved discrimination. Compared to the C-index of the polygenic risk model (demographic and polygenic risk factors only, C-index = 0.795, 95% CI: 0.774–0.818), models that integrated these risk factors with accelerometer-derived metrics yielded non-overlapping CIs of C-index, except in the case of activity composition.

The PD-like gait score was identified as the most influential additional risk factor when combined with demographic and polygenic risk scores, resulting in a C-index of 0.896 (95% CI: 0.876–0.916). By comparison, models incorporating activity composition (average sedentary, light and moderate-vigorous activity time) or overall activity achieved C-indices of 0.817 (95% CI: 0.795–0.839) and 0.855 (95% CI: 0.834–0.876), respectively. The model that exhibited the highest C-index was the one encompassing all explored risk factors, with a C-index of 0.902 (95% CI: 0.883–0.921), compared to the baseline demographic model's C-index of 0.778 (95% CI: 0.755–0.803).

Further improvement in discrimination over the baseline model was also evaluated using continuous NRI, as presented in the final column of Table 6.4. The findings were consistent with those derived from the C-index; the model with the highest continuous NRI was the model integrating all risk factors.

Calibration

The calibration of each model was assessed using the Integrated Calibration Index (ICI), calibration slope, and inspecting the calibration curves for each model. These are displayed in Table 6.4 and in Appendix C.

All models were well-calibrated, showing a close alignment to observed risk, though a slight tendency to overestimate PD risk relative to observed outcomes was observed. This effect was especially evident in Figure C.9, depicting the calibration curve for the model incorporating demographic, polygenic, and PD-like gait score risk factors. Although the majority of predicted risk scores were below 0.5%, some outlying predictions exceeded 90%, which skewed the smoothed calibration curve away from the ideal, particularly in the context of a less common disease.

6.4 Discussion

The results of this study demonstrated that accelerometer-derived metrics significantly improved 6-year predictive risk models for PD when trained and evaluated on the UK Biobank population. Models that incorporated digital measures from wrist-worn accelerometers worn for up to seven days consistently exhibited higher discriminative performance, assessed through C-index and continuous NRI metrics, compared to baseline models using solely demographic and genetic risk factors. These findings suggest a promising new pathway for the earlier identification of individuals at high risk of developing PD.

The PD-like gait score offered a novel digital measure as a risk factor for PD, demonstrating notable improvements in discriminative ability when included in the risk model. Previous research has documented differences in gait styles and patterns at early stages of PD [81, 120]. However, attempts to quantify these

changes in a way that could serve as a predictive algorithm risk factor were limited. This work, therefore, underscores the value of exploring PD-like gait metrics as a potential digital biomarker.

This study's findings align with prior research on motor changes as early markers of PD. Expanding on the framework established by PD risk models such as Predict-PD [151], accelerometer-derived metrics enhanced model discrimination over traditional demographic and genetic predictors [24, 151]. Improved risk prediction using objective activity metrics had been previously proposed based on observed activity changes in diagnosed and late-stage prodromal PD individuals [58, 59, 60, 138]. For example, Van Nimwegen et al. (2011) [28] hypothesised that these activity changes could be for a variety of reasons, such as increasing disability, poorer walking performance, or the presence of fear and anxiety-related factors, such as fear of falls or depression.

Unexpectedly, cumulative activity emerged as a stronger predictor of PD risk than more detailed activity compositions, such as average sedentary, light, and moderate-vigorous activity levels. One possible explanation is that cumulative activity may better encapsulate early motor changes associated with prodromal PD, while the finer classifications may be more susceptible to classification errors or variability unrelated to PD pathology. Overall, models performed best when using a combination of all digital measures alongside demographic and polygenic risk factors, supporting the inclusion of multiple digital measures rather than a single activity metric.

These findings highlight the potential of accelerometer-derived metrics for improving the identification of individuals at risk of PD. While further validation in independent cohorts is necessary before clinical application, these models can aid in building prodromal PD cohorts within research studies using wrist-worn accelerometer data. This, in turn, may facilitate future research into targeted monitoring strategies. The clinical use of these models remains to be established, requiring additional studies to assess their effectiveness in real-world settings.

This study has several strengths. The use of diverse accelerometer-derived metrics allowed for a wide assessment of movement-related predictors. Additionally, the digital measures were developed and validated on external datasets, thus ensuring the robustness and generalisability of the wrist-worn accelerometer data. This approach reinforces the validity of gait metrics as valuable digital biomarkers for identifying PD risk across diverse populations. The evaluation of the risk model performance robustly took into consideration overestimation and accounted for this using shrinkage, following the standard research guidelines[154, 161].

Several limitations must also be acknowledged. The study's scope was limited to a six-year risk model due to the short follow-up period and this resulted in a relatively small number of incident cases (259) in the UK Biobank's physical activity monitoring cohort. Extending the follow-up duration and incorporating primary care data would increase the incident case sample size, thus enhancing the model's robustness and generalisability [70]. Additionally, these PD risk models did not use commonly known PD risk factors, such as early non-motor symptoms like REM sleep behaviour disorder and anosmia that can be extracted from primary care data[9, 162]. This was not possible for this study as primary care data is only available for roughly 45% of UK Biobank participants. Future risk models however could be further refined by including these additional predictors, once primary care data in UK Biobank is available for all participants[162]. Certain non-motor metrics, such as sleep quality, could also be derived from accelerometer data, potentially improving the model's sensitivity to early PD signs. This study, however, focused on movement-related accelerometer-derived digital measures, which have shown greater classification performance and reliability using current machine learning architectures as compared to sleep metrics[80, 163]. It would also be ideal to validate these risk models using external datasets, however, to the best of my knowledge, no other large cohort dataset currently exists that includes demographic, genetic, and accelerometer data alongside a sufficient number of incident PD cases.

6.5 Summary

The inclusion of various accelerometer-derived measures significantly improved the identification of individuals at high risk of PD. Of particular note, the PD-like gait score provided the greatest improvement in the predictive performance of all the accelerometer-derived measures. These findings support the integration of wrist-worn accelerometers in the development of future PD risk models.

Predictor	Predictor label/ dard deviation	Hazard Ratio (95% CI)	C-index (95% CI)
Age (years)	7.8	3.17 (2.76-3.64)	0.75 (0.74-0.76)
Sex	Female	1.00	0.62 (0.60-0.63)
	Male	2.57 (2.09-3.14)	0.62 (0.60-0.63)
Smoking	Never	1.00	0.54 (0.53-0.56)
	Previous	1.27 (1.04-1.55)	0.54 (0.53-0.56)
	Current	0.41 (0.23-0.75)	0.54 (0.53-0.56)
Alcohol	Less than once a week	1.00	0.51 (0.49-0.52)
	Once a week or more	1.07 (0.86-1.34)	0.51 (0.49-0.52)
Coffee	One cup a day or less	1.00	0.51 (0.50-0.52)
	More than one cup a day	1.08 (0.89-1.31)	0.51 (0.50-0.52)
Constipation	No	1.00	0.51 (0.50-0.51)
	Yes	1.45 (0.94-2.23)	0.51 (0.50-0.51)
Depression	No	1.00	0.51 (0.50-0.52)
	Yes	0.94 (0.77-1.16)	0.51 (0.50-0.52)
Family history of Parkinson's disease	No	1.00	0.52 (0.51-0.53)
	Yes	1.92 (1.35-2.74)	0.52 (0.51-0.53)
Parkinson's disease polygenic risk	1.0	1.44 (1.31-1.58)	0.60 (0.58-0.61)
Overall activity (mg)	8.2	0.26 (0.23-0.30)	0.77 (0.76-0.78)
Average daily sedentary time (hours)	2.1	1.47 (1.34-1.62)	0.61 (0.60-0.62)
Average daily light activity time (hours)	2.2	0.95 (0.86-1.05)	0.53 (0.51-0.54)
Average daily moderate-vigorous activity time (minutes)	32.6	0.69 (0.61-0.79)	0.62 (0.60-0.63)
Parkinson's disease-like gait score	0.1	2.32 (2.24-2.41)	0.83 (0.82-0.84)

Table 6.3: Univariate Cox model hazard ratios and C-index for each predictor with labels. The hazard ratio for continuous predictors is per standard deviation, categorical predictors are relative to the first label.

Risk model	C-Index	Calibration slope	ICI ($\times 10^{-4}$)	Nagelkerke’s R^2	Continuous NRI
Demographic factors	0.78 (0.76, 0.80)	0.97 (0.97, 0.97)	3.67 (1.64, 6.25)	0.05 (0.04, 0.06)	-
Polygenic risk	0.80 (0.77, 0.82)	0.96 (0.96, 0.96)	3.53 (1.57, 6.51)	0.06 (0.05, 0.07)	0.34 (0.23, 0.44)
Overall activity	0.85 (0.83, 0.88)	0.98 (0.98, 0.98)	4.82 (1.81, 9.56)	0.10 (0.09, 0.12)	0.78 (0.66, 0.91)
Activity composition	0.82 (0.79, 0.84)	0.96 (0.96, 0.96)	3.58 (1.04, 7.83)	0.07 (0.06, 0.09)	0.48 (0.34, 0.62)
All activities	0.87 (0.85, 0.89)	0.94 (0.94, 0.94)	4.51 (1.37, 9.76)	0.12 (0.10, 0.13)	0.90 (0.78, 1.01)
PD-like gait score	0.90 (0.88, 0.92)	0.99 (0.99, 0.99)	9.17 (2.59, 17.60)	0.19 (0.16, 0.21)	0.96 (0.82, 1.09)
All features	0.90 (0.88, 0.92)	0.98 (0.98, 0.98)	4.09 (-0.49, 11.87)	0.20 (0.17, 0.22)	1.04 (0.92, 1.16)

Table 6.4: Summary table of model evaluation metrics across different Parkinson’s disease risk models trained and validated on the UK Biobank population. Optimism-adjusted mean and 95% confidence interval shown across 5 imputations and 1,000 bootstraps merged using Rubin’s rule[154]. Continuous NRI is calculated relative to the Demographic factors model.

7

Discussion

Contents

7.1	Summary of findings	125
7.2	Ethical considerations	128
7.2.1	Ethical research conduct	128
7.2.2	Ethical implications of this research	129
7.3	Limitations	131
7.4	Future Work	133
7.5	Conclusion	135

7.1 Summary of findings

This thesis aimed to address several key questions concerning the use of wrist-worn accelerometry and advanced machine learning techniques in understanding and predicting Parkinson’s disease (PD). Specifically, it focused on developing generalisable activity recognition models, assessing the role of reduced activity levels in PD risk, distinguishing between the free-living gait of healthy and PD populations, and exploring whether accelerometer-derived measures could improve PD risk prediction models. Through the analyses presented across different chapters, this research contributes to the understanding of how digital measures can be leveraged to assess and predict PD.

In Chapter 3, the primary goal was to develop improved activity recognition models that are generalisable across both healthy and PD populations. The model `pypi:actinet`, a Human Activity Recognition (HAR) model using a self-supervised ResNet-18[80, 91] for predicting activity intensity labels based on wrist-worn accelerometry data, outperformed the existing HAR model, `pypi:accelerometer`. Notably, `pypi:actinet` exhibited consistent performance across different age groups and sexes, demonstrating its robustness and reliability. The model’s consistency suggests that it can be applied across diverse populations.

Following the development of `pypi:actinet`, further exploration was conducted using self-supervised ResNet-18 models[80, 91] to classify walking behaviours in both healthy and PD populations. This analysis revealed that models trained solely on healthy populations exhibited poorer performance when tested on individuals with more advanced PD motor impairments, based on higher Unified Parkinson’s Disease Rating Scale (UPDRS) Part III scores[14]. This finding highlights a critical limitation of training models on homogeneous datasets and deploying them on populations with varying disease severities. When models were trained on mixed datasets comprising both healthy individuals and PD patients, the performance improved. These models were able to generalise effectively across all populations, showing no dip in performance, even in those with higher PD severity. This result underscores the importance of inclusive training datasets that encompass a range of motor abilities to ensure reliable performance in clinical applications, particularly in motor-altering conditions like PD.

Chapter 4 delved into the relationship between activity levels, measured using median daily step counts, and incident PD in the UK Biobank Physical Activity Monitoring population. It was observed that taking fewer daily steps was associated with incident PD, suggesting that reduced physical activity could be an early marker of PD. However, despite this early association, there was no evidence for daily step count as a modifiable risk factor for PD. This is because the strength of the association between step count and incident PD diminished in cases diagnosed six years after the study period. This observation indicates that while reduced

activity may signal early disease onset, it does not appear to be a direct causal factor. Instead, reduced activity might be reflective of early, subtle motor changes associated with prodromal PD.

In Chapter 5, using the PPMI-Verily dataset, I explored the potential of machine learning models to distinguish between the free-living gait of different populations; healthy controls, diagnosed PD, and prodromal PD. The models were able to capture subtle differences in walking behaviour, particularly of those diagnosed with PD. For example, diagnosed PD participants exhibited lower amplitude fluctuations in acceleration, perhaps an indicator of reduced arm swings during walking, a hallmark of Parkinsonian gait[81, 120]. This ability to differentiate between the gait patterns of various PD statuses demonstrates the potential of accelerometry-derived digital measures as biomarkers for PD. These findings highlight the opportunity to develop digital measures that provide a continuous, objective score for comparing an individual's walking behaviour to established patterns in PD populations. Such measures could be used for early detection, monitoring disease progression, and even assessing treatment efficacy in clinical trials.

Chapter 6 combined the findings from previous work covered in this thesis, by exploring how the digital measures extracted from wrist-worn accelerometers could improve PD risk prediction models. Incorporating these digital measures into traditional risk models that comprised demographic, behavioural, and genetic factors was found to enhance the predictive accuracy of the 6-year risk algorithms. Among the digital measures, the PD-like gait score, developed in Chapter 5, emerged as the biggest contributor to improved risk prediction. This underscores the value of extracting disease-specific digital measures for the development of digital biomarkers. The use of digital measures not only enhances the early detection of PD but also may offer a scalable and non-invasive tool for long-term monitoring.

7.2 Ethical considerations

7.2.1 Ethical research conduct

Undertaking this research project required careful attention to various ethical considerations. Research fundamentally aims to contribute to the pool of generalisable knowledge[164], and this process must be conducted in an ethically sound manner, particularly in health-related research, which can be deeply personal for those involved.

My studentship during this research was funded by GSK through the Engineering and Physical Sciences Research Council (EPSRC). GSK is a pharmaceutical company with a vested interest in exploring the use of digital measures in clinical trials, particularly for discovering novel treatments for diseases such as PD. Given this relationship, I took several steps to mitigate potential conflicts of interest. Firstly, I publicly disclosed GSK’s funding support and the possibility of any perceived conflicts throughout the production of this thesis. Secondly, I ensured that no member of GSK influenced the research process, findings, nor presentation of results. Finally, all datasets and models used in this study are publicly available to ensure transparency and reproducibility of the findings.

Ethical considerations were paramount in the handling of participant data. The careful handling of healthcare data of participants is essential to maintain their privacy[165, 166]. All datasets used in this study were accompanied by explicit consent from participants for the use of their data in health research. I restricted the use of demographic and health data to what was directly relevant to the research aims. Digital measures extracted from accelerometer data were confined to aspects related to gait and activity behaviour, avoiding more sensitive information that could be extracted. The data was securely handled through platforms such as the University of Oxford’s Biomedical Research Computing cluster, GSK’s secure Spark cluster, and the UK Biobank’s Research Analytics Platform. This ensured that no sensitive information was stored on personal devices, thus minimising the risk of theft or unauthorised access.

Special care was taken in the evaluation and presentation of findings. I sought advice from experts in clinical neuroscience and population health epidemiology to ensure the findings, particularly those concerning the association between reduced physical activity and PD, were accurately and responsibly communicated. Furthermore, I delivered a lightning talk to a Patient and Public Involvement (PPI) group to gain feedback on how best to present the potential of PD prediction, given the current lack of a cure. This input helped me refine how I communicated the intentions and limitations of predictive modelling. A further learning point from this interaction is to involve external stakeholders earlier in research project, to allow for greater impact.

7.2.2 Ethical implications of this research

The use of advanced machine learning techniques and wearable technologies in healthcare research introduces several ethical implications that need to be carefully considered[164, 167, 168]. Recent regulations, such as the EU Artificial Intelligence Act (AI Act)[169], underscore the growing need for AI research to be purpose-driven and to provide clear societal benefits. The AI Act, which classifies AI systems based on their risk levels, places particular emphasis on high-risk AI systems, such as those used in healthcare, demanding rigorous governance, transparency, and ethical oversight[170]. This regulatory framework reflects an increasing recognition that AI research should not be pursued merely for its novelty but must address real-world challenges and demonstrate clear utility. As such, new regulations highlight the ethical obligation to not only innovate responsibly, but to ensure that AI technologies are designed and implemented in a way that prioritises human welfare, fairness, and accountability[164, 167, 168, 170].

One key ethical concern is the bias in model performance across different populations[168]. The datasets used in this study, such as the UK Biobank[66] and PPMI[75] cohorts, are disproportionately composed of affluent, White individuals, and valid questions could be posed on the generalisability of the models on less affluent, non-White individuals. This bias may result in suboptimal healthcare

outcomes for underrepresented minorities. Therefore, there is an ethical imperative to ensure that future models are trained and evaluated on more diverse populations to provide equitable healthcare solutions for all[168].

The reliance on wrist-worn accelerometry also raises ethical considerations regarding accessibility and inclusivity[168]. Certain populations, such as those lacking or have limited use of upper limbs or those requiring walking aids, may not benefit from these models, as the accelerometer data collected from wrist-worn devices might not accurately capture their movement patterns. Additionally, the increased promotion of smartwatches for health monitoring could exacerbate health inequalities, as these devices may be too expensive or technologically inaccessible to some segments of the population, particularly older adults or those from lower socioeconomic backgrounds[168].

Another concern is the potential for misuse or unintended consequences of making the models open-source. While transparency and reproducibility are key to scientific progress, making machine learning models publicly available could allow for their use in unethical or harmful ways. For instance, the models developed may be repurposed for surveillance or discriminatory practices. This raises significant concerns about data privacy and the potential exploitation of vulnerable populations, emphasising the need for strict governance and ethical oversight in the use of digital health models[171, 172].

Furthermore, machine learning models, particularly those using self-supervision and large datasets, often operate as black boxes, meaning their decision-making processes are not fully transparent or interpretable, and may lead to unforeseen failure cases[173]. Moreover, training these models on large datasets requires significant computational resources, contributing to high energy consumption[174]. Given the current global climate crisis, the environmental impact of training large machine learning models must be considered as an ethical issue. Encouraging the development of more energy-efficient algorithms is a critical step toward reducing the carbon footprint of AI-driven healthcare research[174].

This research must navigate the balance between digital health innovations and traditional clinical assessments. While digital biomarkers offer the promise of continuous and non-invasive monitoring, they should not be seen as replacements for established clinical tools without sufficient evidence. For example, the U.S. Food and Drug Administration (FDA) rejected Verily’s proposal to use a digital watch as a replacement for the UPDRS Part III score in PD assessments, citing concerns that digital measures may not fully capture meaningful aspects of patient function in day-to-day life[175]. This highlights the importance of ensuring that digital biomarkers complement rather than replace traditional methods, respecting the complexities of patient care[168].

7.3 Limitations

This section addresses the key limitations of the research project. While the findings provide valuable insights, several constraints within the methodology and data must be acknowledged. These limitations influence the interpretation of the results and highlight areas where further research may be required.

This study relied heavily on data from the UK Biobank, which, while one of the largest health research resources globally, presents several challenges. Most notably, the UK Biobank cohort is disproportionately composed of individuals who are White, and of higher socioeconomic status compared to the general UK population. This is also true for the PPMI-Verily cohort, when compared to the US population. This lack of diversity limits the generalisability of the findings to the broader population, particularly to underrepresented groups. The skewed demographics mean that the models trained on this population may not perform as well when applied to more diverse populations, such as those from different ethnic backgrounds or varying economic circumstances.

Moreover, the follow-up time in the UK Biobank study is limited, which constrains the ability to fully investigate pre-clinical or prodromal stages of PD. Parkinson’s is a slowly progressing disease, and longer follow-up periods are required to observe individuals in the earliest, pre-diagnostic stages. Although the study

aimed to explore digital markers during these early phases, the limited follow-up window meant that it was difficult to fully capture the onset of disease and distinguish behaviour from those already with the disease. Consequently, the study is less equipped to make definitive conclusions about long-term pre-clinical changes in physical activity.

Another limitation of the UK Biobank dataset is the limited health information. Following recommendations from Bush et al. (2018)[71], incident PD was identified using hospital episode and death register data. However, PD can also be identifying in GP records' primary care data. Clifton et al. (2024)[70] has shown that the number of identified cases of PD can roughly double with the inclusion of primary care data. However, this data is only provided in UK Biobank up to 2017, for roughly 45% of the study population[67, 70]. This study therefore would benefit from the additional inclusion of primary care data. These health records may also offer greater insight into PD specific medical notes, such as sub-type, severity and regularity of symptoms.

Another significant limitation is the observational nature of the analyses, which restricts the ability to draw causal inferences from the observed associations between physical activity levels and PD risk. While the research showed correlations between digital measures of PD-like gait and reduced physical activity to future PD risk, it remains unclear whether these measures are a direct contributing factor to the disease. Further experimental or interventional studies would be required to uncover the causal mechanisms that could explain the associations found.

In terms of the methodologies used for measuring activity and behaviour, the reliance on wrist-worn accelerometers presents its own challenges. While accelerometry is a useful tool for tracking movement, it remains an imperfect measure of activity. One limitation is that accelerometers capture movement data from the wrist, which may not always accurately reflect the full range of an individual's physical activity. For instance, moderate-vigorous activity was the most difficult activity intensity to reliably identify, partially due to limited training samples, but also as wrist movements are not always directly correlated with overall

activity intensity. Furthermore, people who use walking aids may have limited arm swing, leading the accelerometer to record less movement than is actually occurring. These scenarios contribute to a lack of precision in the classification of activities, potentially leading to misclassification of activity types or intensities.

The lack of external validation presents another limitation within this thesis. The models and findings from this study were not tested on independent datasets from other populations or research cohorts. Without this validation, it is difficult to ascertain how well the developed models would perform in different settings, such as other geographic locations or clinical environments. The absence of external validation introduces uncertainty regarding the robustness and generalisability of the models, highlighting the need for future studies to replicate and validate these findings with external datasets.

7.4 Future Work

This section outlines potential avenues for future research that can build on the findings and address the limitations identified in this study. Expanding the scope of the work will help enhance the robustness, generalisability, and clinical relevance of the results. Several key areas for future work are discussed below.

A primary recommendation for future research is to extend the follow-up time in cohort studies. The current study, constrained by the follow-up window available in the UK Biobank, with a median follow-up of the physical activity monitoring study of 7.8 years. This follow-up period could not fully capture long-term preclinical changes in physical activity. Given the study was conducted between 2013 and 2015, in a few years, follow-up may extend to assess 10-year risk. Extending follow-up periods to 10 or 20 years would allow researchers to better assess the risk of developing PD disease over a longer time frame. Such longer-term data would help identify subtle changes in physical activity that may occur decades before a formal diagnosis, offering valuable insights into early disease markers/modifiable risk factors.

Another important area of future work is the inclusion of external validation datasets. This study's findings are based on data from the UK Biobank and

PPMI-Verily cohorts, both of which have limited diversity in terms of ethnicity and socioeconomic status. To ensure the models and conclusions are broadly applicable, it is essential to validate them using independent datasets from other populations and geographic locations. Incorporating more diverse data sources will enhance the generalisability of the models and help mitigate concerns related to demographic biases. External validation is also crucial for testing the robustness of the developed models in different clinical or community settings, ensuring they are suitable for broader application. Such datasets could include the NHANES[115] or China Kadoorie Biobank(CKB)[176] accelerometer datasets.

In addition to extending follow-up times and incorporating diverse datasets, future research could focus on monitoring disease progression in individuals diagnosed with PD. While this study primarily focused on identifying risk factors and digital markers in pre-clinical stages, there is significant potential to use wrist-worn accelerometry for tracking the trajectory of disease progression. Continuous monitoring could provide a more detailed picture of how motor symptoms evolve over time and how this correlates with other clinical indicators. Such insights could help improve patient management and inform personalised treatment plans by identifying key points where interventions could slow disease progression.

Another promising direction for future research is the development of wrist-worn accelerometry as a digital biomarker for PD. While this study demonstrated the effectiveness of digital measures in identifying disease risk, there is significant potential to refine these measures further for use in both diagnosis and monitoring. For instance, applying machine learning algorithms to accelerometer data could help detect subtle movement patterns indicative of early or prodromal PD. These digital biomarkers could complement traditional clinical assessments by providing a less invasive, continuous monitoring tool, thereby improving early detection and patient outcomes. An example of a successful digital biomarker is 95th percentile stride velocity (SV95C), which has been effectively used for monitoring Duchenne Muscular Dystrophy (DMD)[177]. The results found within this thesis provide

evidence for strong associations between various digital measures and PD, offering the possibility of elevating these digital measures to digital biomarkers.

7.5 Conclusion

This thesis evaluated the potential of digital measures extracted from wrist-worn accelerometers for Parkinson’s disease prediction. The research explored machine learning models, including self-supervised learning approaches, to uncover changes in activity behaviour during prodromal stages of Parkinson’s disease. By leveraging accelerometer data from large datasets such as the UK Biobank and PPMI-Verily cohort, the study has demonstrated that digital measures hold promise as tools for disease prediction. Key findings include the identification of lower physical activity as a potential early marker for future Parkinson’s diagnosis. In addition, a PD-like gait score, a digital measure of the comparison of walking behaviour to diagnosed PD participants in the PPMI-Verily study, showed great predictive performance over traditional risk factors and other digital measures. The findings are promising, showing the potential for wrist-worn accelerometers to improve PD-risk monitoring.

Appendices



Comparison of Accelerometer and ActiNet

A.1 HMM matrices

Below shows the prior, transition, and emission matrices from the training of Accelerometer and ActiNet models on the first 100 participants of the Capture-24 dataset. All matrices are ordered by the activity states: **Sleep**, **Sedentary**, **Light** and **Moderate-vigorous physical activity (MVPA)**. Transition matrices represent the probability of moving from a state S at time t to a state at time $t + 1$. Emission matrices represent the probability of observing an activity intensity O at time t given the state S at time t .

Accelerometer

Prior matrix:

$$\begin{array}{c} \textit{Sleep} \quad \textit{Sedentary} \quad \textit{Light} \quad \textit{MVPA} \\ \left[\begin{array}{cccc} 0.359 & 0.409 & 0.188 & 0.044 \end{array} \right] \end{array}$$

Transition matrix:

$$\begin{array}{c} S(t) \backslash S(t+1) \\ \begin{array}{c} \textit{Sleep} \\ \textit{Sedentary} \\ \textit{Light} \\ \textit{MVPA} \end{array} \end{array} \begin{array}{c} \textit{Sleep} \quad \textit{Sedentary} \quad \textit{Light} \quad \textit{MVPA} \\ \left[\begin{array}{cccc} 1.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.987 & 0.013 & 0.000 \\ 0.000 & 0.029 & 0.964 & 0.007 \\ 0.000 & 0.004 & 0.029 & 0.967 \end{array} \right] \end{array}$$

Emission matrix:

$S(t) \backslash O(t)$	<i>Sleep</i>	<i>Sedentary</i>	<i>Light</i>	<i>MVPA</i>
<i>Sleep</i>	0.868	0.107	0.018	0.007
<i>Sedentary</i>	0.106	0.642	0.190	0.062
<i>Light</i>	0.017	0.181	0.592	0.210
<i>MVPA</i>	0.006	0.061	0.219	0.714

ActiNet

Prior matrix:

<i>Sleep</i>	<i>Sedentary</i>	<i>Light</i>	<i>MVPA</i>
0.409	0.389	0.164	0.038

Transition matrix:

$S(t) \backslash S(t+1)$	<i>Sleep</i>	<i>Sedentary</i>	<i>Light</i>	<i>MVPA</i>
<i>Sleep</i>	1.000	0.000	0.000	0.000
<i>Sedentary</i>	0.000	0.987	0.013	0.000
<i>Light</i>	0.000	0.032	0.963	0.005
<i>MVPA</i>	0.000	0.003	0.019	0.978

Emission matrix:

$S(t) \backslash O(t)$	<i>Sleep</i>	<i>Sedentary</i>	<i>Light</i>	<i>MVPA</i>
<i>Sleep</i>	0.937	0.056	0.006	0.001
<i>Sedentary</i>	0.060	0.749	0.155	0.036
<i>Light</i>	0.011	0.167	0.620	0.202
<i>MVPA</i>	0.007	0.036	0.264	0.693

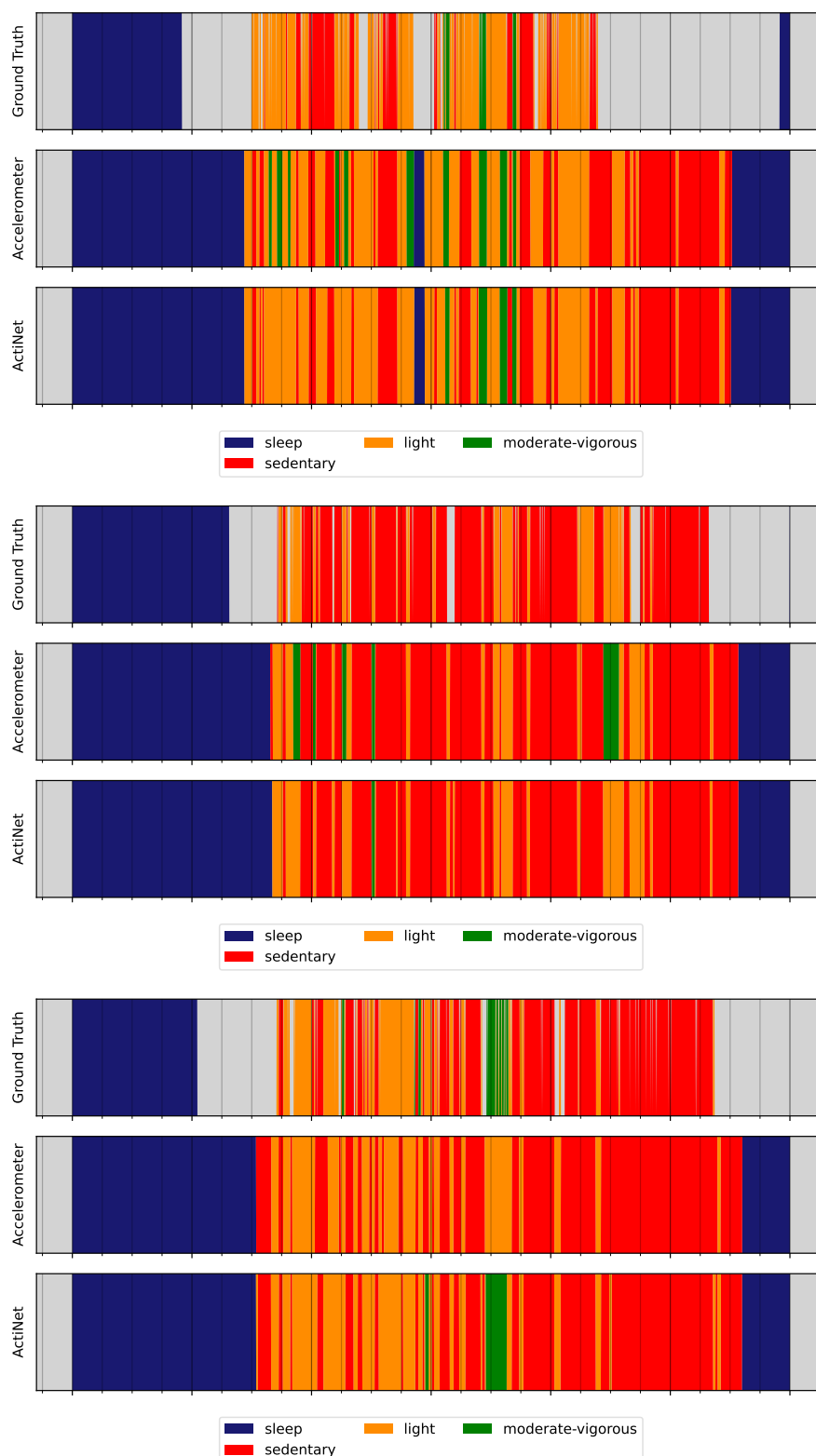


Figure A.1: Comparison of `pypi:accelerometer` and `pypi:actinet` models trained on first 100 Capture-24 participants, deployed on sample test Capture-24 participants. Grey regions in ground truth labels indicate a lack of annotations during data collection.

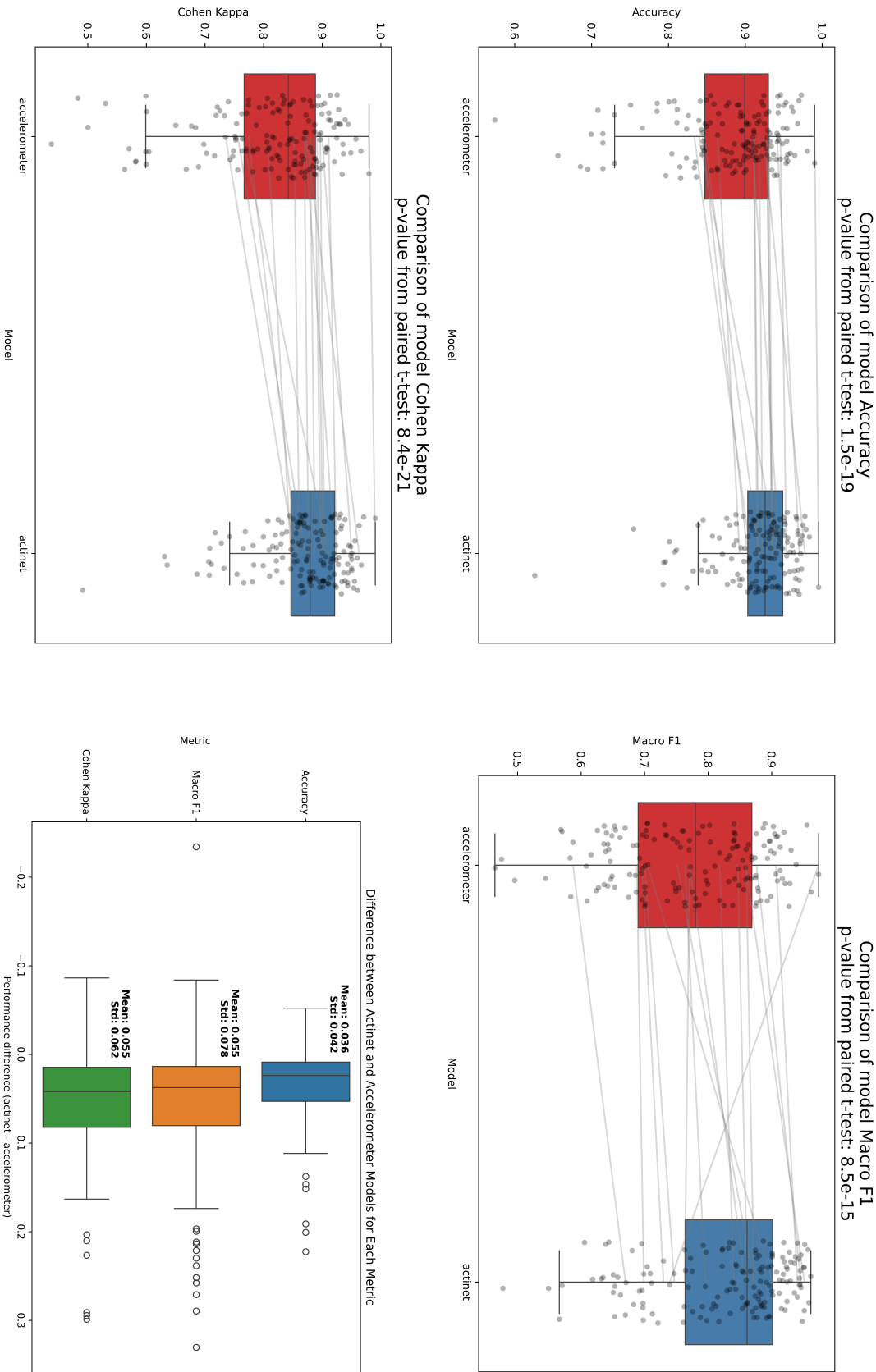


Figure A.2: Visualisation of differences in performance metrics of `ypyi:actinet` and `ypyi:accelerometer`.

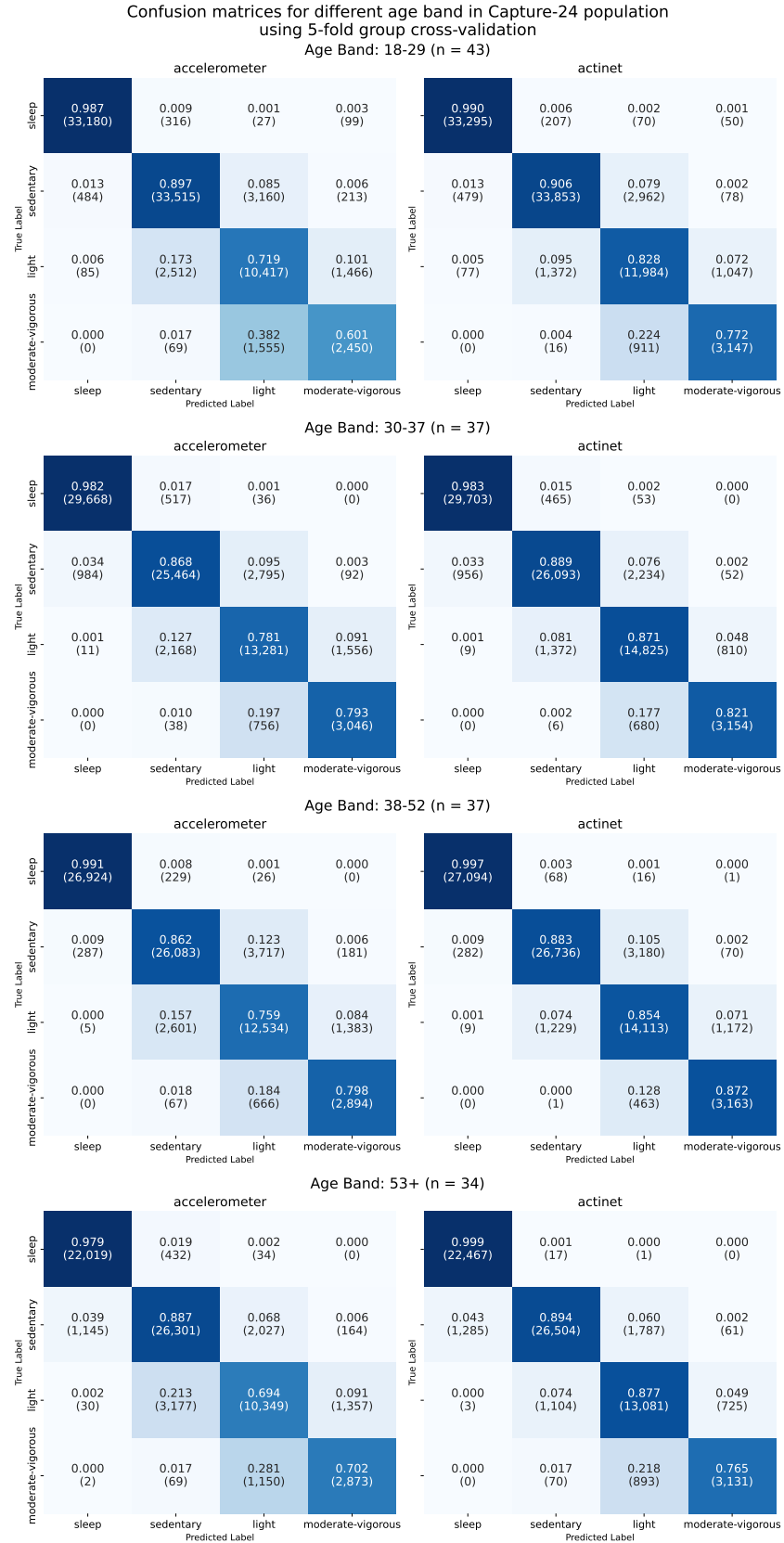


Figure A.3: Confusion matrices for activity intensity classification outputs of `pypi:accelerometer` and `pypi:actinet` for different age bands of Capture-24 population.

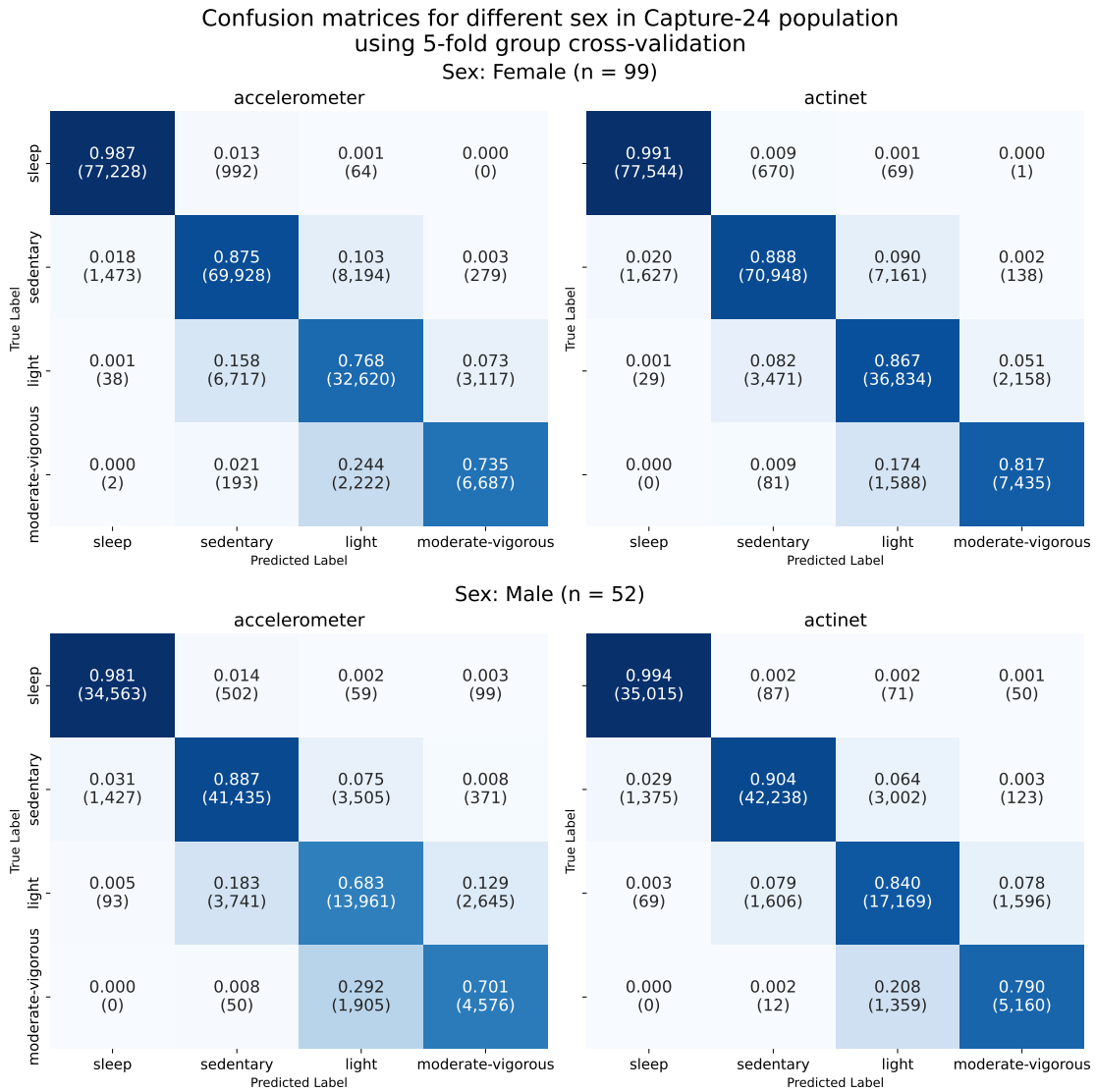


Figure A.4: Confusion matrices for activity intensity classification outputs of `pypi:accelerometer` and `pypi:actinet` for males and females in Capture-24 population.

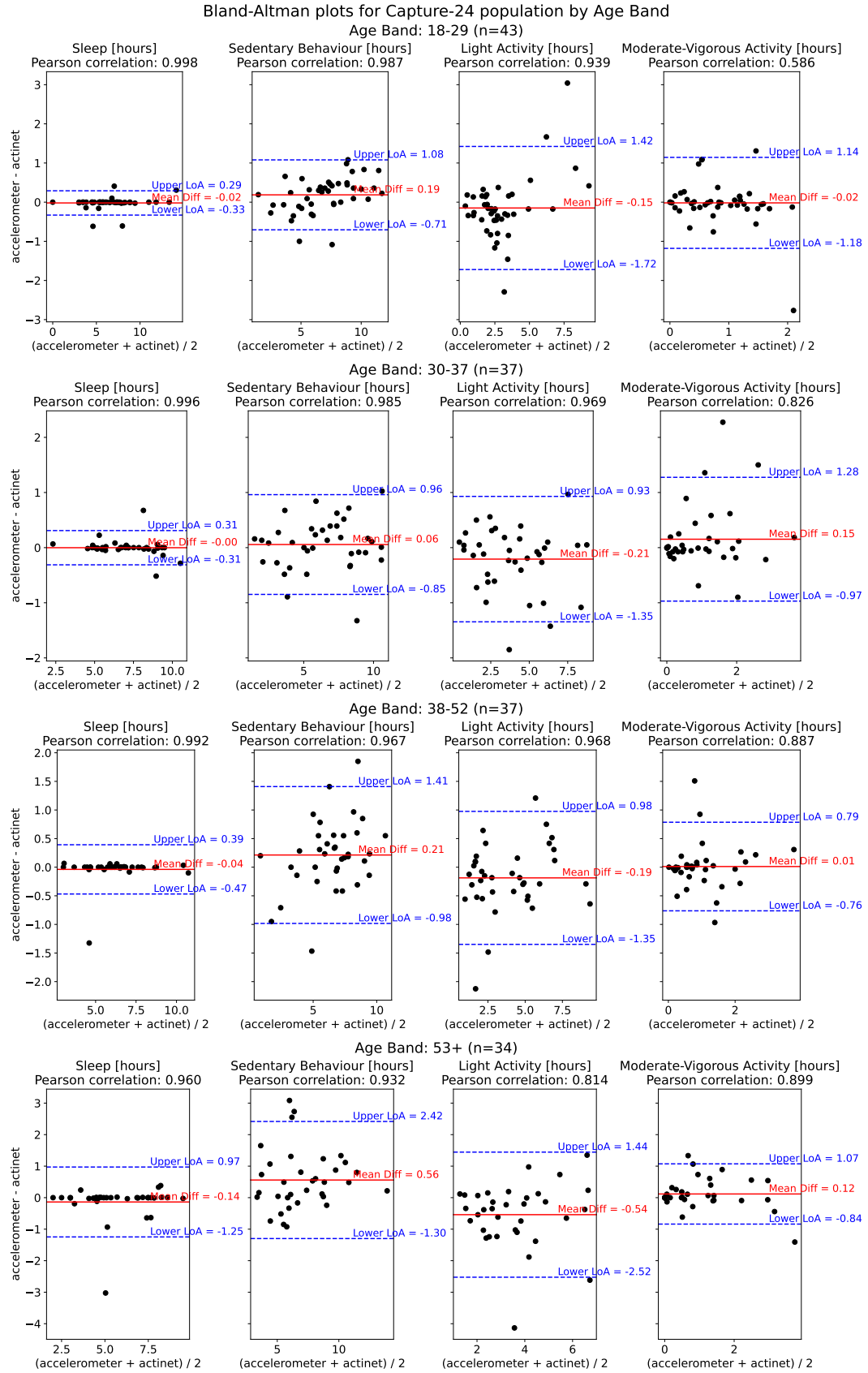


Figure A.5: Bland-Altman plots comparing model activity intensity classification outputs of pypi:accelerometer and pypi:actinet for different age bands of Capture-24 population.

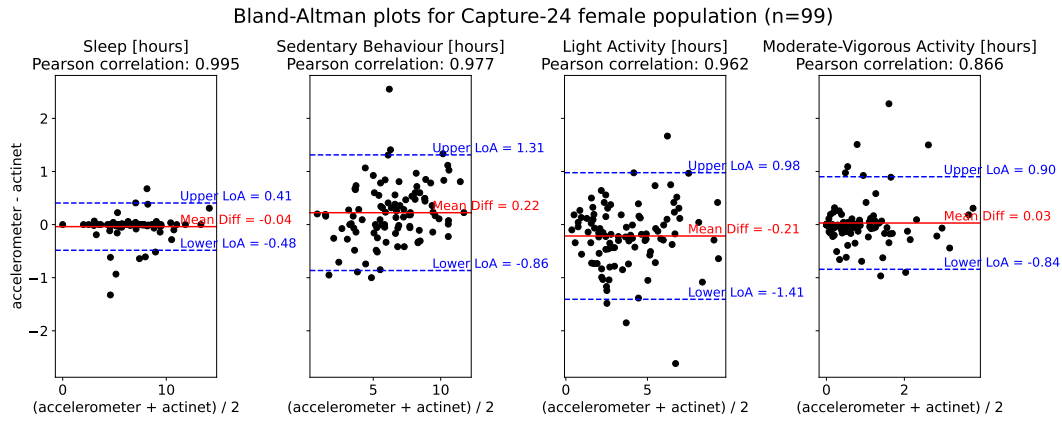


Figure A.6: Bland-Altman plots comparing model activity intensity classification outputs of `pypi:accelerometer` and `pypi:actinet` for the female Capture-24 population.

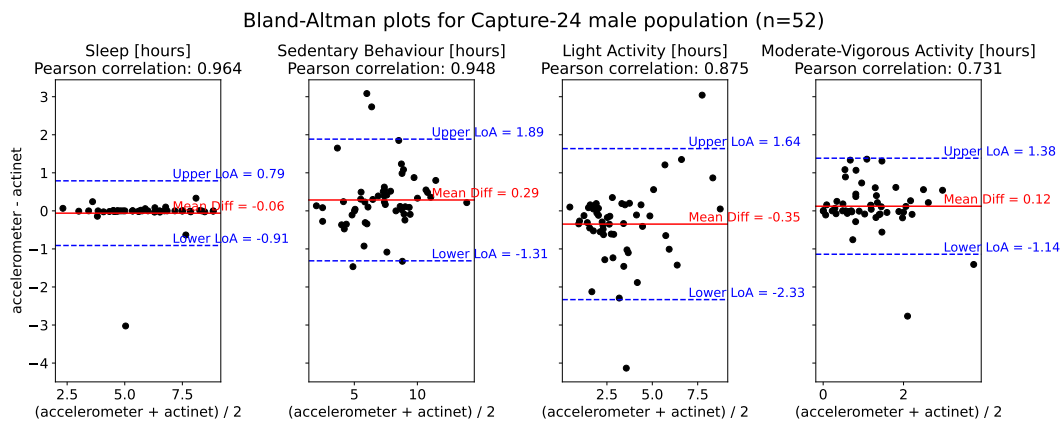


Figure A.7: Bland-Altman plots comparing model activity intensity classification outputs of `pypi:accelerometer` and `pypi:actinet` for the male Capture-24 population.

Fold	Test Participant IDs
1	P001, P003, P006, P010, P012, P013, P017, P025, P030, P044, P056, P057, P059, P065, P070, P075, P079, P080, P082, P090, P092, P114, P115, P116, P117, P130, P131, P142, P145
2	P004, P008, P014, P015, P016, P021, P023, P024, P028, P029, P032, P035, P037, P038, P043, P047, P048, P050, P074, P081, P085, P093, P094, P098, P099, P112, P118, P121, P127, P128, P146
3	P026, P036, P039, P040, P042, P045, P046, P049, P052, P055, P064, P068, P069, P071, P072, P088, P091, P095, P100, P101, P103, P106, P109, P111, P126, P141, P144, P148, P149, P150, P151
4	P002, P009, P011, P018, P019, P027, P031, P051, P053, P054, P060, P061, P067, P076, P077, P078, P083, P086, P108, P110, P113, P119, P122, P123, P132, P133, P135, P136, P138, P139, P147
5	P005, P007, P020, P022, P033, P034, P041, P058, P062, P063, P066, P073, P084, P087, P089, P096, P097, P102, P104, P105, P107, P120, P124, P125, P129, P134, P137, P140, P143

Table A.1: Test Participant IDs for each fold of the stratified group cross-validation on the Capture-24 dataset.

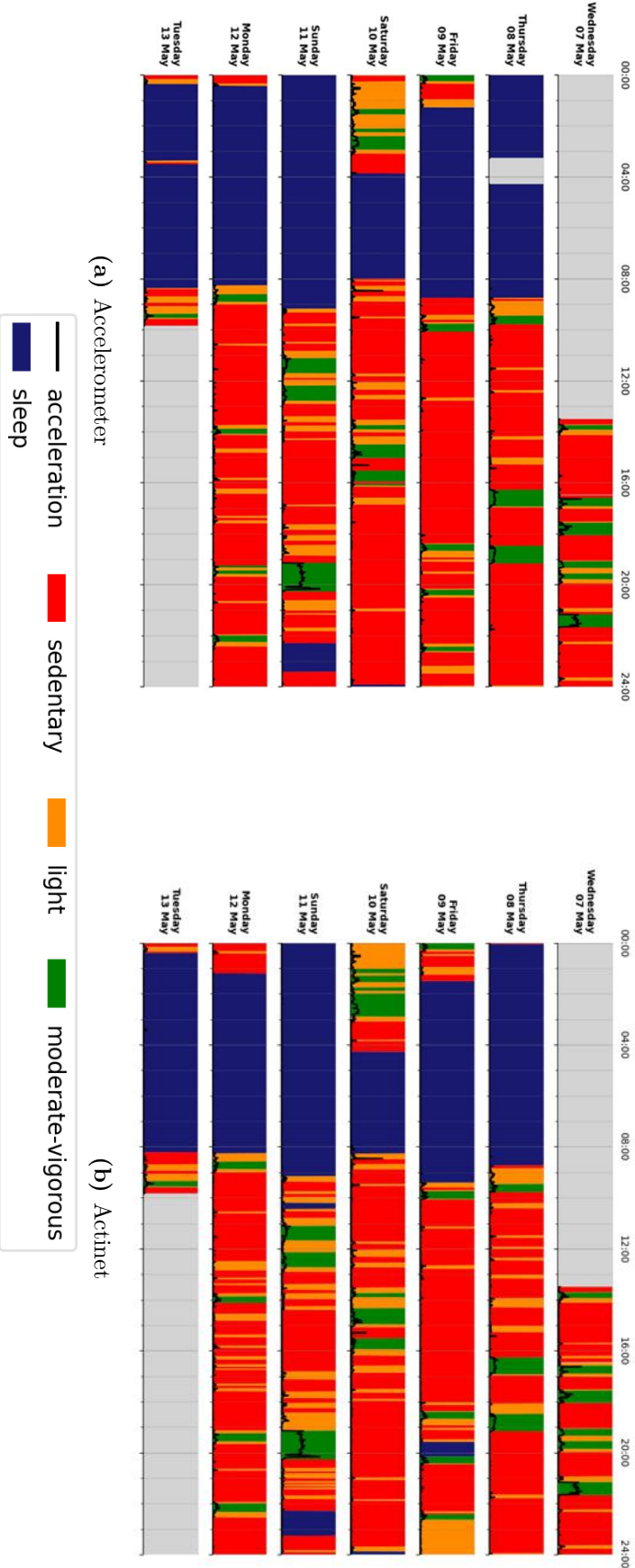


Figure A.8: Deployment of `pypi:accelerometer` and `pypi:actinet` on sample participant data.

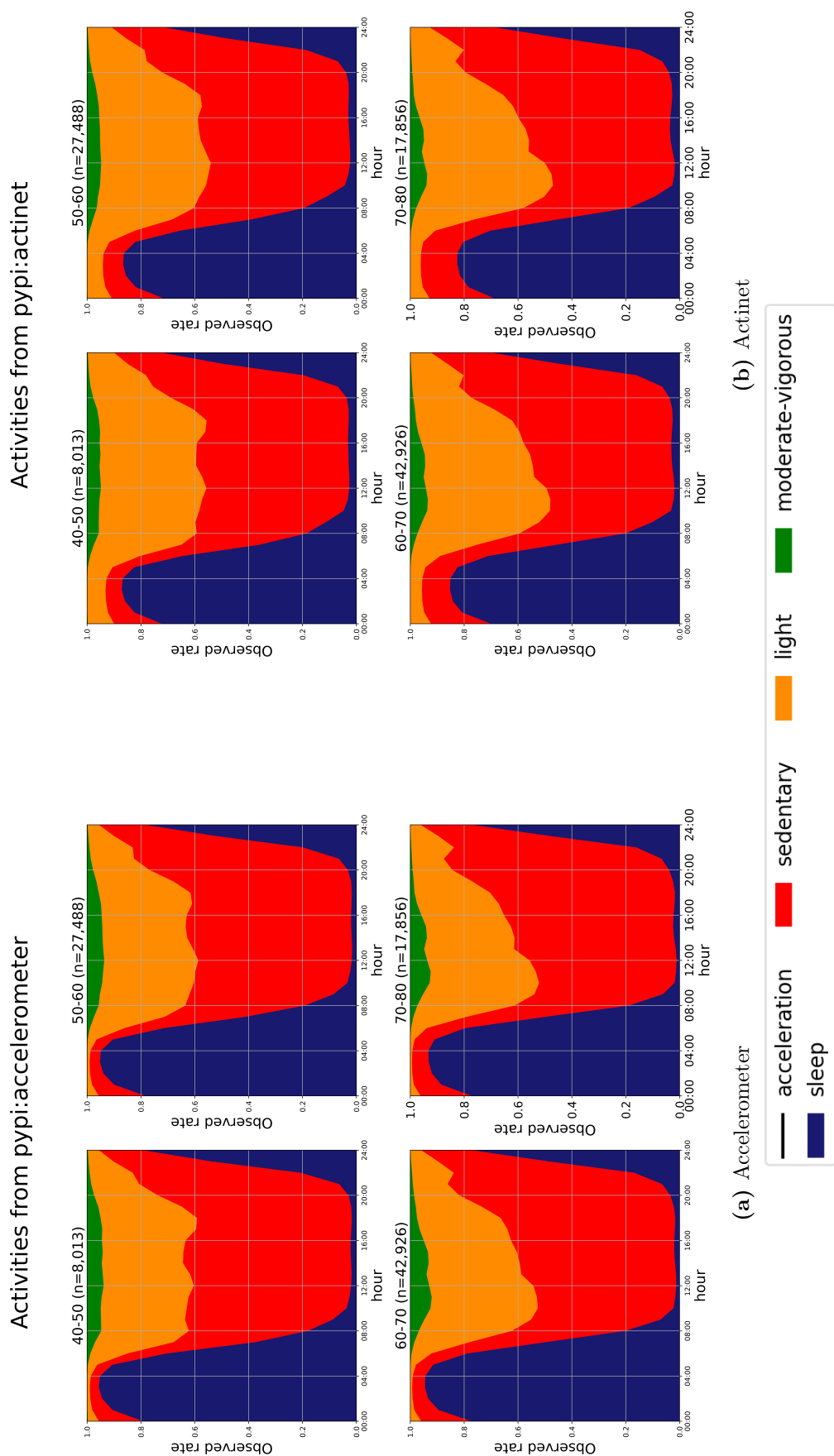


Figure A.9: Observed rate of activities from predictions made by the deployment of Accelerometer and ActiNet on UK Biobank age subpopulations.

Activities from pyapi:accelerometer

Activities from pyapi:actinet

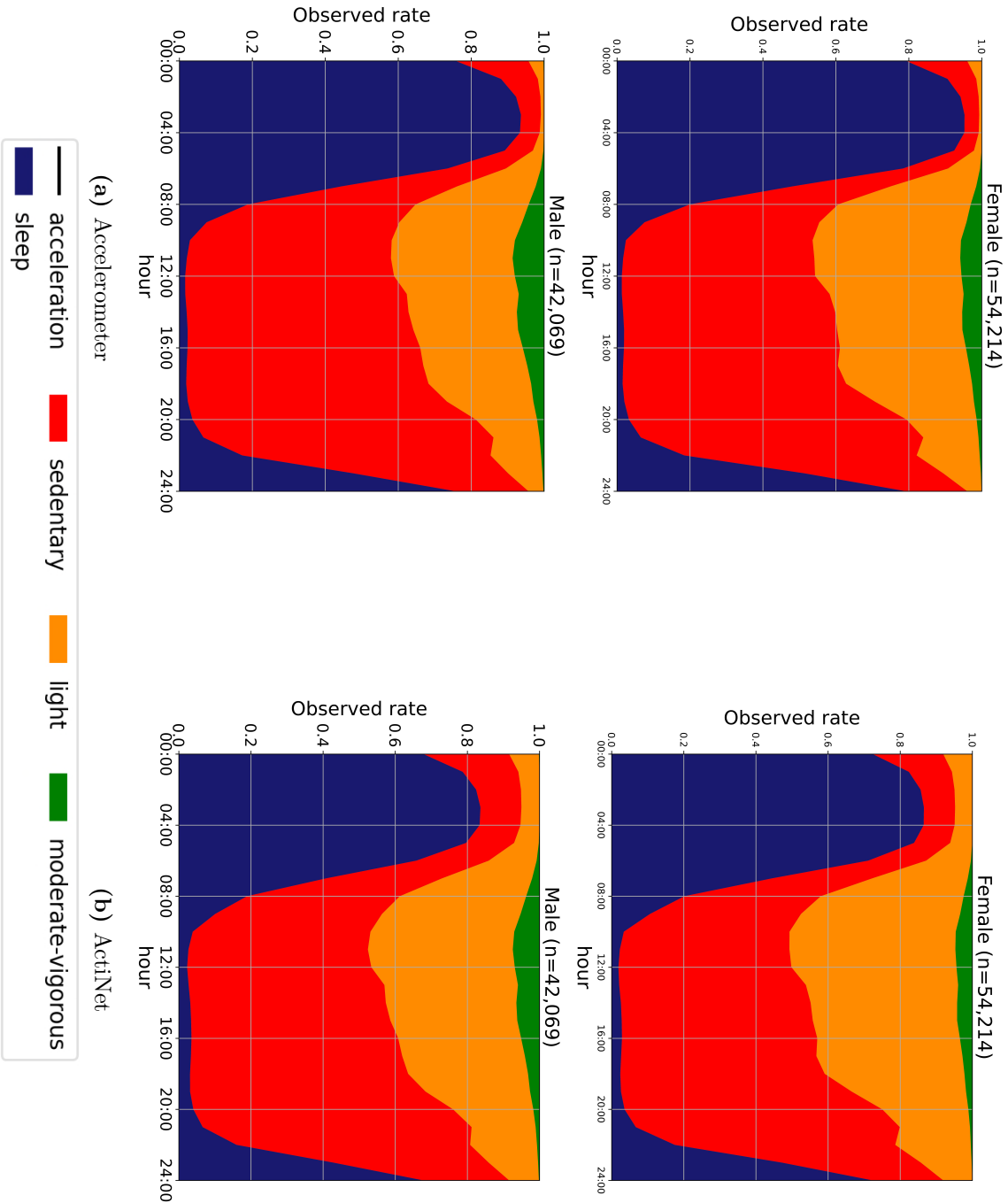


Figure A.10: Observed rate of activities from predictions made by the deployment of Accelerometer and ActiNet on UK Biobank sex based subpopulations.

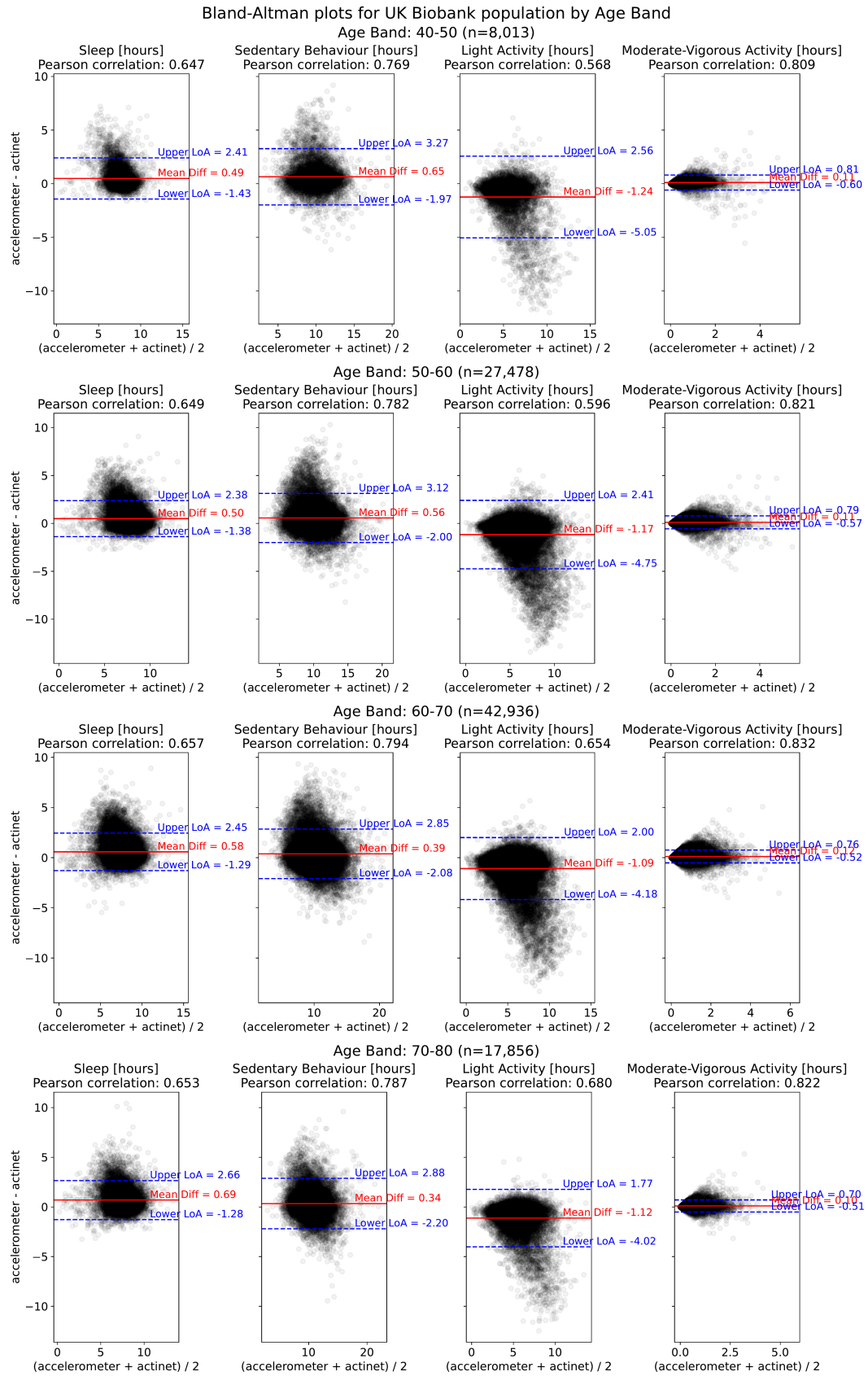


Figure A.11: Bland-Altman plots comparing model prediction outputs of `pypi:accelerometer` and `pypi:actinet` for different Age Bands of UK Biobank physical activity monitoring population.

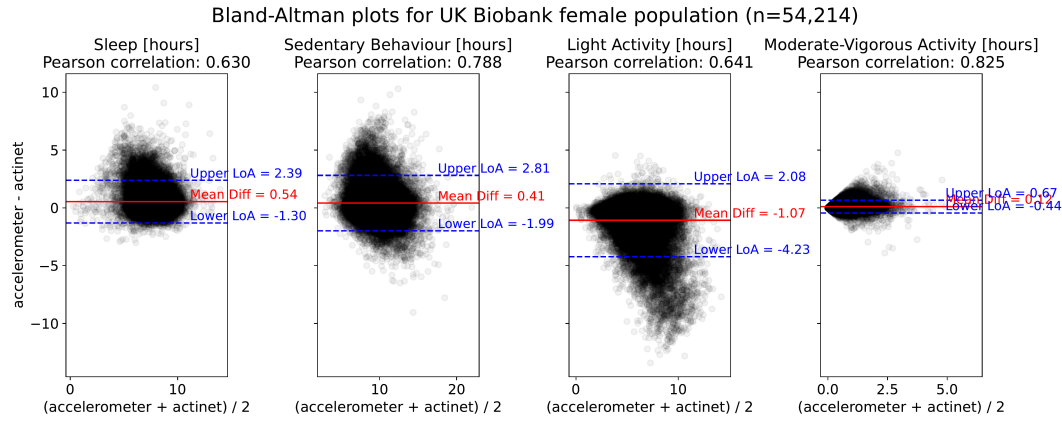


Figure A.12: Bland-Altman plots comparing model prediction outputs of `pypi:accelerometer` and `pypi:actinet` for the UK Biobank physical activity monitoring female population.

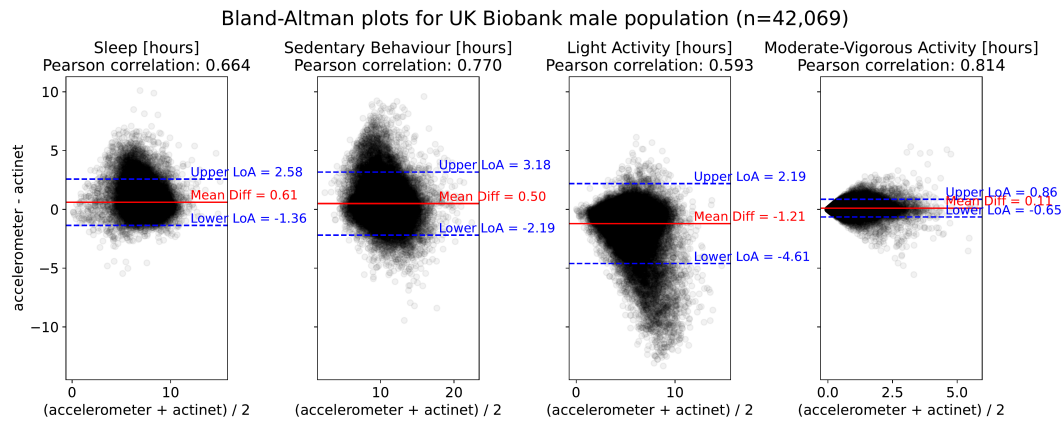


Figure A.13: Bland-Altman plots comparing model prediction outputs of `pypi:accelerometer` and `pypi:actinet` for the UK Biobank physical activity monitoring male population.

B

Median daily step count and incident
Parkinson's disease

Characteristic	Source	Notes	UK Biobank field	Coding Notes
OUTCOME				
Age at first PD event	Hospital episode statistics (HES) records	First ICD-10 code (G20) in HES data.	Derived from 2000	
Age loss-to-follow up	Death Registry		Derived from 100093	
EXPOSURE				
Median Daily Step count	Accelerometry	Derived using the "stepcount" tool, version 2.1.5. Details can be found at: https://github.com/OxWearables/stepcount .		Reported either as divided into quintiles: <6,276, 6,276-8,145, 8,146-9,966, 9,967-12,368, 12,369+, or as per-1000 daily steps
EXCLUSION VARIABLES				
Prior Parkinsonism	HES Records	PD-related ICD-10 codes in hospital records before accelerometry wear date: G20-26, G35, G47.	Derived from 2000	
	Baseline	PD-related self-reported disease.	Derived from 20002	
	Algorithmically derived first instance of PD (G20)	Date G20 first reported (Parkinson's disease) prior to the accelerometry wear date.	Derived from 131022	
ADJUSTMENT VARIABLES				
Age	Baseline	Attained age was the underlying timescale in survival analyses; participants entered the study at the end of accelerometer wear.	Derived from 90011, 34, 52	

Characteristic	Source	Notes	UK Biobank field	Coding Notes
Sex	Baseline		Derived from 31	Female, Male
Ethnicity	Baseline		Derived from 21000	White, non-white
Townsend Deprivation Index	Baseline	Townsend Deprivation Index of address at time of UKB baseline assessment. Raw value used in model, most deprived UK population quintile from 2011 census presented in Table 4.1	Derived from 22189	UK population quintiles from 2011 census: <-2.928, [-2.928, -1.509), [-1.509, 0.219), [0.219, 2.516), ≥2.516 [72]
UK recruitment centre geographic region	Baseline		Derived from 54	London, Midlands, Yorkshire, northeast, northwest, southeast, southwest England, Scotland, Wales
Education/Qualifications	Baseline	Selected as highest educational achievement. Labels are ordered by highest to lowest value.	Derived from 6138	College/University degree, NVQ/HND/HNC or equivalent, Other professional qualification, A levels/AS levels or equivalent, GCSEs/CSEs/O levels or equivalent, No qualification [73]
Employment status	Baseline		Derived from 6142, 20119	Employed, Not employed
Smoking status	Baseline		Derived from 20116	Never, previous, current
Alcohol consumption	Baseline		Derived from 1558	Never, <3 times per week, ≥3 times per week

Characteristic	Source	Notes	UK Biobank field	Coding Notes
Coffee intake	Baseline		Derived from 1498	Non-drinker, Drinker - up to 2 cups, Drinker - more than 2 cups
ADJUSTMENT VARIABLES – SENSITIVITY ANALYSIS				
Body mass index	Baseline		Derived from 21001	
Average sleep duration	Accelerometry		Derived from 40046	
Type 2 Diabetes	HES Records	ICD-10 code (E11) in HES data prior to end of accelerometer wear.	Derived from 2000	
	Baseline	Type 2 Diabetes self-reported disease at baseline.	Derived from 20002, 2443	
Depression	HES Records	ICD-10 code (F32, 33) in HES data prior to end of accelerometer wear.	Derived from 2000	
	Baseline	Depression self-reported disease at baseline.	Derived from 20002, 2090, 2100	
Constipation	HES Records	ICD-10 code (K59.0) in HES data prior to accelerometer wear.	Derived from 2000	
	Baseline	Use of laxatives in self-reported medication use at baseline.	Derived from 6154	
Bladder dysfunction	HES Records	ICD-10 code (N31) in HES data prior to accelerometer wear.	Derived from 2000	
	Baseline	Incontinence self-reported disease at baseline.	Derived from 20002	

Characteristic	Source	Notes	UK Biobank field	Coding Notes
Neurological conditions	HES Records	Any of the following ICD10 codes in HES data prior to accelerometer wear: A80, A80.[0-4,8,9], D32, D36, E1[0-4].4, F00, F00.[0-2,9], F01, F01.[0-3,8,9], F02, F02.[0-4,8], F03, F05, F05.1, G0[0-4,7], G12, G25.0, G3[0-2, 5-7], G30.[0,1,8,9], G31.1, G4[0,3,4], G5[0-3,7], G50.[0,1,8,9], G51.0, G6[0-2], G61.0, G70.0, G72, G8[0-4], G82.[2-4], G83.[0-4,8,9], G9[0-9], G93.[0-9], G96.[0,1,8,9], H53, H81, H81.[0-4,8,9], H83, H9[0,1,3], H91.[0-3,8,9], H93.[0-3,8,9], I67, I67.[0-9], K56, M79, M79.[0-4,6-9], N08, N08.[0,8,9], N13, N13.[0-9], Q05, R29.0, R41.81, S0[0-9], S[1-9]4, S[1-3]4.0, S24.1, S34.1, S34.3, T0[4,6,7], T14, T24, T84, T90, T90.[3-5,8,9], T91, T91.3	Derived from 2000	

Characteristic	Source	Notes	UK Biobank field	Coding Notes
	Baseline	Any of the following self-reported diseases at baseline: infection of nervous system, brain abscess/intracranial abscess, encephalitis, meningitis, spinal abscess, cranial nerve problem/palsy, bell's palsy/facial nerve palsy, trigeminal neuralgia, spinal cord disorder, paraplegia, spina bifida, peripheral nerve disorder, peripheral neuropathy, acute infective polyneuritis/guillain-barre syndrome, trapped nerve/compressed nerve, diabetic neuropathy/ulcers, chronic/degenerative neurological problem, motor neurone disease, myasthenia gravis, multiple sclerosis, parkinsons, dementia, alzheimers, cognitive impairment, other demyelinating disease, epilepsy, migraine, cerebral palsy, other neurological problem, headaches, benign / essential tremor, polio / poliomyelitis, meningioma / benign meningeal tumour, benign neuroma, neurological injury/trauma, head injury, spinal injury, peripheral nerve injury	Derived from 20002	

Characteristic	Source	Notes	UK Biobank field	Coding Notes
DESCRIPTIVE VARIABLES – DESCRIPTIVE TABLES				
Wear season	UK Biobank	Winter, Spring, Summer, Autumn	Derived from 90001	

Table B.1: Definition of variables from the UK Biobank data for the association between median daily step count and incident Parkinson's disease.

Section	Subsection	Summary	Code	Pages
Title and Abstract	Title and Abstract	Indicate study's design in title/abstract	1a	74
Title and Abstract	Title and Abstract	Provide informative summary of study	1b	N/A
Introduction	Background/rationale	Explain scientific background and rationale	2	74
Introduction	Objectives	State specific objectives and hypotheses	3	74
Methods	Study design	Present key elements of study design	4	75
Methods	Setting	Describe setting, locations, and relevant dates	5	75
Methods	Participants	Give eligibility criteria and selection methods	6a	76
Methods	Participants	Provide matching criteria for matched studies	6b	N/A
Methods	Variables	Define outcomes, exposures, confounders, and effect modifiers	7	76, 76
Methods	Data sources measurement	Give sources and methods of data assessment	8	76, 76
Methods	Bias	Describe efforts to address bias	9	76
Methods	Study size	Explain how study size was determined	10	N/A
Methods	Quantitative variables	Explain handling of quantitative variables	11	75, 76
Methods	Statistical methods	Describe all statistical methods used	12a	76

Section	Subsection	Summary	Code	Pages
Methods	Statistical methods	Describe methods for examining subgroups and interactions	12b	76
Methods	Statistical methods	Explain how missing data were addressed	12c	76
Methods	Statistical methods	Explain how loss to follow-up was addressed	12d	76
Methods	Statistical methods	Describe any sensitivity analyses	12e	76
Results	Participants	Report numbers at each stage of the study	13a	76
Results	Participants	Give reasons for non-participation	13b	N/A
Results	Participants	Consider use of a flow diagram	13c	78
Results	Descriptive data	Provide characteristics of study participants	14a	81
Results	Descriptive data	Indicate number of participants with missing data	14b	81
Results	Descriptive data	Summarise follow-up time	14c	80
Results	Outcome data	Report numbers of outcome events or summary measures	15	80
Results	Main results	Give unadjusted and adjusted estimates with precision	16a	80, 82, 164
Results	Main results	Report category boundaries when continuous variables were categorized	16b	159
Results	Main results	Translate estimates of relative risk into absolute risk if relevant	16c	N/A
Results	Other analyses	Report other analyses done	17	74
Discussion	Key results	Summarise key results with reference to objectives	18	74
Discussion	Limitations	Discuss the limitations of the study	19	74
Discussion	Interpretation	Provide overall interpretation of results	20	74
Discussion	Generalisability	Discuss the generalisability of the study results	21	74

Section	Subsection	Summary	Code	Pages
Other information	Funding	Give the source of funding and role of funders	22	153

Table B.2: Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines checklist for reporting of findings with this work. N/A indicates checklist item is not applicable to the contents of this thesis.

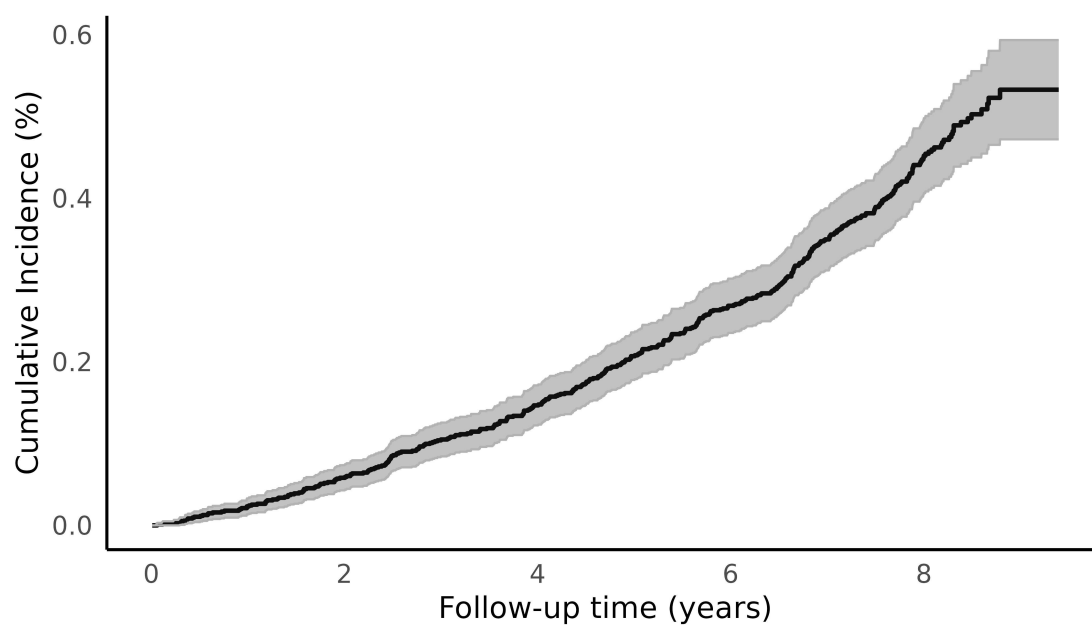


Figure B.1: Kaplan-Meier plot for cumulative incidence of Parkinson's disease in UK Biobank Physical Activity Monitoring population.

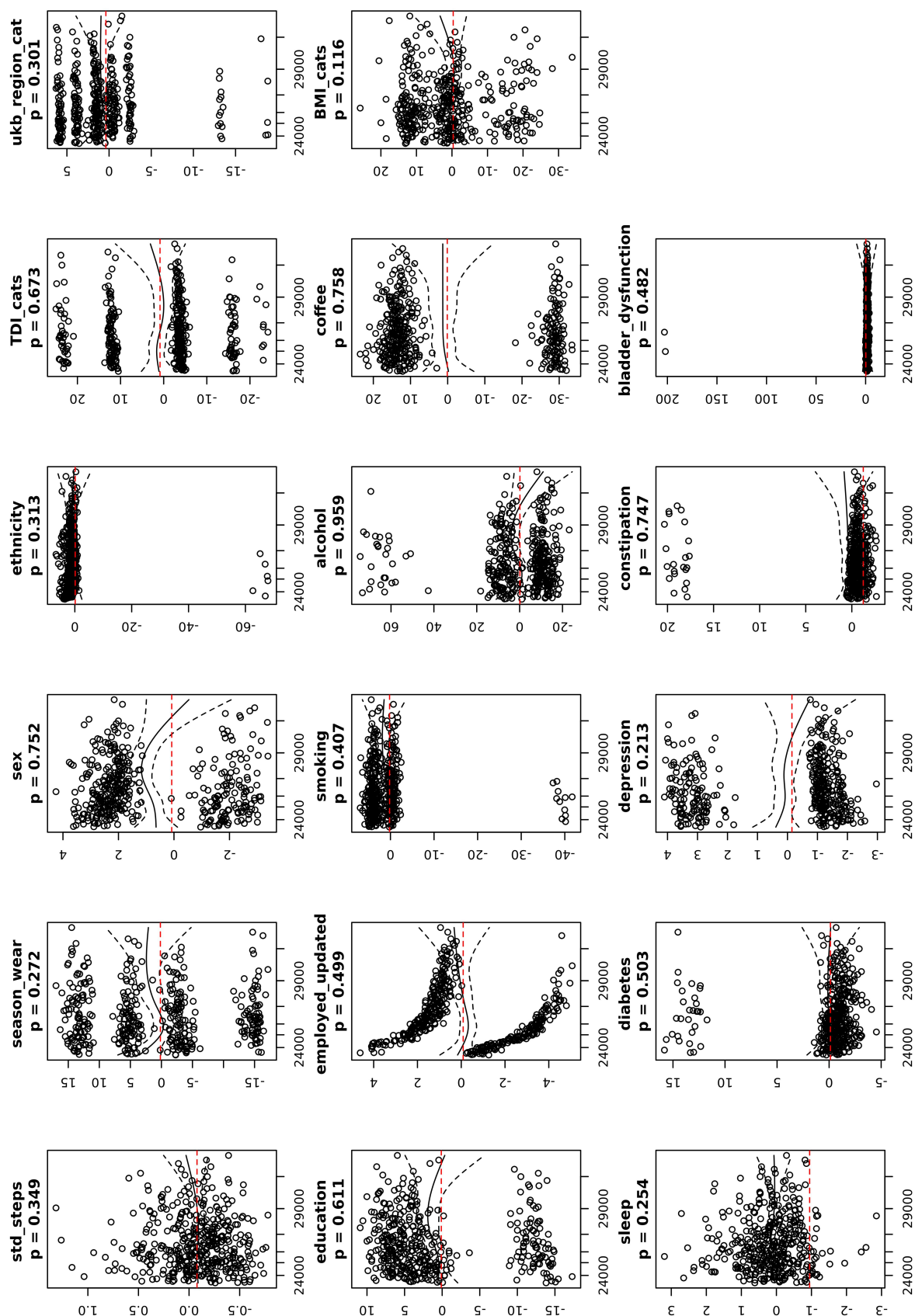


Figure B.2: Plot for Schoenfeld residuals to ensure proportional hazards assumption is met.

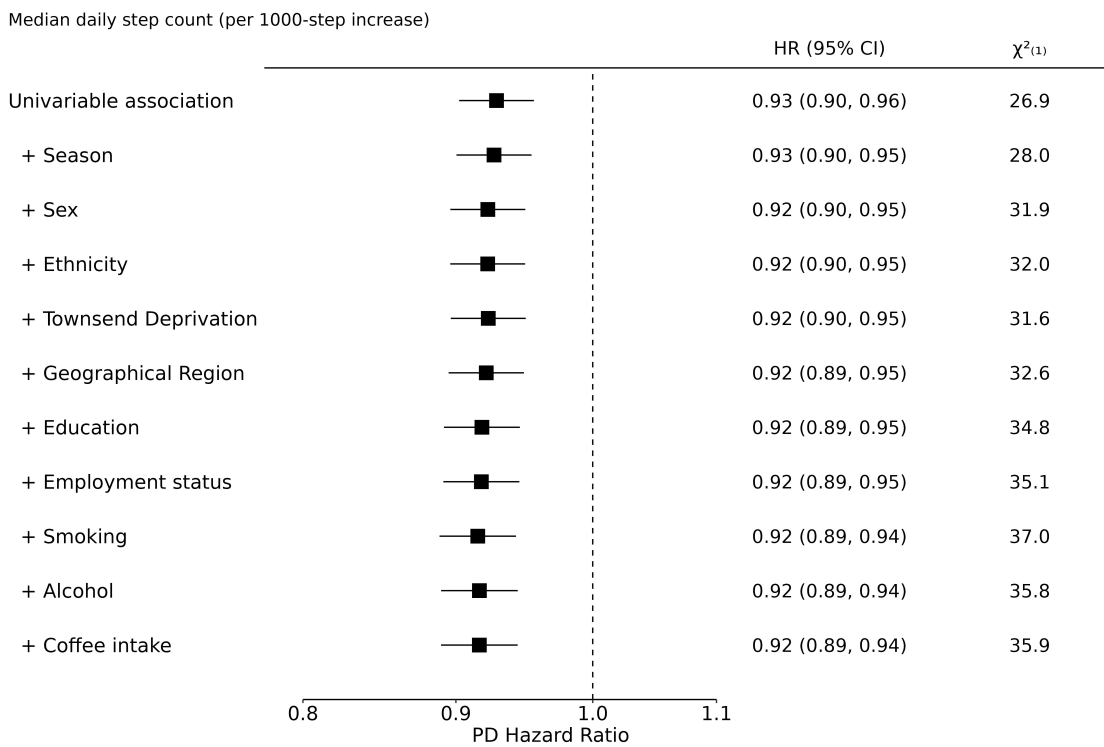


Figure B.3: The sequential adjustment of the association between a per 1000-step increase in imputed median daily step count and incident Parkinson's disease with age as timescale using all periods of follow-up. HR=Hazard ratio, CI=Confidence interval, $\chi^2_{(1)}$ =chi squared using likelihood ratio test for one degree of freedom.

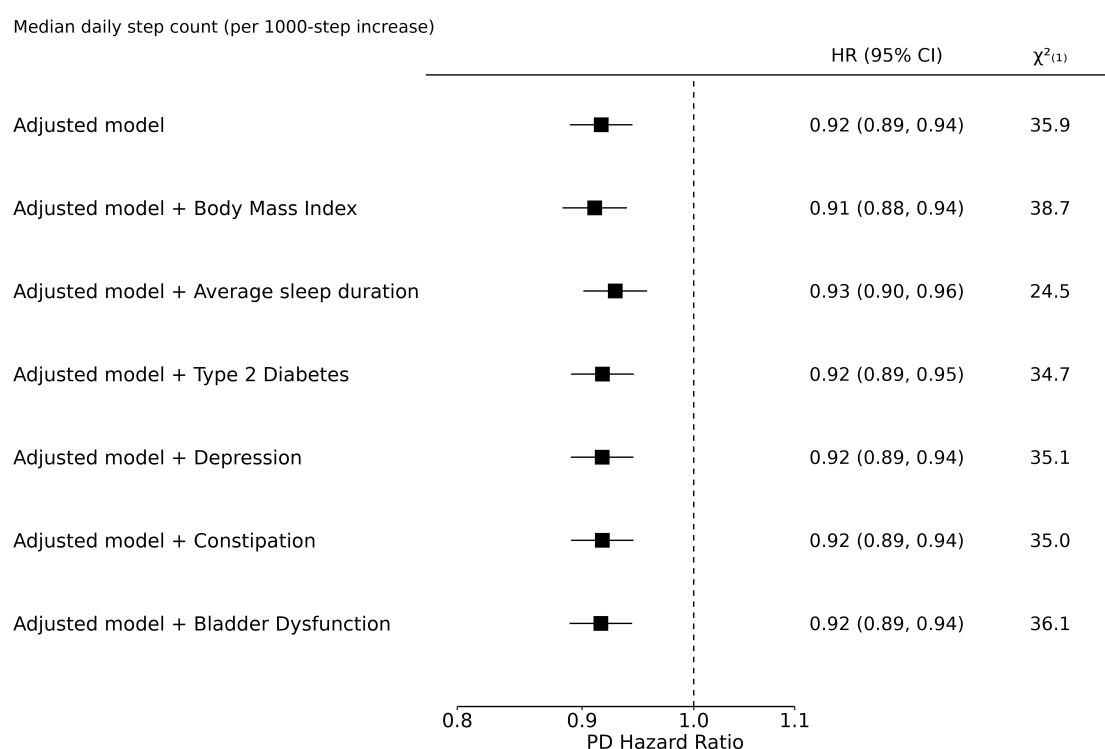


Figure B.4: The association between a per 1000-step increase in imputed median daily step count and incident Parkinson's disease with age as timescale using all periods of follow-up after adjustment for potential confounding factors. The adjusted model is adjusted for season, sex, ethnicity, Townsend deprivation, geographic region, education, employment status, smoking, alcohol and coffee intake. HR=Hazard ratio, CI=Confidence interval, $\chi^2_{(1)}$ =chi squared using likelihood ratio test for one degree of freedom.

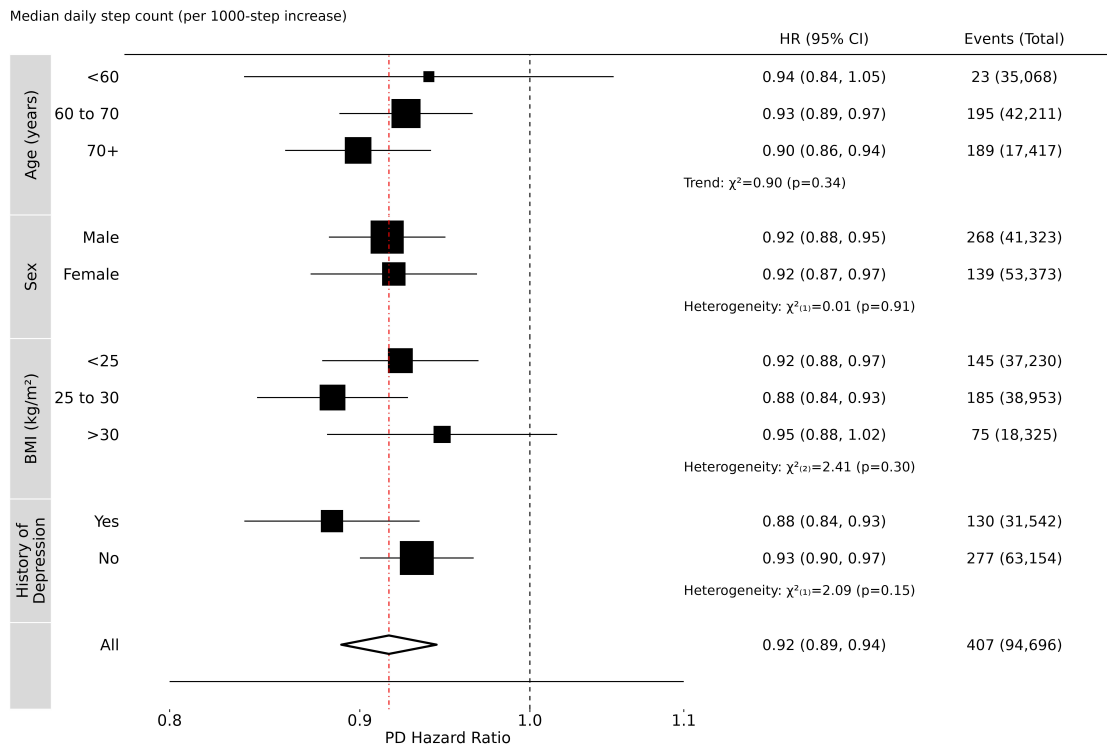


Figure B.5: The association between a per 1000-step increase in imputed median daily step count and incident Parkinson's disease with age as timescale using all periods of follow-up within various subpopulations. Models are adjusted for season, sex (except for sex-based sub-populations), ethnicity, Townsend deprivation, geographic region, education, employment status, smoking, alcohol and coffee intake. For each category, the area of the square is inversely proportional to the log HR, which also determines the 95% CI.

BMI=Body Mass Index, HR=Hazard ratio, CI=Confidence interval.

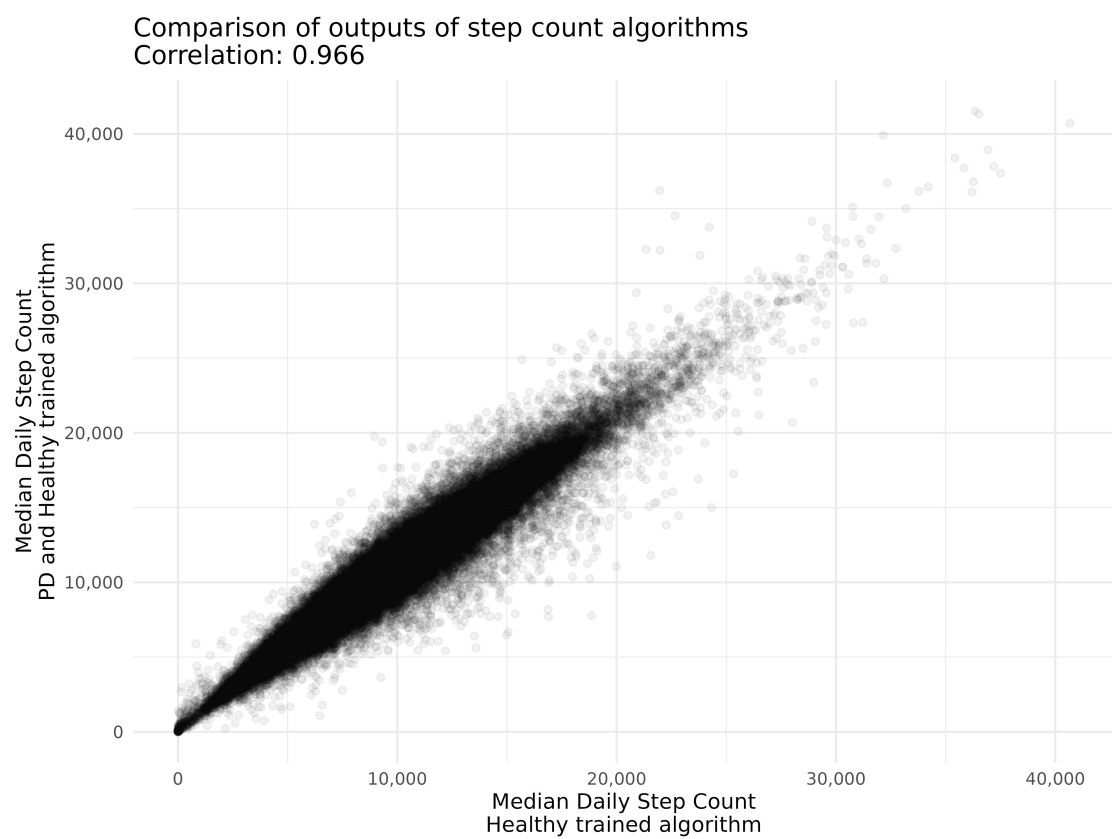


Figure B.6: Comparison of median daily step count output by step count algorithms trained only on healthy participants, and trained on both healthy and Parkinson's disease participants.

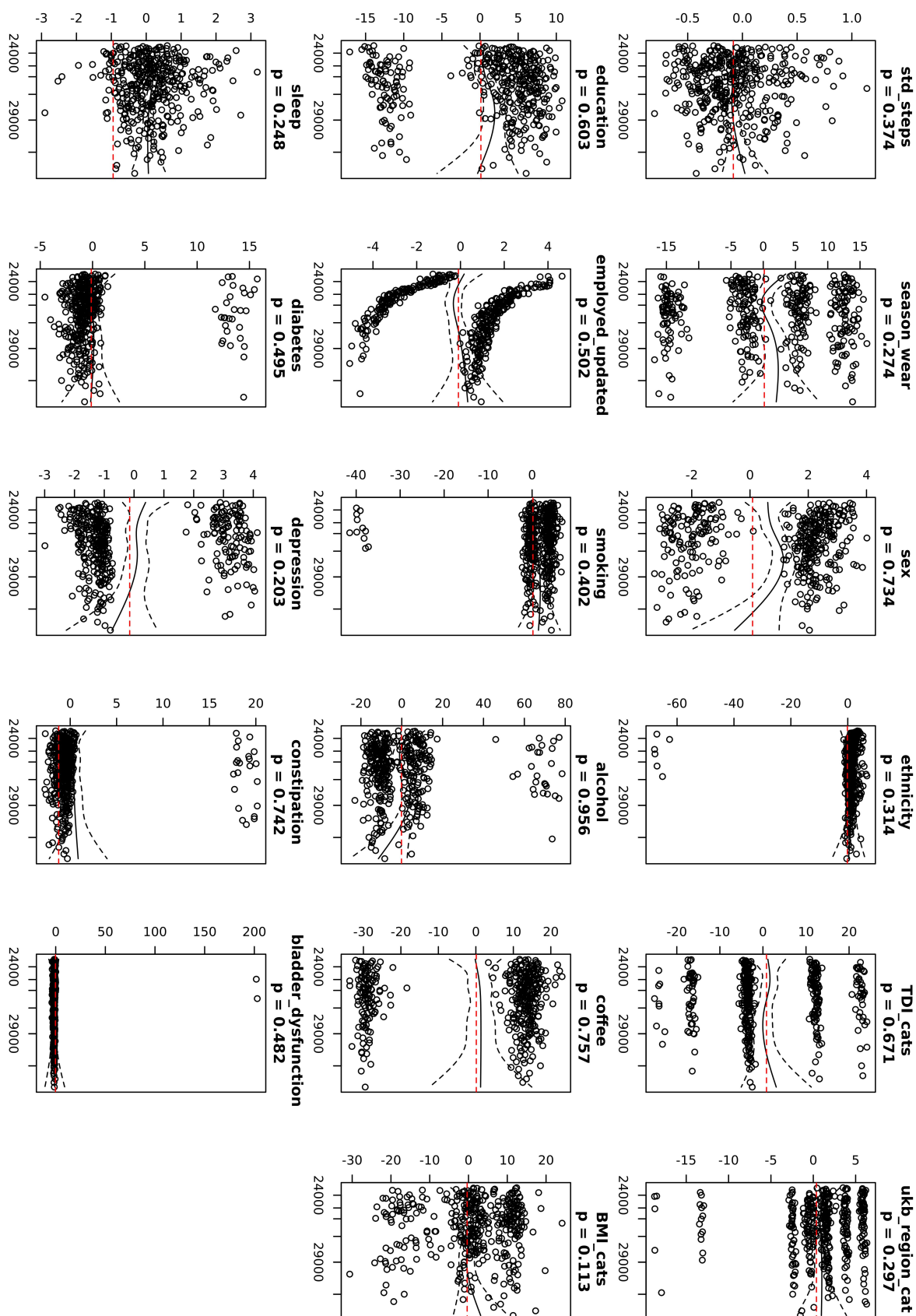


Figure B.7: Plot for schoenfeld residuals to ensure proportional hazards assumption is met when using PD trained algorithm.

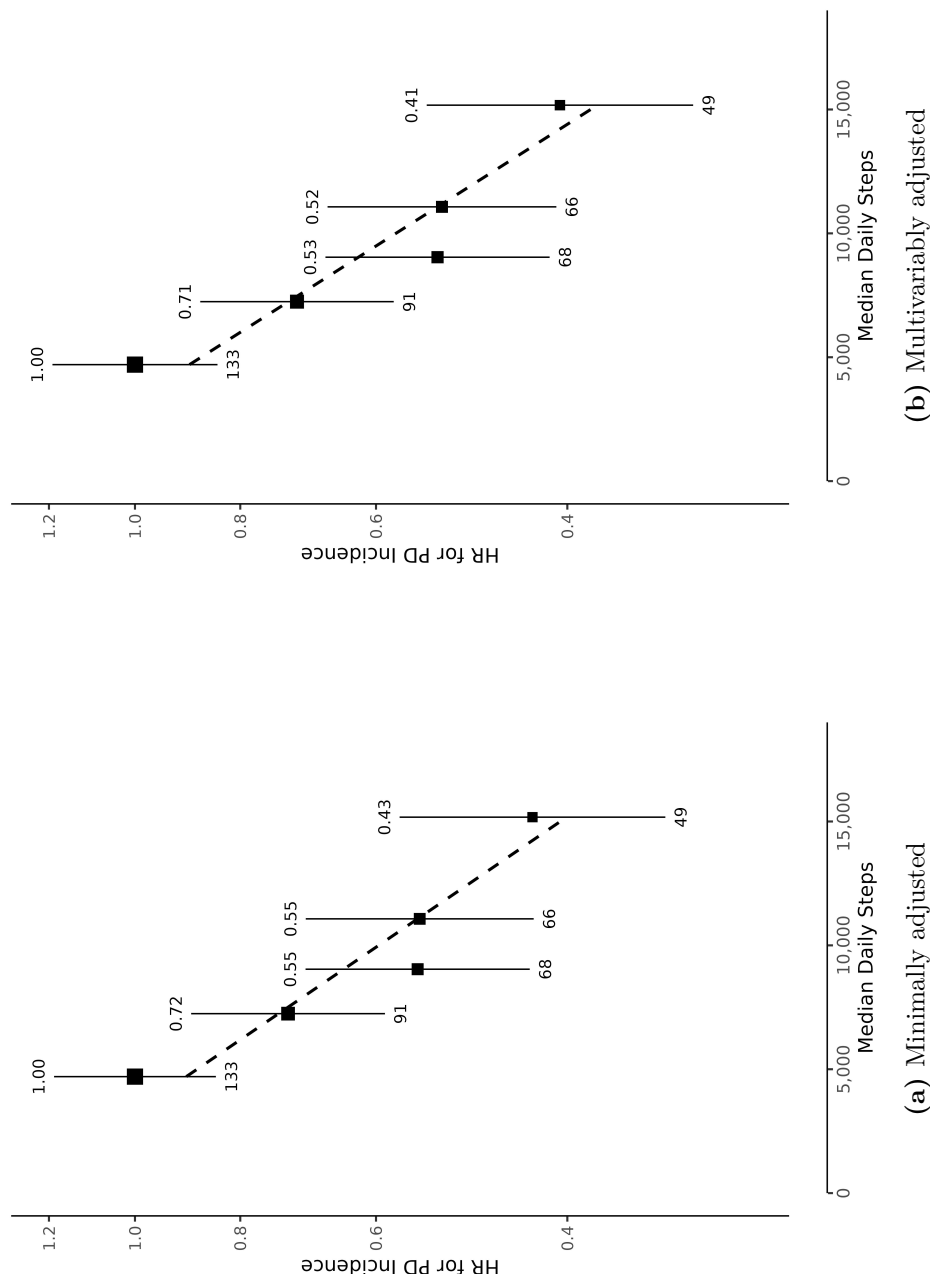


Figure B.8: Association between quintiles of median daily step count using PD trained algorithm and incident Parkinson's disease. Hazard ratios (HR) and 95% confidence intervals were calculated using age as a timescale. HR is above and the number of events is plotted below each data point. a) Minimally adjustment model, adjusted only for season of wear and sex. b) Multivariably adjusted model, adjusted for season of wear, sex, ethnicity, Townsend deprivation index, geographic region, education, employment, alcohol intake, smoking status and coffee intake. The HR (95% CI) per 1,000 increase in daily steps for the multivariably adjusted model is 0.91 (0.89 - 0.94). For each category, the area of the square is inversely proportional to the log HR, which also determines the 95% CI.

PD-trained algorithm sensitivity analysis

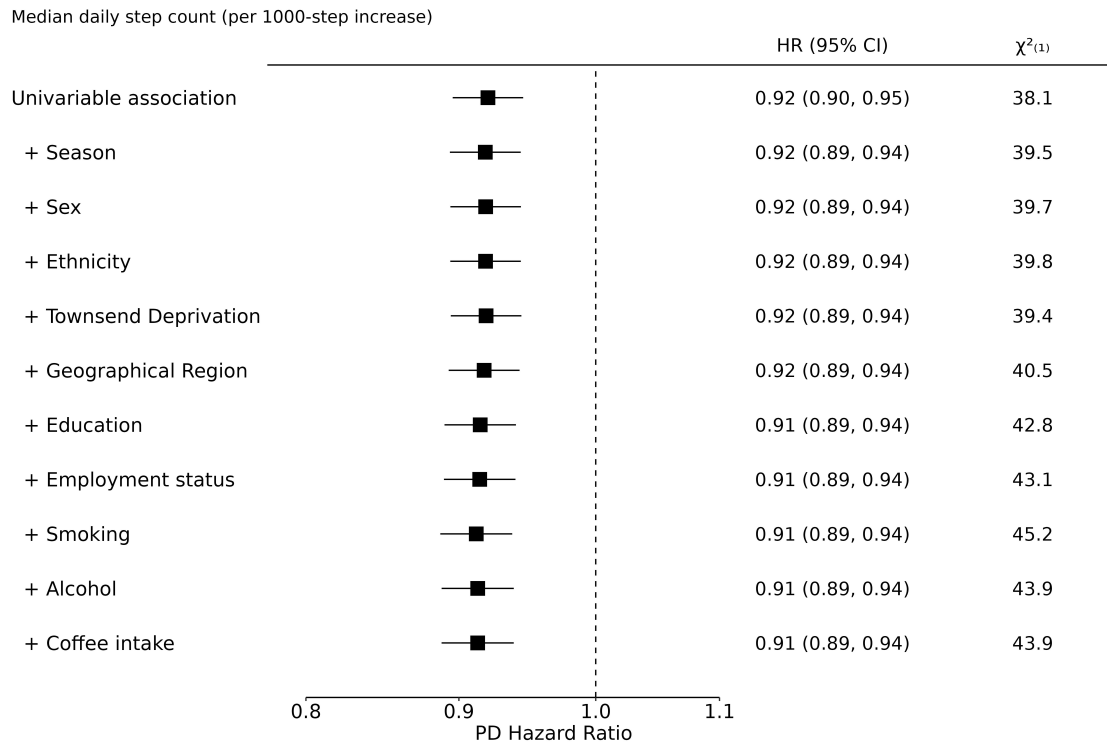


Figure B.9: The sequential adjustment of the association between a per 1000-step increase in imputed median daily step count using PD trained algorithm and incident Parkinson's disease, with age as timescale using all periods of follow-up. HR=Hazard ratio, CI=Confidence interval, $\chi^2_{(1)}$ =chi squared using likelihood ratio test for one degree of freedom, to test significance of median daily step count term.

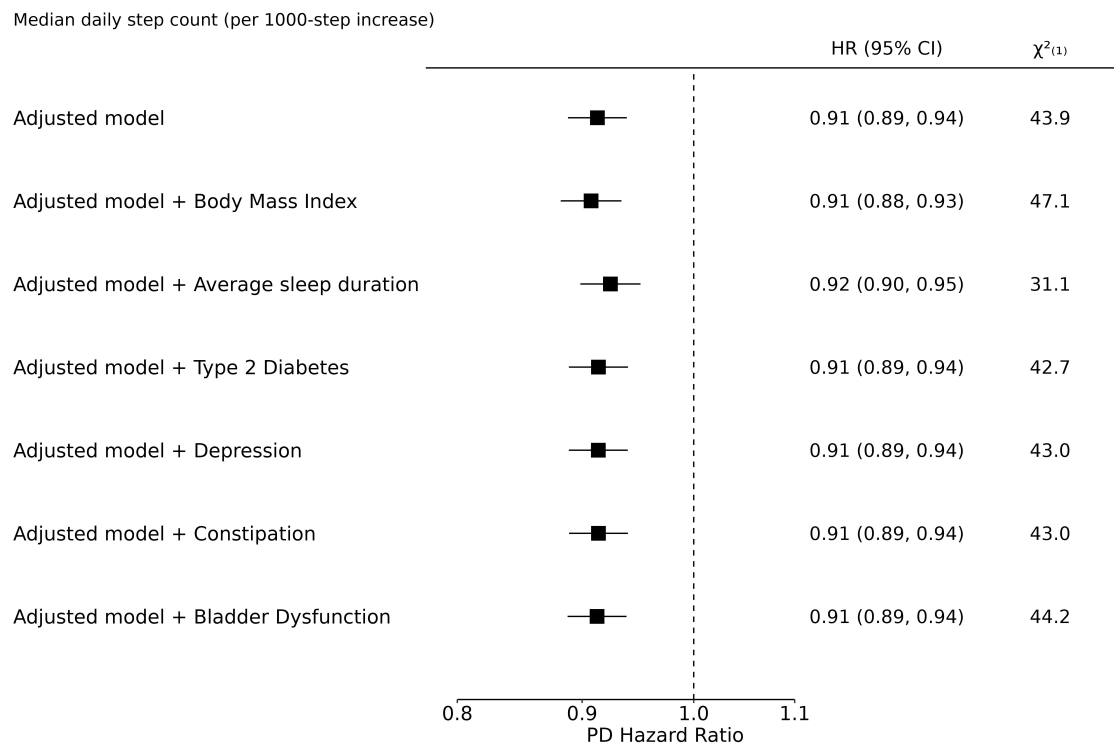


Figure B.10: The association between a per 1000-step increase in imputed median daily step count using PD trained algorithm and incident Parkinson's disease with age as timescale using all periods of follow-up after adjustment for potential confounding factors. The adjusted model is adjusted for season, sex, ethnicity, Townsend deprivation, geographic region, education, employment status, smoking, alcohol and coffee intake. HR=Hazard ratio, CI=Confidence interval, $\chi^2_{(1)}$ =chi squared using likelihood ratio test for one degree of freedom, to test significance of median daily step count term.

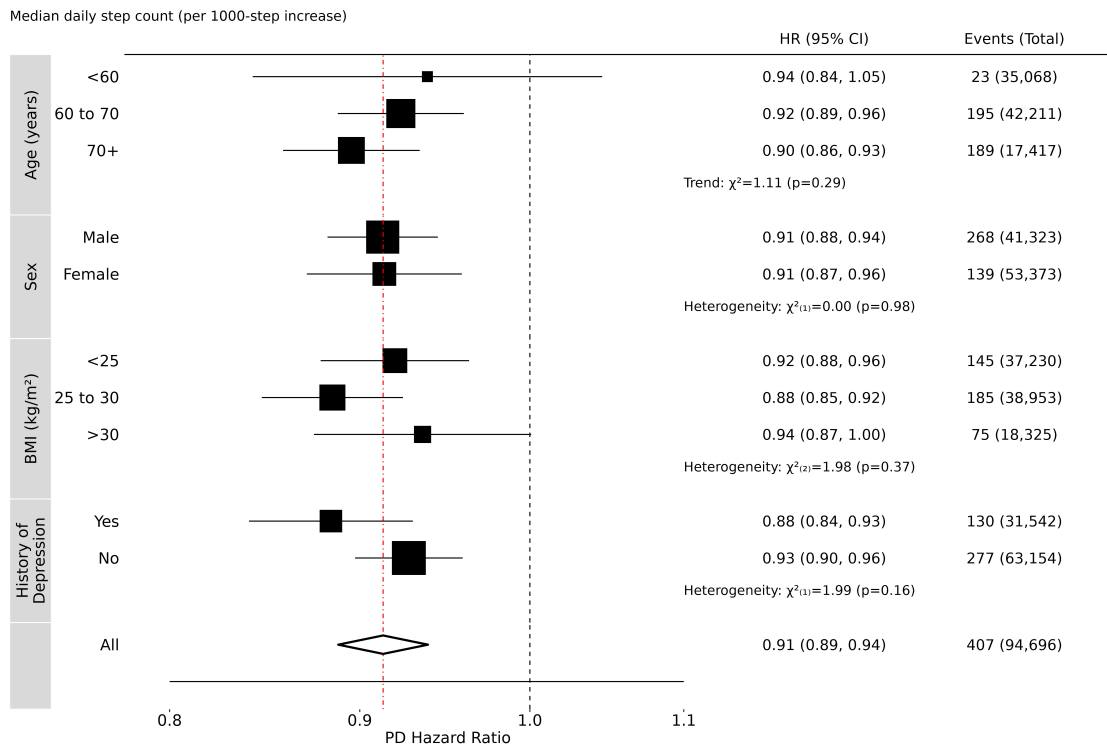


Figure B.11: The association between a per 1000-step increase in imputed median daily step count using PD trained algorithm and incident Parkinson's disease with age as timescale using all periods of follow-up within various subpopulations. Models are adjusted for season, sex (except for sex-based sub-populations), ethnicity, Townsend deprivation, geographic region, education, employment status, smoking, alcohol and coffee intake. For each category, the area of the square is inversely proportional to the log HR, which also determines the 95% CI.

BMI=Body Mass Index, HR=Hazard ratio, CI=Confidence interval.

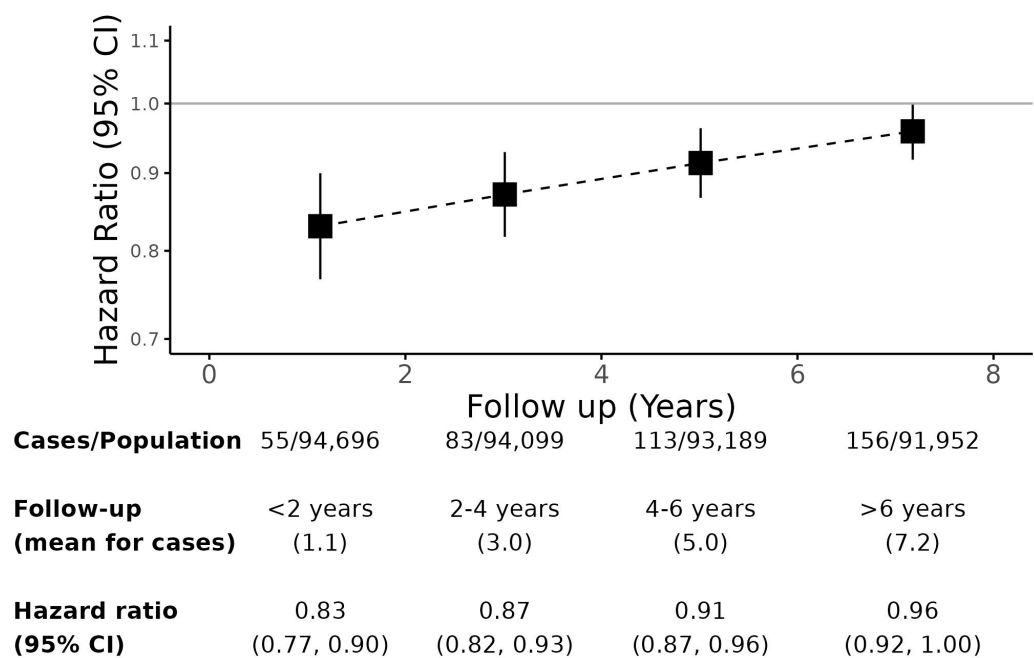


Figure B.12: Hazard ratio for incident Parkinson's disease by period of follow-up per 1000-step increase in daily steps using PD trained algorithm. Models use age as timescale, and are adjusted for season of wear, sex, Townsend deprivation index, geographic region of assessment centre, educational attainment, employment, smoking, alcohol and coffee consumption. CI=confidence interval.

B.1 Power Calculations

For the subset of the population with at least 6 years of follow-up, of which 156 incident PD cases were found, an expected hazard ratio of 0.942 per 1000-step increase in median daily steps would be required to achieve a power of 0.800 or greater. This was calculated using the default function for calculating power for continuous exposures (`powerEpiCont.default`) from the `powerSurvEpi` R package[178].

C

Parkinson's disease risk prediction

Characteristic	Source	Notes	UK Biobank field	Coding Notes
OUTCOME/EXCLUSION VARIABLES				
Parkinson's disease	HES records	PD-related ICD-10 codes in hospital records before accelerometer wear date: G20-26, G35, G47	Derived from 2000	
	Baseline	PD-related self-reported disease	Derived from 20002	
	Algorithmically derived first instance of PD (G20)	Date G20 first reported (Parkinson's disease) prior to the accelerometer wear date	Derived from 131022	
Age loss-to-follow up	Death Registry		Derived from 100093	
PREDICTOR VARIABLES				
Age	Baseline	Attained age was the underlying timescale in survival analyses; participants entered the study at the end of accelerometer wear.	Derived from 90011, 34, 52	
Sex	Baseline		31	Female, Male
Smoking status	Baseline		Derived from 20116	Never, previous, current
Alcohol consumption	Baseline		Derived from 1558	Less than once a week, Once a week or more
Coffee intake	Baseline		Derived from 1498	One a day or less, More than once a day
Depression	HES records	ICD-10 code (F32, 33) in HES data prior to end of accelerometer wear.	Derived from 2000	
Constipation	Baseline	Self-reported disease at baseline	Derived from 20002, 2090, 2100	

Characteristic	Source	Notes	UK Biobank field	Coding Notes
	HES records	ICD10 code (K59.0) in HES data prior to accelerometer wear	Derived from 2000	
Polygenic risk score	Genomics	Polygenic score for PD for all UK Biobank participants	26260	
Overall activity	Physical activity measurement	Euclidean norm minus one, averaged over monitoring period	Derived from 90001	
Average sedentary time	Physical activity measurement	Proportion of day sedentary predicted using pypi:actinet	Derived from 90001	
Average light activity time	Physical activity measurement	Proportion of day in light activity predicted using pypi:actinet	Derived from 90001	
Average moderate-vigorous activity time	Physical activity measurement	Proportion of day in moderate-vigorous activity predicted using pypi:actinet	Derived from 90001	
PD-like gait score	Physical activity measurement	Measure of similarity to the walking of diagnosed PD participants of the PPMI-Verily study	Derived from 90001	

Table C.1: Variables used in Parkinson's disease risk prediction analysis.

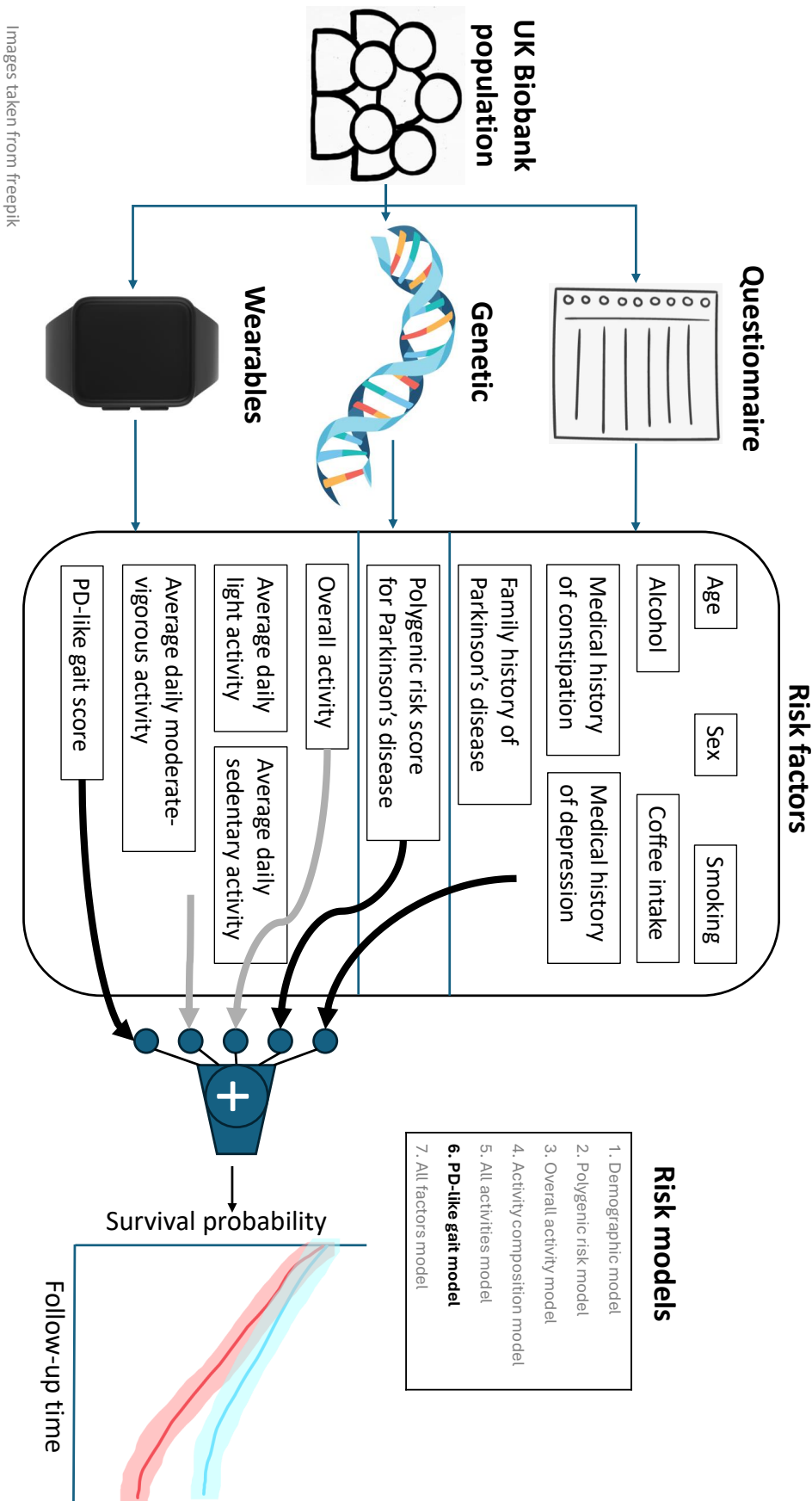


Figure C.1: Illustration of Parkinson's disease risk analysis in UK Biobank physical activity monitoring population.

Images taken from freepik

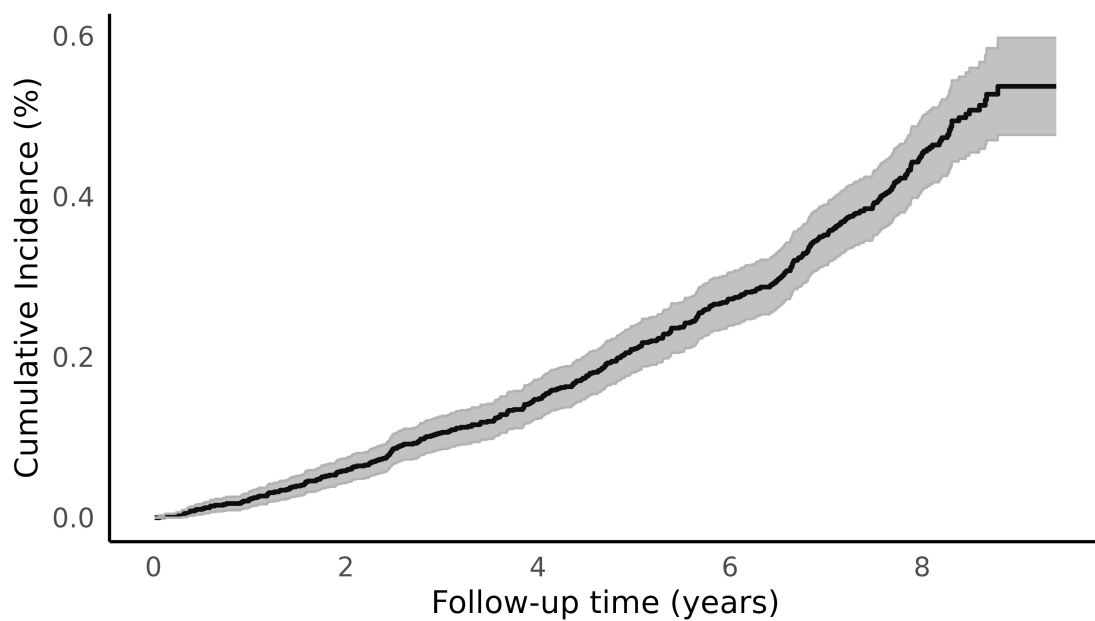


Figure C.2: Kaplan-Meier plot for incidence probability for Parkinson's disease risk analysis in UK Biobank physical activity monitoring population.

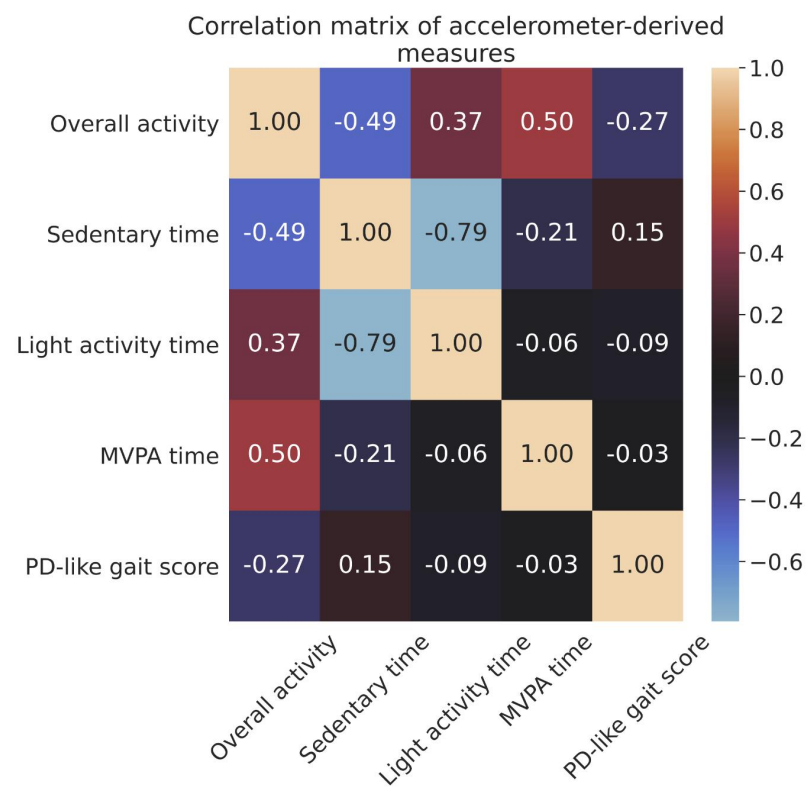


Figure C.3: Correlation of accelerometer-derived predictors for Parkinson's disease risk analysis.

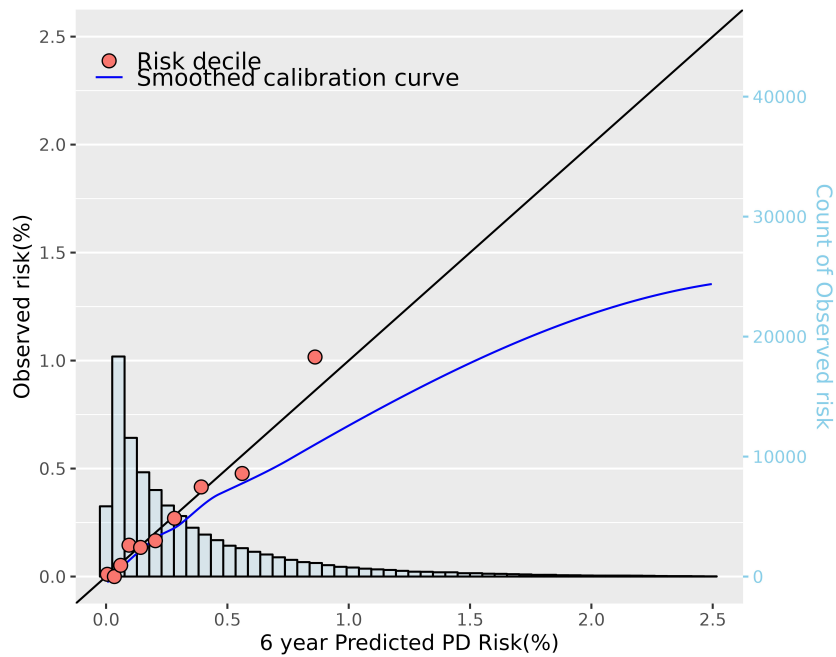


Figure C.4: Calibration curve for Parkinson's disease risk analysis using demographic factors; age, sex, alcohol, smoking, coffee, depression, constipation and family history of PD.

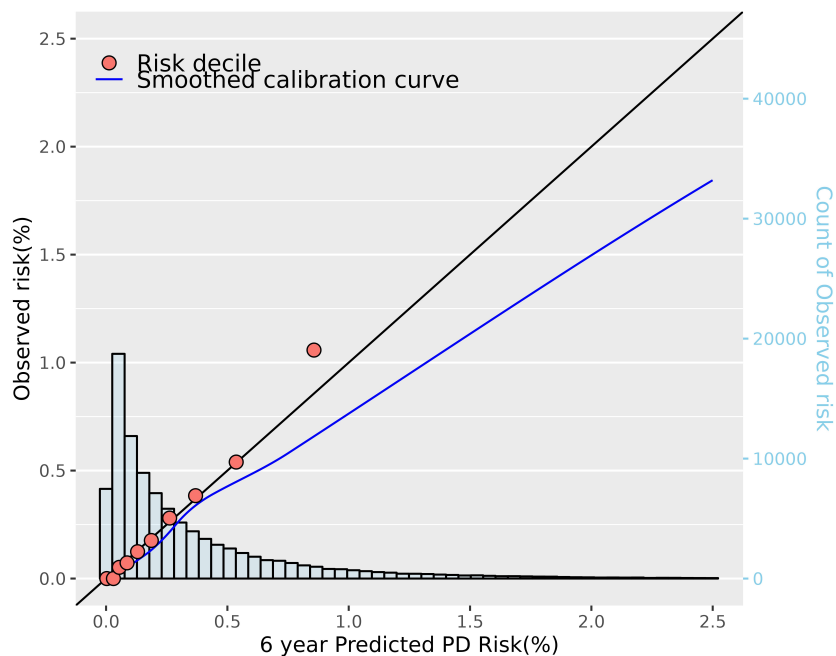


Figure C.5: Calibration curve for Parkinson's disease risk analysis using demographic and genetic factors; age, sex, alcohol, smoking, coffee, depression, constipation, family history of PD and polygenic risk of PD.

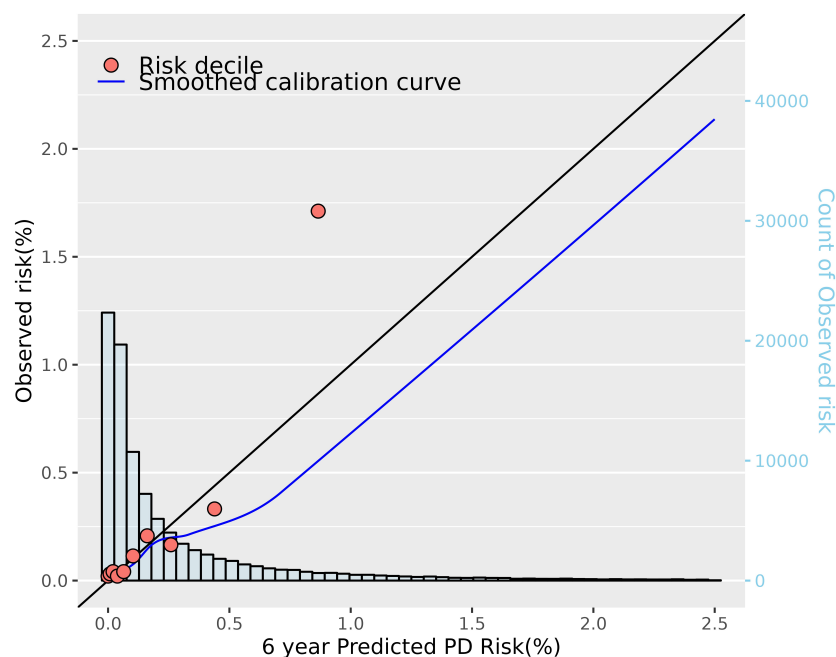


Figure C.6: Calibration curve for Parkinson's disease risk analysis using demographic, genetic and accelerometer-derived risk factors; age, sex, alcohol, smoking, coffee, depression, constipation, family history of PD and polygenic risk of PD.

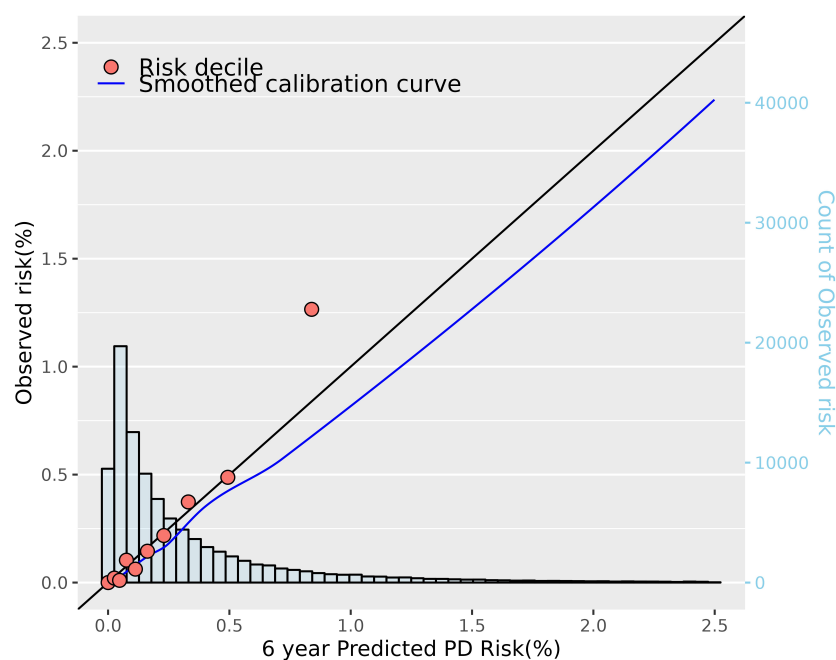


Figure C.7: Calibration curve for Parkinson's disease risk analysis using demographic, genetic and accelerometer-derived risk factors; age, sex, alcohol, smoking, coffee, depression, constipation, family history of PD and polygenic risk of PD.

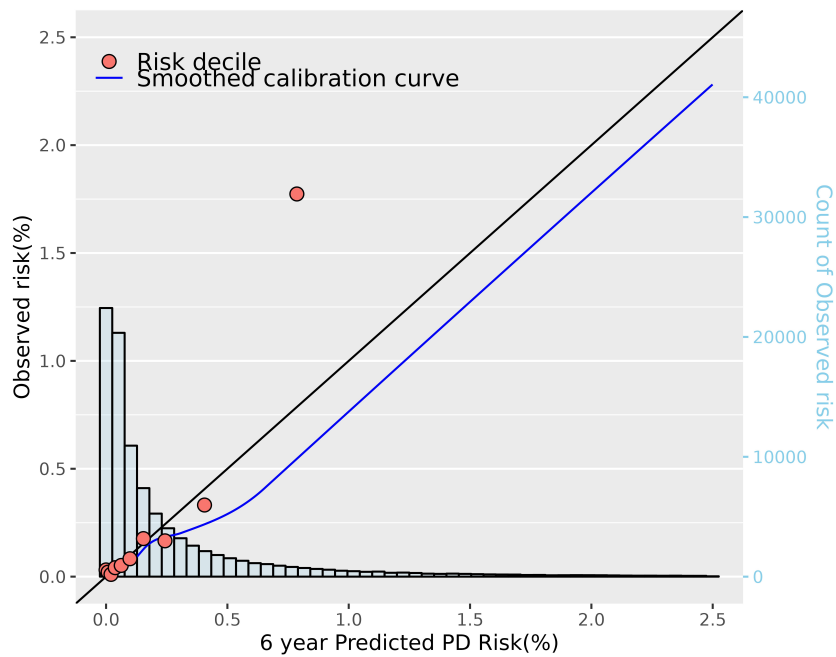


Figure C.8: Calibration curve for Parkinson's disease risk analysis using demographic, genetic and accelerometer-derived risk factors; age, sex, alcohol, smoking, coffee, depression, constipation, family history of PD, polygenic risk of PD, overall activity, average sedentary time, average light activity time and average moderate-vigorous activity time.

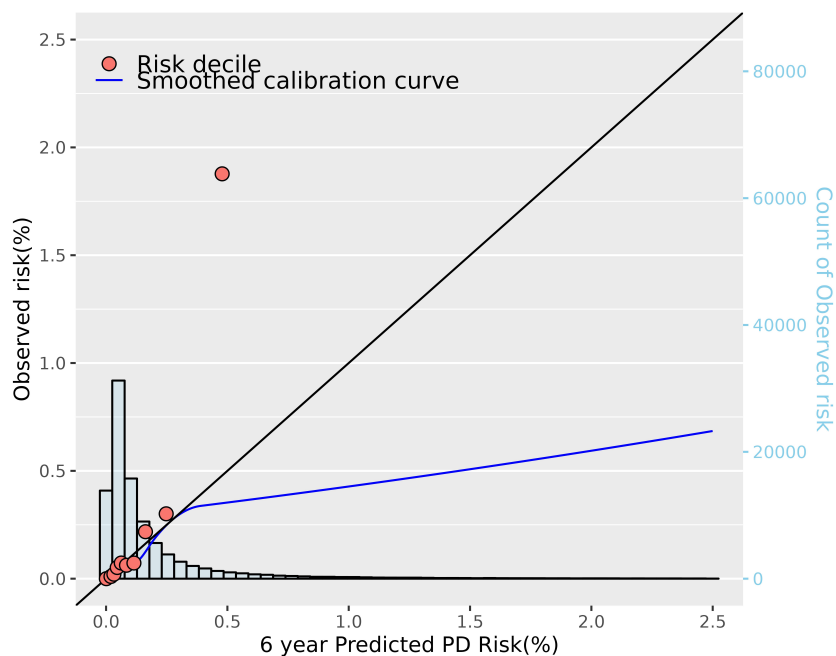


Figure C.9: Calibration curve for Parkinson's disease risk analysis using demographic, genetic and accelerometer-derived risk factors; age, sex, alcohol, smoking, coffee, depression, constipation, family history of PD, polygenic risk of PD and Parkinson's disease-like gait score.

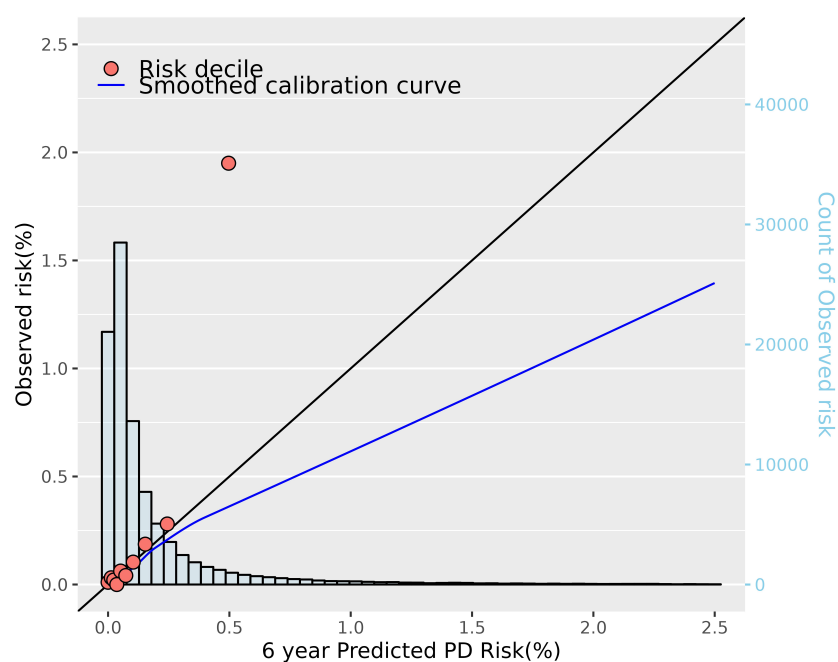


Figure C.10: Calibration curve for Parkinson's disease risk analysis using demographic, genetic and accelerometer-derived risk factors; age, sex, alcohol, smoking, coffee, depression, constipation, family history of PD, polygenic risk of PD, overall activity, average sedentary time, average light activity time, average moderate-vigorous activity time and Parkinson's disease-like gait score.

Risk factor	Beta Estimate	Standard Error	p-value
Demographic model: Baseline survival = 0.99856991850672			
Age (years)	0.140	0.011	3.63e-27 ***
Sex	1.014	0.139	4.18e-12 ***
Alcohol (Once a week or more)	-0.132	0.147	0.369
Smoking (current) smoker	-0.739	0.344	0.033 *
Previous smoker	-0.337	0.132	0.011 *
Coffee (more than once a day)	0.004	0.125	0.973
Depression	0.201	0.135	0.138
Constipation	0.145	0.310	0.641
Family history of PD	0.540	0.228	0.018 *
Polygenic model: Baseline survival = 0.998669448403106			
Age (years)	0.140	0.011	3.83e-27 ***
Sex	1.015	0.139	3.96e-12 ***
Alcohol (Once a week or more)	-0.136	0.147	0.354
Smoking (current) smoker	-0.735	0.344	0.034 *
Previous smoker	-0.327	0.132	0.014 *
Coffee (more than once a day)	0.005	0.125	0.967
Depression	0.202	0.135	0.137
Constipation	0.136	0.310	0.662
Family history of PD	0.471	0.228	0.040 *
Polygenic risk of PD	0.404	0.061	2.63e-10 ***
Overall activity model: Baseline survival = 0.999299312345833			
Age (years)	0.098	0.011	8.82e-16 ***
Sex	0.788	0.140	5.02e-8 ***
Alcohol (Once a week or more)	0.117	0.148	0.429
Smoking (current) smoker	-1.142	0.345	0.001 **
Previous smoker	-0.401	0.132	0.003 **

Risk factor	Beta Estimate	Standard Error	p-value
Coffee (more than once a day)	0.019	0.125	0.878
Depression	0.077	0.136	0.569
Constipation	-0.153	0.311	0.623
Family history of PD	0.486	0.228	0.034 *
Polygenic risk of PD	0.415	0.062	1.30e-10 ***
Overall activity (mg)	-0.171	0.012	6.49e-35 ***
Activity composition model: Baseline survival = 0.998808821396873			
Age (years)	0.128	0.011	3.69e-24 ***
Sex	1.021	0.145	2.17e-11 ***
Alcohol (Once a week or more)	-0.056	0.148	0.705
Smoking (current) smoker	-0.925	0.345	0.008 **
Previous smoker	-0.389	0.133	0.004 **
Coffee (more than once a day)	-0.003	0.126	0.981
Depression	0.139	0.136	0.307
Constipation	0.011	0.311	0.971
Family history of PD	0.470	0.228	0.041 *
Polygenic risk of PD	0.413	0.062	1.49e-10 ***
Average sedentary time	6.991	1.023	6.55e-11 ***
Average light activity time	5.656	1.042	1.38e-7 ***
Average moderate-vigorous activity time	-12.055	4.142	0.004 **
All activities model: Baseline survival = 0.999281716840156			
Age (years)	0.095	0.011	4.05e-15 ***
Sex	0.715	0.144	1.38e-16 ***
Alcohol (Once a week or more)	0.078	0.149	0.600
Smoking (current) smoker	-1.123	0.345	0.001 **
Previous smoker	-0.376	0.133	0.005 **
Coffee (more than once a day)	-0.004	0.126	0.974

Risk factor	Beta Estimate	Standard Error	p-value
Depression	0.089	0.137	0.515
Constipation	-0.172	0.312	0.583
Family history of PD	0.474	0.228	0.039 *
Polygenic risk of PD	0.406	0.062	2.80e-10 ***
Overall activity (mg)	-0.198	0.013	3.90e-39 ***
Average sedentary time	4.027	1.016	9.67e-5 ***
Average light activity time	6.771	0.876	2.79e-13 ***
Average moderate-vigorous activity time	13.296	2.654	1.05e-6 ***
Gait score model: Baseline survival = 0.999244300866579			
Age (years)	0.074	0.011	5.51e-11 ***
Sex	0.553	0.140	1.01e-4 ***
Smoking (current)	-1.310	0.347	0.0002 ***
Previous smoker	-0.313	0.133	0.019 *
Family history of PD	0.423	0.228	0.065
Coffee (more than once a day)	0.143	0.126	0.259
Alcohol (Once a week or more)	0.094	0.147	0.525
Constipation	-0.159	0.312	0.610
Depression	0.064	0.137	0.638
Polygenic risk of PD	0.361	0.061	1.14e-8 ***
PD-like gait score	8.522	0.261	1.54e-91 ***
All features model: Baseline survival = 0.999382353705463			
Age (years)	0.061	0.011	3.81e-8 ***
Sex	0.591	0.145	7.27e-5 ***
Smoking (current)	-1.332	0.350	0.0002 ***
Previous smoker	-0.278	0.133	0.037 *
Family history of PD	0.418	0.228	0.067
Coffee (more than once a day)	0.113	0.126	0.372

Risk factor	Beta Estimate	Standard Error	p-value
Alcohol (Once a week or more)	0.082	0.148	0.582
Constipation	-0.208	0.313	0.503
Depression	0.081	0.137	0.556
Polygenic risk of PD	0.368	0.063	1.81e-8 ***
Overall activity (mg)	-0.099	0.013	3.24e-14 ***
Average sedentary time	3.954	1.018	1.39e-4 ***
Average light activity time	5.948	0.885	8.27e-12 ***
Average moderate-vigorous activity time	-9.661	2.529	1.78e-4 ***
PD-like gait score	7.505	0.307	2.14e-67 ***

Table C.2: Parameters for trained Cox regression PD risk models. * for p-values <0.05, ** <0.01, *** <0.001, indicating significance

Section/Topic	Item	Development/ evaluation	Checklist item	Pages
TITLE				
<i>Title</i>	1	D;E	Identify the study as developing or evaluating the performance of a multivariable prediction model, the target population, and the outcome to be predicted	107
ABSTRACT				
<i>Abstract</i>	2	D;E	See TRIPOD+AI for Abstracts checklist	N/A
INTRODUCTION				
<i>Background</i>	3a	D;E	Explain the healthcare context and rationale for developing or evaluating the prediction model, including references to existing models	107
	3b	D;E	Describe the target population and the intended purpose of the prediction model	107
	3c	D;E	Describe any known health inequalities between sociodemographic groups	107
<i>Objectives</i>	4	D;E	Specify the study objectives, including whether the study describes the development or validation of a prediction model	109
METHODS				
<i>Data</i>	5a	D;E	Describe the sources of data separately for the development and evaluation datasets, the rationale for using these data	
	5b	D;E	Specify the dates of the collected participant data, including start and end of participant accrual	109
<i>Participants</i>	6a	D;E	Specify key elements of the study setting, including the number and location of centres	
	6b	D;E	Describe the eligibility criteria for study participants	109
	6c	D;E	Give details of any treatments received and how they were handled during model development or evaluation	109

Section/Topic	Item	Development/ evaluation	Checklist item	Pages
<i>Data preparation</i>	7	D;E	Describe any data pre-processing and quality checking	109
	8a	D;E	Clearly define the outcome that is being predicted and the time horizon, including how and when assessed	109, 177
	8b	D;E	If outcome assessment requires subjective interpretation, describe the qualifications and demographic characteristics of the outcome assessors	N/A
	8c	D;E	Report any actions to blind assessment of the outcome to be predicted	N/A
<i>Predictors</i>	9a	D	Describe the choice of initial predictors and any pre-selection of predictors before model building	110
	9b	D;E	Clearly define all predictors, including how and when they were measured	110, 177
	9c	D;E	If predictor measurement requires subjective interpretation, describe the qualifications and demographic characteristics of the predictor assessors	N/A
<i>Sample size</i>	10	D;E	Explain how the study size was arrived at, and justify that the study size was sufficient	109
<i>Missing data</i>	11	D;E	Describe how missing data were handled. Provide reasons for omitting any data	112
<i>Analytical methods</i>	12a	D	Describe how the data were used in the analysis, including whether the data were partitioned	112
	12b	D	Describe how predictors were handled in the analyses (functional form, rescaling, transformation, etc.)	112
	12c	D	Specify the type of model, rationale, all model-building steps, and method for internal validation	112
	12d	D;E	Describe if and how heterogeneity in estimates of model parameter values was handled	112

Section/Topic	Item	Development/ evaluation	Checklist item	Pages
	12e	D;E	Specify all measures and plots used to evaluate model performance, and, if relevant, to compare models	112
	12f	E	Describe any model updating arising from the evaluation	N/A
	12g	E	For model evaluation, describe how the model predictions were calculated	112
<i>Class imbalance</i>	13	D;E	If class imbalance methods were used, describe how this was done and any methods to recalibrate the model	N/A
<i>Fairness</i>	14	D;E	Describe any approaches used to address model fairness and their rationale	N/A
<i>Model output</i>	15	D	Specify the output of the prediction model, including any classification or thresholds	187

Table C.3: TRIPOD-AI Checklist for PD risk prediction analysis in UK Biobank.

References

- [1] George DeMaagd and Ashok Philip. “Parkinson’s Disease and Its Management: Part 1: Disease Entity, Risk Factors, Pathophysiology, Clinical Presentation, and Diagnosis”. In: *P & T : a peer-reviewed journal for formulary management* 40.8 (2015), pp. 504–532. ISSN: 1052-1372. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4517533/>.
- [2] Antonina Kouli, Kelli M. Torsney, and Wei-Li Kuan. “Parkinson’s Disease: Etiology, Neuropathology, and Pathogenesis”. In: *Parkinson’s Disease: Pathogenesis and Clinical Aspects*. Ed. by Thomas B. Stoker and Julia C. Greenland. Brisbane (AU): Codon Publications, 2018. ISBN: 978-0-9944381-6-4. pmid: 30702842. URL: <http://www.ncbi.nlm.nih.gov/books/NBK536722/> (visited on 07/17/2024).
- [3] Fatemeh N. Emamzadeh and Andrei Surguchov. “Parkinson’s Disease: Biomarkers, Treatment, and Risk Factors”. In: *Frontiers in Neuroscience* 12 (2018). DOI: 10.3389/fnins.2018.00612. pmid: 30214392. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6125353/> (visited on 06/02/2024).
- [4] Rana Zia Ur Rehman et al. “Selecting Clinically Relevant Gait Characteristics for Classification of Early Parkinson’s Disease: A Comprehensive Machine Learning Approach”. In: *Scientific Reports* 9.1 (1 Nov. 21, 2019), p. 17269. ISSN: 2045-2322. DOI: 10.1038/s41598-019-53656-7. URL: <https://www.nature.com/articles/s41598-019-53656-7> (visited on 05/30/2022).
- [5] E. Ray Dorsey et al. “The Emerging Evidence of the Parkinson Pandemic”. In: *Journal of Parkinson’s Disease* 8 (Suppl 1 2018), S3–S8. ISSN: 1877-7171. DOI: 10.3233/JPD-181474. pmid: 30584159.
- [6] N. Maserejian, L. Vinikoor-Imler, and A. Dilley. *Estimation of the 2020 Global Population of Parkinson’s Disease (PD)*. MDS Abstracts. 2020. URL: <https://www.mdsabstracts.org/abstract/estimation-of-the-2020-global-population-of-parkinsons-disease-pd/> (visited on 05/30/2022).
- [7] Parkinson’s UK. *Incidence and Prevalence of Parkinson’s Disease: Summary Report*. Parkinson’s UK, 2018. URL: <https://www.parkinsons.org.uk/sites/default/files/2018-01/CS2960%20Incidence%20and%20prevalence%20report%20branding%20summary%20report.pdf>.
- [8] Parkinson’s UK. *Prevalence and Incidence of Parkinson’s Disease: Detailed Report*. Parkinson’s UK, 2018. URL: https://www.parkinsons.org.uk/sites/default/files/2018-01/Prevalence%20%20Incidence%20Report%20Latest_Public_2.pdf.

- [9] Ronald B. Postuma and Daniela Berg. “Advances in Markers of Prodromal Parkinson Disease”. In: *Nature Reviews Neurology* 12.11 (2016), pp. 622–634. ISSN: 1759-4766. DOI: 10.1038/nrneurol.2016.152. URL: <https://doi.org/10.1038/nrneurol.2016.152>.
- [10] Rajan S and Kaas B. “Parkinson’s Disease: Risk Factor Modification and Prevention”. In: *Seminars in neurology* 42.5 (Oct. 2022). ISSN: 1098-9021. DOI: 10.1055/s-0042-1758780. pmid: 36427528. URL: <https://pubmed.ncbi.nlm.nih.gov/36427528/> (visited on 06/02/2024).
- [11] Eduardo Tolosa et al. “Challenges in the Diagnosis of Parkinson’s Disease”. In: *The Lancet Neurology* 20.5 (May 1, 2021), pp. 385–397. ISSN: 1474-4422, 1474-4465. DOI: 10.1016/S1474-4422(21)00030-2. pmid: 33894193. URL: [https://www.thelancet.com/journals/lanneur/article/PIIS1474-4422\(21\)00030-2/abstract](https://www.thelancet.com/journals/lanneur/article/PIIS1474-4422(21)00030-2/abstract) (visited on 09/01/2024).
- [12] *Parkinson’s Disease: Challenges, Progress, and Promise / National Institute of Neurological Disorders and Stroke*. National Institute of Neurological Disorders and Stroke. Jan. 30, 2023. URL: <https://www.ninds.nih.gov/current-research/focus-disorders/parkinsons-disease-research/parkinsons-disease-challenges-progress-and-promise> (visited on 06/02/2024).
- [13] Melissa J. Armstrong and Michael S. Okun. “Time for a New Image of Parkinson Disease”. In: *JAMA Neurology* 77.11 (Nov. 1, 2020), p. 1345. ISSN: 2168-6149. DOI: 10.1001/jamaneurol.2020.2412. URL: <https://jamanetwork.com/journals/jamaneurology/fullarticle/2768703> (visited on 06/02/2024).
- [14] Christopher G. Goetz et al. “Movement Disorder Society-sponsored Revision of the Unified Parkinson’s Disease Rating Scale (MDS-UPDRS): Scale Presentation and Clinimetric Testing Results”. In: *Movement Disorders* 23.15 (Nov. 15, 2008), pp. 2129–2170. ISSN: 0885-3185. DOI: 10.1002/mds.22340. URL: <https://doi.org/10.1002/mds.22340> (visited on 08/29/2024).
- [15] *Australian Parkinson’s Genetics Study*. QIMR Berghofer. URL: <https://www.qimrberghofer.edu.au/apgs/> (visited on 07/18/2024).
- [16] Werner Poewe et al. “Levodopa in the Treatment of Parkinson’s Disease: An Old Drug Still Going Strong”. In: *Clinical Interventions in Aging* 5 (2010), pp. 229–238. ISSN: 1176-9092. pmid: 20852670. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2938030/> (visited on 07/31/2024).
- [17] Andres M. Lozano et al. “Deep Brain Stimulation: Current Challenges and Future Directions”. In: *Nature reviews. Neurology* 15.3 (Mar. 2019), pp. 148–160. ISSN: 1759-4758. DOI: 10.1038/s41582-018-0128-2. pmid: 30683913. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6397644/> (visited on 05/30/2022).
- [18] Lorraine V. Kalia and Anthony E. Lang. “Parkinson’s Disease”. In: *The Lancet* 386.9996 (Aug. 29, 2015), pp. 896–912. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(14)61393-3. pmid: 25904081. URL: [https://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(14\)61393-3/abstract](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(14)61393-3/abstract) (visited on 09/01/2024).

- [19] Amy Reeve, Eve Simcox, and Doug Turnbull. “Ageing and Parkinson’s Disease: Why Is Advancing Age the Biggest Risk Factor?” In: *Ageing Research Reviews* 14.100 (Mar. 2014), pp. 19–30. ISSN: 1568-1637. DOI: 10.1016/j.arr.2014.01.004. pmid: 24503004. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3989046/> (visited on 06/02/2024).
- [20] Silvia Cerri, Liudmila Mus, and Fabio Blandini. “Parkinson’s Disease in Women and Men: What’s the Difference?” In: *Journal of Parkinson’s Disease* 9.3 (2019), p. 501. DOI: 10.3233/JPD-191683. pmid: 31282427. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6700650/> (visited on 06/02/2024).
- [21] E. Ray Dorsey and Bastiaan R. Bloem. “Parkinson’s Disease Is Predominantly an Environmental Disease”. In: *Journal of Parkinson’s Disease* 14.3 (Apr. 23, 2024), pp. 451–465. ISSN: 1877-7171. DOI: 10.3233/JPD-230357. pmid: 38217613. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11091623/> (visited on 09/01/2024).
- [22] Alexandra Zirra et al. “Gender Differences in the Prevalence of Parkinson’s Disease”. In: *Movement Disorders Clinical Practice* 10.1 (Jan. 2023), p. 86. DOI: 10.1002/mdc3.13584. pmid: 36699001. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9847309/> (visited on 06/02/2024).
- [23] Sara Bandres-Ciga et al. “Genetics of Parkinson’s Disease: An Introspection of Its Journey towards Precision Medicine”. In: *Neurobiology of Disease* 137 (Apr. 1, 2020), p. 104782. ISSN: 0969-9961. DOI: 10.1016/j.nbd.2020.104782. URL: <https://www.sciencedirect.com/science/article/pii/S0969996120300577> (visited on 06/02/2024).
- [24] Benjamin Meir Jacobs et al. “Parkinson’s Disease Determinants, Prediction and Gene–Environment Interactions in the UK Biobank”. In: *Journal of Neurology, Neurosurgery & Psychiatry* 91.10 (Oct. 1, 2020), p. 1046. DOI: 10.1136/jnnp-2020-323646. URL: <http://jnnp.bmj.com/content/91/10/1046.abstract>.
- [25] Kenneth N. Rose, Michael A. Schwarzschild, and Stephen N. Gomperts. “Clearing the Smoke: What Protects Smokers from Parkinson’s Disease?” In: *Movement Disorders* 39.2 (2024), pp. 267–272. ISSN: 1531-8257. DOI: 10.1002/mds.29707. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/mds.29707> (visited on 07/18/2024).
- [26] Chao Wang et al. “Association between Cigarette Smoking and Parkinson’s Disease: A Neuroimaging Study”. In: *Therapeutic Advances in Neurological Disorders* 15 (Apr. 18, 2022), p. 17562864221092566. ISSN: 1756-2856. DOI: 10.1177/17562864221092566. pmid: 35464739. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9019319/> (visited on 07/18/2024).

- [27] Xiangpeng Ren and Jiang-Fan Chen. “Caffeine and Parkinson’s Disease: Multiple Benefits and Emerging Mechanisms”. In: *Frontiers in Neuroscience* 14 (Dec. 17, 2020), p. 602697. ISSN: 1662-4548. DOI: 10.3389/fnins.2020.602697. pmid: 33390888. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7773776/> (visited on 07/18/2024).
- [28] Marlies van Nimwegen et al. “Physical Inactivity in Parkinson’s Disease.” In: *Journal of neurology* 258.12 (Dec. 2011), pp. 2214–2221. ISSN: 1432-1459 0340-5354. DOI: 10.1007/s00415-011-6097-7. pmid: 21614433.
- [29] Xuexian Fang et al. “Association of Levels of Physical Activity With Risk of Parkinson Disease: A Systematic Review and Meta-analysis.” In: *JAMA network open* 1.5 (Sept. 7, 2018), e182421. ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2018.2421. pmid: 30646166.
- [30] Ann D. Cohen et al. “Neuroprotective Effects of Prior Limb Use in 6-hydroxydopamine-treated Rats: Possible Role of GDNF”. In: *Journal of Neurochemistry* 85.2 (Apr. 2003), pp. 299–305. ISSN: 0022-3042, 1471-4159. DOI: 10.1046/j.1471-4159.2003.01657.x. URL: <https://onlinelibrary.wiley.com/doi/10.1046/j.1471-4159.2003.01657.x> (visited on 01/16/2024).
- [31] Shawne A. Neeper et al. “Physical Activity Increases mRNA for Brain-Derived Neurotrophic Factor and Nerve Growth Factor in Rat Brain”. In: *Brain Research* 726.1-2 (July 1996), pp. 49–56. ISSN: 00068993. DOI: 10.1016/0006-8993(96)00273-9. URL: <https://linkinghub.elsevier.com/retrieve/pii/0006899396002739> (visited on 01/16/2024).
- [32] Lisanne J. Dommershuijsen, Agnita J. W. Boon, and M. Kamran Ikram. “Probing the Pre-diagnostic Phase of Parkinson’s Disease in Population-Based Studies”. In: *Frontiers in Neurology* 12 (2021). DOI: 10.3389/fneur.2021.702502. pmid: 34276552. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8284316/> (visited on 06/02/2024).
- [33] Jinhu Chen et al. “Meta-Analysis: Overweight, Obesity, and Parkinson’s Disease”. In: *International Journal of Endocrinology* 2014 (2014), pp. 1–7. ISSN: 1687-8337, 1687-8345. DOI: 10.1155/2014/203930. URL: <http://www.hindawi.com/journals/ije/2014/203930/> (visited on 09/26/2023).
- [34] Yun-Liang Wang et al. “Body Mass Index and Risk of Parkinson’s Disease: A Dose-Response Meta-Analysis of Prospective Studies”. In: *PLoS ONE* 10.6 (June 29, 2015), e0131778. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0131778. pmid: 26121579. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4488297/> (visited on 07/18/2024).
- [35] Silvana S. Bettiol et al. “Alcohol Consumption and Parkinson’s Disease Risk: A Review of Recent Findings”. In: *Journal of Parkinson’s Disease* 5.3 (2015), pp. 425–442. ISSN: 1877-718X. DOI: 10.3233/JPD-150533. pmid: 26406123.

- [36] Fereshteh Najafi et al. “Association between Socioeconomic Status and Parkinson’s Disease: Findings from a Large Incident Case–Control Study”. In: *BMJ Neurology Open* 5.1 (Feb. 15, 2023), e000386. ISSN: 2632-6140. DOI: 10.1136/bmjno-2022-000386. pmid: 36817512. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9933671/> (visited on 07/18/2024).
- [37] Annie Amjad and Gillian Granville. *Ethnicity and Parkinson’s: What We Know so Far*. Parkinson’s UK. Nov. 30, 2020. URL: <https://www.parkinsons.org.uk/research/research-blog/research-explained/ethnicity-and-parkinsons-what-we-know-so-far> (visited on 07/18/2024).
- [38] William G Dixon et al. “Charting a Course for Smartphones and Wearables to Transform Population Health Research”. In: *Journal of medical Internet research* 25 (Feb. 7, 2023), e42449. ISSN: 1439-4456. DOI: 10.2196/preprints.42449. pmid: 36749628. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7614184/> (visited on 07/29/2024).
- [39] Lydia G. Roos and George M. Slavich. “Wearable Technologies for Health Research: Opportunities, Limitations, and Practical and Conceptual Considerations”. In: *Brain, Behavior, and Immunity* 113 (Oct. 1, 2023), pp. 444–452. ISSN: 0889-1591. DOI: 10.1016/j.bbi.2023.08.008. URL: <https://www.sciencedirect.com/science/article/pii/S0889159123002313> (visited on 06/03/2024).
- [40] Stefano Canali, Viola Schiaffonati, and Andrea Aliverti. “Challenges and Recommendations for Wearable Devices in Digital Health: Data Quality, Interoperability, Health Equity, Fairness”. In: *PLOS Digital Health* 1.10 (Oct. 2022). DOI: 10.1371/journal.pdig.0000104. pmid: 36812619. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9931360/> (visited on 06/03/2024).
- [41] Sophie Huhn et al. “The Impact of Wearable Technologies in Health Research: Scoping Review”. In: *JMIR mHealth and uHealth* 10.1 (Jan. 2022). DOI: 10.2196/34384. pmid: 35076409. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8826148/> (visited on 06/03/2024).
- [42] Vini Vijayan et al. “Review of Wearable Devices and Data Collection Considerations for Connected Health”. In: *Sensors (Basel, Switzerland)* 21.16 (Aug. 19, 2021), p. 5589. ISSN: 1424-8220. DOI: 10.3390/s21165589. pmid: 34451032. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8402237/> (visited on 06/03/2024).
- [43] *Wearable Medical Devices Market Size / Research Report [2030]*. Fortune Business Insights. May 13, 2024. URL: <https://www.fortunebusinessinsights.com/industry-reports/wearable-medical-devices-market-101070> (visited on 06/03/2024).

- [44] Fangyu Liu, Amal A. Wanigatunga, and Jennifer A. Schrack. “Assessment of Physical Activity in Adults Using Wrist Accelerometers”. In: *Epidemiologic Reviews* 43.1 (2021), p. 65. DOI: 10.1093/epirev/mxab004. pmid: 34215874. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8900289/> (visited on 06/03/2024).
- [45] Erik K. Antonsson and Robert W. Mann. “The Frequency Content of Gait”. In: *Journal of Biomechanics* 18.1 (Jan. 1, 1985), pp. 39–47. ISSN: 0021-9290. DOI: 10.1016/0021-9290(85)90043-0. URL: <https://www.sciencedirect.com/science/article/pii/0021929085900430> (visited on 08/21/2024).
- [46] Ping Yi Chan et al. “Motion Characteristics of Subclinical Tremors in Parkinson’s Disease and Normal Subjects”. In: *Scientific Reports* 12.1 (Mar. 7, 2022), p. 4021. ISSN: 2045-2322. DOI: 10.1038/s41598-022-07957-z. URL: <https://www.nature.com/articles/s41598-022-07957-z> (visited on 08/21/2024).
- [47] Vincent T. van Hees et al. “Autocalibration of accelerometer data for free-living physical activity assessment using local gravity and temperature: an evaluation on four continents”. In: *Journal of Applied Physiology (Bethesda, Md.: 1985)* 117.7 (Oct. 1, 2014), pp. 738–744. ISSN: 1522-1601. DOI: 10.1152/japplphysiol.00421.2014.
- [48] Vincent T. van Hees et al. “Separating Movement and Gravity Components in an Acceleration Signal and Implications for the Assessment of Human Daily Physical Activity”. In: *PLoS ONE* 8.4 (Apr. 23, 2013). Ed. by Michael Müller, e61691. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0061691. URL: <https://dx.plos.org/10.1371/journal.pone.0061691> (visited on 10/11/2022).
- [49] Matthew Willetts et al. “Statistical Machine Learning of Sleep and Physical Activity Phenotypes from Sensor Data in 96,220 UK Biobank Participants”. In: *Scientific Reports* 8.1 (May 21, 2018), p. 7961. ISSN: 2045-2322. DOI: 10.1038/s41598-018-26174-1. URL: <https://doi.org/10.1038/s41598-018-26174-1>.
- [50] Rosemary Walmsley et al. “Reallocation of Time between Device-Measured Movement Behaviours and Risk of Incident Cardiovascular Disease”. In: *British Journal of Sports Medicine* (Sept. 6, 2021), bjsports-2021-104050. DOI: 10.1136/bjsports-2021-104050. URL: <http://bjsm.bmj.com/content/early/2022/04/06/bjsports-2021-104050.abstract>.
- [51] Charalampos Sotirakis et al. “Identification of Motor Progression in Parkinson’s Disease Using Wearable Sensors and Machine Learning”. In: *npj Parkinson’s Disease* 9.1 (Oct. 7, 2023), p. 142. ISSN: 2373-8057. DOI: 10.1038/s41531-023-00581-2. URL: <https://doi.org/10.1038/s41531-023-00581-2>.
- [52] Leo Breiman. “Random Forests”. In: *Machine Learning* 45.1 (Oct. 1, 2001), pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. URL: <https://doi.org/10.1023/A:1010933404324> (visited on 06/05/2022).

- [53] Roozbeh Atri et al. “Deep Learning for Daily Monitoring of Parkinson’s Disease Outside the Clinic Using Wearable Sensors.” In: *Sensors (Basel, Switzerland)* 22.18 (Sept. 9, 2022). ISSN: 1424-8220. DOI: 10.3390/s22186831. pmid: 36146181.
- [54] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *Nature* 521.7553 (May 2015). Publisher: Nature Publishing Group, pp. 436–444. ISSN: 1476-4687. DOI: 10.1038/nature14539. URL: <https://www.nature.com/articles/nature14539> (visited on 03/22/2025).
- [55] Aiden Doherty et al. “Large Scale Population Assessment of Physical Activity Using Wrist Worn Accelerometers: The UK Biobank Study”. In: *PLOS ONE* 12.2 (Feb. 1, 2017), e0169649. DOI: 10.1371/journal.pone.0169649. URL: <https://doi.org/10.1371/journal.pone.0169649>.
- [56] James R Williamson et al. “Detecting Parkinson’s Disease from Wrist-Worn Accelerometry in the U.K. Biobank”. In: *Sensors (Basel, Switzerland)* 21.6 (Mar. 14, 2021), p. 2047. ISSN: 1424-8220. DOI: 10.3390/s21062047. PMID: 33799420. URL: <https://pubmed.ncbi.nlm.nih.gov/33799420>.
- [57] J. A. Hanley and B. J. McNeil. “The meaning and use of the area under a receiver operating characteristic (ROC) curve”. In: *Radiology* 143.1 (Apr. 1982), pp. 29–36. ISSN: 0033-8419. DOI: 10.1148/radiology.143.1.7063747.
- [58] Ann-Kathrin Schalkamp et al. “Wearable Movement-Tracking Data Identify Parkinson’s Disease Years before Clinical Diagnosis”. In: *Nature Medicine* 29.8 (Aug. 1, 2023), pp. 2048–2056. ISSN: 1546-170X. DOI: 10.1038/s41591-023-02440-2. URL: <https://doi.org/10.1038/s41591-023-02440-2>.
- [59] Mengyi Liu et al. “Association of Accelerometer-Measured Physical Activity Intensity, Sedentary Time, and Exercise Time with Incident Parkinson’s Disease”. In: *npj Digital Medicine* 6.1 (Nov. 28, 2023), p. 224. ISSN: 2398-6352. DOI: 10.1038/s41746-023-00969-7. URL: <https://doi.org/10.1038/s41746-023-00969-7>.
- [60] Angela Zhao et al. “Evaluating the Prediction Performance of Objective Physical Activity Measures for Incident Parkinson’s Disease in the UK Biobank”. In: *Journal of Neurology* (Aug. 24, 2023). ISSN: 1432-1459. DOI: 10.1007/s00415-023-11939-0. URL: <https://doi.org/10.1007/s00415-023-11939-0>.
- [61] Fabian Lin et al. “Association of Physical Activity Pattern and Risk of Parkinson’s Disease”. In: *npj Digital Medicine* 7.1 (May 23, 2024), p. 137. ISSN: 2398-6352. DOI: 10.1038/s41746-024-01135-3. URL: <https://doi.org/10.1038/s41746-024-01135-3>.
- [62] *Parkinson’s Progression Markers Initiative*. PPMI. URL: <https://www.ppmi-info.org/>.
- [63] Pavan A. Vaswani, Thomas F. Tropea, and Nabila Dahodwala. “Overcoming Barriers to Parkinson Disease Trial Participation: Increasing Diversity and Novel Designs for Recruitment and Retention”. In: *Neurotherapeutics* 17.4 (Oct. 2020), pp. 1724–1735. ISSN: 1933-7213. DOI: 10.1007/s13311-020-00960-0. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7851248/> (visited on 03/20/2025).

- [64] Cathie Sudlow et al. “UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age”. In: *PLOS Medicine* 12.3 (Mar. 31, 2015), e1001779. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1001779. URL: <https://dx.plos.org/10.1371/journal.pmed.1001779> (visited on 02/13/2024).
- [65] World Health Organization. *ICD-10 : international statistical classification of diseases and related health problems : tenth revision*. Edition: 2nd ed Pages: Spanish version, 1st edition published by PAHO as Publicación Científica 544. 2004.
- [66] UK Biobank. *UK Biobank*. May 21, 2024. URL: <https://www.ukbiobank.ac.uk> (visited on 06/26/2024).
- [67] *Data Providers and Dates of Data Availability*. UKBiobank. URL: https://biobank.ndph.ox.ac.uk/ukb/exinfo.cgi?src=Data_providers_and_dates (visited on 07/19/2024).
- [68] UK Biobank. *UK Biobank Enable Your Research*. June 24, 2021. URL: <https://www.ukbiobank.ac.uk/enable-your-research> (visited on 02/13/2024).
- [69] Mike A Nalls et al. “Identification of Novel Risk Loci, Causal Insights, and Heritable Risk for Parkinson’s Disease: A Meta-Analysis of Genome-Wide Association Studies”. In: *The Lancet Neurology* 18.12 (Dec. 1, 2019), pp. 1091–1102. ISSN: 1474-4422. DOI: 10.1016/S1474-4422(19)30320-5. URL: <https://www.sciencedirect.com/science/article/pii/S1474442219303205> (visited on 06/05/2022).
- [70] Lei Clifton et al. “Assessing the Importance of Primary Care Diagnoses in the UK Biobank”. In: *European Journal of Epidemiology* 39.2 (Feb. 2024), pp. 219–229. ISSN: 1573-7284. DOI: 10.1007/s10654-023-01095-0. pmid: 38225527.
- [71] Kathryn Bush et al. *Definitions of Parkinson’s Disease and the Major Causes of Parkinsonism, UK Biobank Phase 1 Outcomes Adjudication*. Mar. 2018. URL: https://biobank.ndph.ox.ac.uk/showcase/showcase/docs/alg_outcome_pdp.pdf.
- [72] Sanah Yousaf and Amy Bonsall. “UK Townsend Deprivation Scores from 2011 Census Data”. Sept. 7, 2017. URL: https://s3-eu-west-1.amazonaws.com/statistics.digitalresources.jisc.ac.uk/dkan/files/Townsend_Deprivation_Scores/UK%20Townsend%20Deprivation%20Scores%20from%202011%20census%20data.pdf.
- [73] Alice R Carter et al. “Educational Attainment as a Modifier for the Effect of Polygenic Scores for Cardiovascular Risk Factors: Cross-Sectional and Prospective Analysis of UK Biobank”. In: *International Journal of Epidemiology* 51.3 (Feb. 3, 2022), pp. 885–897. ISSN: 0300-5771. DOI: 10.1093/ije/dyac002. pmid: 35134953. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9189971/> (visited on 07/22/2024).

- [74] E. Ray Dorsey and William J. Marks Jr. “Verily and Its Approach to Digital Biomarkers”. In: *Digital Biomarkers* 1.1 (July 14, 2017), pp. 96–99. ISSN: 2504-110X. DOI: 10.1159/000476051. pmid: 32095750. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7015368/> (visited on 07/31/2024).
- [75] Kenneth Marek et al. “The Parkinson’s Progression Markers Initiative (PPMI) - Establishing a PD Biomarker Cohort”. In: *Annals of Clinical and Translational Neurology* 5.12 (Dec. 2018), pp. 1460–1477. ISSN: 2328-9503. DOI: 10.1002/acn3.644. pmid: 30564614.
- [76] S. Chan Chang et al. “Capture-24: Activity Tracker Dataset for Human Activity Recognition”. In: (2021). URL: <https://ora.ox.ac.uk/objects/uuid:99d7c092-d865-4a19-b096-cc16440cd001> (visited on 07/05/2022).
- [77] Shing Chan et al. “CAPTURE-24: A Large Dataset of Wrist-Worn Activity Tracker Data Collected in the Wild for Human Activity Recognition”. In: *Scientific Data* 11.1 (1 Oct. 16, 2024), pp. 1–11. ISSN: 2052-4463. DOI: 10.1038/s41597-024-03960-3. URL: <https://www.nature.com/articles/s41597-024-03960-3> (visited on 10/23/2024).
- [78] Alessandra Angelucci, Stefano Canali, and Andrea Aliverti. “Digital Technologies for Step Counting: Between Promises of Reliability and Risks of Reductionism”. In: *Frontiers in Digital Health* 5 (Dec. 13, 2023), p. 1330189. ISSN: 2673-253X. DOI: 10.3389/fdgth.2023.1330189. pmid: 38152629. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10751316/> (visited on 07/20/2024).
- [79] Scott R. Small et al. “Self-Supervised Machine Learning to Characterise Step Counts from Wrist-Worn Accelerometers in the UK Biobank”. In: *Medicine & Science in Sports & Exercise* (May 15, 2024). ISSN: 0195-9131. URL: https://journals.lww.com/acsm-msse/fulltext/9900/self_supervised_machine_learning_to_characterise.543.aspx.
- [80] Hang Yuan et al. “Self-Supervised Learning for Human Activity Recognition Using 700,000 Person-Days of Wearable Data”. In: *npj Digital Medicine* 7.1 (Apr. 12, 2024), pp. 1–10. ISSN: 2398-6352. DOI: 10.1038/s41746-024-01062-3. URL: <https://www.nature.com/articles/s41746-024-01062-3> (visited on 09/01/2024).
- [81] Víctor Navarro-López et al. “Arm-Swing Kinematics in Parkinson’s Disease: A Systematic Review and Meta-Analysis”. In: *Gait & Posture* 98 (Oct. 2022), pp. 85–95. ISSN: 1879-2219. DOI: 10.1016/j.gaitpost.2022.08.017. pmid: 36088898.
- [82] Rana Zia Ur Rehman et al. “Investigating the Impact of Environment and Data Aggregation by Walking Bout Duration on Parkinson’s Disease Classification Using Machine Learning”. In: *Frontiers in Aging Neuroscience* 14 (2022). ISSN: 1663-4365. DOI: 10.3389/fnagi.2022.808518. URL: <https://www.frontiersin.org/article/10.3389/fnagi.2022.808518>.
- [83] Phil Snyder and Meghasyam Tummalacherla. *MJFF Levodopa Response Study*. Synapse, 2019. DOI: 10.7303/SYN20681023. URL: <https://repo-prod.prod.sagebase.org/repo/v1/doi/locate?id=syn20681023&type=ENTITY>.

- [84] Alex V. Rowlands et al. “Enhancing the Value of Accelerometer-Assessed Physical Activity: Meaningful Visual Comparisons of Data-Driven Translational Accelerometer Metrics”. In: *Sports Medicine - Open* 5.1 (Dec. 5, 2019), p. 47. ISSN: 2198-9761. DOI: 10.1186/s40798-019-0225-9. URL: <https://doi.org/10.1186/s40798-019-0225-9> (visited on 08/30/2024).
- [85] James W. White et al. “Comparison of Raw Accelerometry Data from ActiGraph, Apple Watch, Garmin, and Fitbit Using a Mechanical Shaker Table”. In: *PLOS ONE* 19.3 (Mar. 29, 2024), e0286898. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0286898. URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0286898> (visited on 08/30/2024).
- [86] S. Butterworth. “On the Theory of Filter Amplifiers”. In: *Experimental Wireless and the Wireless Engineer* 7 (1930). Accessed: 2025-04-02, pp. 536–541. URL: https://www.changpuak.ch/electronics/downloads/On_the_Theory_of_Filter_Amplifiers.pdf.
- [87] Rinat Khusainov et al. “Real-Time Human Ambulation, Activity, and Physiological Monitoring: Taxonomy of Issues, Techniques, Applications, Challenges and Limitations”. In: *Sensors (Basel, Switzerland)* 13.10 (Sept. 25, 2013), pp. 12852–12902. ISSN: 1424-8220. DOI: 10.3390/s131012852. pmid: 24072027. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3859040/> (visited on 07/22/2024).
- [88] Séverine Sabia et al. “Association Between Questionnaire- and Accelerometer-Assessed Physical Activity: The Role of Sociodemographic Factors”. In: *American Journal of Epidemiology* 179.6 (Mar. 15, 2014), pp. 781–790. ISSN: 0002-9262. DOI: 10.1093/aje/kwt330. URL: <https://doi.org/10.1093/aje/kwt330> (visited on 03/23/2025).
- [89] Henry Friday Nweke et al. “Deep Learning Algorithms for Human Activity Recognition Using Mobile and Wearable Sensor Networks: State of the Art and Research Challenges”. In: *Expert Systems with Applications* 105 (Sept. 1, 2018), pp. 233–261. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2018.03.056. URL: <https://www.sciencedirect.com/science/article/pii/S0957417418302136> (visited on 10/25/2024).
- [90] Zawar Hussain, Quan Z. Sheng, and Wei Emma Zhang. “A Review and Categorization of Techniques on Device-Free Human Activity Recognition”. In: *Journal of Network and Computer Applications* 167 (Oct. 1, 2020), p. 102738. ISSN: 1084-8045. DOI: 10.1016/j.jnca.2020.102738. URL: <https://www.sciencedirect.com/science/article/pii/S1084804520302125> (visited on 10/25/2024).
- [91] Kaiming He et al. *Identity Mappings in Deep Residual Networks*. July 25, 2016. DOI: 10.48550/arXiv.1603.05027. arXiv: 1603.05027[cs]. URL: <http://arxiv.org/abs/1603.05027> (visited on 04/14/2025).
- [92] Grace O. Dibben et al. “Physical Activity Assessment by Accelerometry in People with Heart Failure”. In: *BMC Sports Science, Medicine and Rehabilitation* 12.1 (Aug. 12, 2020), p. 47. ISSN: 2052-1847. DOI: 10.1186/s13102-020-00196-7. URL: <https://doi.org/10.1186/s13102-020-00196-7> (visited on 08/23/2024).

- [93] Christina B. Dillon et al. “Number of Days Required to Estimate Habitual Activity Using Wrist-Worn GENEActiv Accelerometer: A Cross-Sectional Study”. In: *PLoS ONE* 11.5 (May 5, 2016), e0109913. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0109913. pmid: 27149674. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4858250/> (visited on 08/23/2024).
- [94] Dale W. Esliger et al. “Validation of the GENEActiv Accelerometer”. In: *Medicine & Science in Sports & Exercise* 43.6 (June 2011), p. 1085. ISSN: 0195-9131. DOI: 10.1249/MSS.0b013e31820513be. URL: https://journals.lww.com/acsm-msse/Fulltext/2011/06000/Validation_of_the_GENEActiv_Accelerometer.22.aspx (visited on 08/23/2024).
- [95] Maria Hildebrand et al. “Age Group Comparability of Raw Accelerometer Output from Wrist- and Hip-Worn Monitors”. In: *Medicine & Science in Sports & Exercise* 46.9 (Sept. 2014), p. 1816. ISSN: 0195-9131. DOI: 10.1249/MSS.0000000000000289. URL: https://journals.lww.com/acsm-msse/Fulltext/2014/09000/Age_Group_Comparability_of_Raw_Accelerometer.17.aspx (visited on 08/23/2024).
- [96] Maria Hildebrand et al. “Evaluation of Raw Acceleration Sedentary Thresholds in Children and Adults”. In: *Scandinavian Journal of Medicine & Science in Sports* 27.12 (2017), pp. 1814–1823. ISSN: 1600-0838. DOI: 10.1111/sms.12795. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/sms.12795> (visited on 08/23/2024).
- [97] Jairo H. Migueles et al. “Calibration and Cross-Validation of Accelerometer Cut-Points to Classify Sedentary Time and Physical Activity from Hip and Non-Dominant and Dominant Wrists in Older Adults”. In: *Sensors (Basel, Switzerland)* 21.10 (May 11, 2021), p. 3326. ISSN: 1424-8220. DOI: 10.3390/s21103326. pmid: 34064790. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8150960/> (visited on 08/23/2024).
- [98] George J. Sanders et al. “Evaluation of Wrist and Hip Sedentary Behaviour and Moderate-to-Vigorous Physical Activity Raw Acceleration Cutpoints in Older Adults”. In: *Journal of Sports Sciences* 37.11 (June 3, 2019), pp. 1270–1279. ISSN: 0264-0414. DOI: 10.1080/02640414.2018.1555904. pmid: 30558487. URL: <https://doi.org/10.1080/02640414.2018.1555904> (visited on 08/23/2024).
- [99] R. Walmsley. “Device-Measured 24-Hour Movement Behaviours and Risk of Incident Cardiovascular Disease”. <http://purl.org/dc/dcmitype/Text>. University of Oxford, 2022. URL: <https://ora.ox.ac.uk/objects/uuid:17cbcd2e-8cb5-4964-b497-16ede2b981eb> (visited on 08/21/2024).
- [100] Katherine Ellis et al. “Hip and Wrist Accelerometer Algorithms for Free-Living Behavior Classification”. In: *Medicine and Science in Sports and Exercise* 48.5 (May 2016), pp. 933–940. ISSN: 1530-0315. DOI: 10.1249/MSS.0000000000000840. pmid: 26673126.
- [101] Toby G. Pavey et al. “Field Evaluation of a Random Forest Activity Classifier for Wrist-Worn Accelerometer Data”. In: *Journal of Science and Medicine in Sport* 20.1 (Jan. 2017), pp. 75–80. ISSN: 1878-1861. DOI: 10.1016/j.jsams.2016.06.003. pmid: 27372275.

- [102] Alok Kumar Chowdhury et al. “Ensemble Methods for Classification of Physical Activities from Wrist Accelerometry”. In: *Medicine and Science in Sports and Exercise* 49.9 (Sept. 2017), pp. 1965–1973. ISSN: 1530-0315. DOI: 10.1249/MSS.0000000000001291. pmid: 28419025.
- [103] Matthew N. Ahmadi, Toby G. Pavey, and Stewart G. Trost. “Machine Learning Models for Classifying Physical Activity in Free-Living Preschool Children”. In: *Sensors (Basel, Switzerland)* 20.16 (Aug. 5, 2020), p. 4364. ISSN: 1424-8220. DOI: 10.3390/s20164364. pmid: 32764316.
- [104] L. Rabiner and B. Juang. “An introduction to hidden Markov models”. In: *IEEE ASSP Magazine* 3.1 (Jan. 1986). Conference Name: IEEE ASSP Magazine, pp. 4–16. ISSN: 1558-1284. DOI: 10.1109/MASSP.1986.1165342. URL: <https://ieeexplore.ieee.org/document/1165342> (visited on 03/22/2025).
- [105] Leland McInnes, John Healy, and James Melville. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction”. Version 3. In: (2018). DOI: 10.48550/ARXIV.1802.03426. URL: <https://arxiv.org/abs/1802.03426> (visited on 05/29/2022).
- [106] G.D. Forney. “The Viterbi Algorithm”. In: *Proceedings of the IEEE* 61.3 (Mar. 1973), pp. 268–278. ISSN: 1558-2256. DOI: 10.1109/PROC.1973.9030.
- [107] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. Jan. 30, 2017. DOI: 10.48550/arXiv.1412.6980. arXiv: 1412.6980 [cs]. URL: <http://arxiv.org/abs/1412.6980> (visited on 03/22/2025).
- [108] *OOB Errors for Random Forests*. scikit-learn. URL: https://scikit-learn/stable/auto_examples/ensemble/plot_ensemble_oob.html (visited on 03/24/2025).
- [109] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. “Random Forests”. In: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Ed. by Trevor Hastie, Robert Tibshirani, and Jerome Friedman. New York, NY: Springer, 2009, pp. 587–604. ISBN: 978-0-387-84858-7. DOI: 10.1007/978-0-387-84858-7_15. URL: https://doi.org/10.1007/978-0-387-84858-7_15 (visited on 03/24/2025).
- [110] *StratifiedGroupKFold*. scikit-learn. URL: https://scikit-learn/stable/modules/generated/sklearn.model_selection.StratifiedGroupKFold.html (visited on 03/24/2025).
- [111] J. M. Bland and D. G. Altman. “Statistical methods for assessing agreement between two methods of clinical measurement”. In: *Lancet (London, England)* 1.8476 (Feb. 8, 1986), pp. 307–310. ISSN: 0140-6736.
- [112] *11. Correlation and regression | The BMJ*. The BMJ | The BMJ: leading general medical journal. Research. Education. Comment. Section: Statistics. Oct. 28, 2020. URL: <https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/11-correlation-and-regression> (visited on 03/24/2025).
- [113] D. Levine, Jim Richards, and M. Whittle. *Whittle’s Gait Analysis*. Elsevier, July 6, 2012. 192 pp. ISBN: 978-0-7020-4265-2. URL: <http://www.elsevier.com/books/whittles-gait-analysis/levine/978-0-7020-5199-9> (visited on 08/27/2024).

- [114] Massimiliano Zanin et al. “Gait Analysis under the Lens of Statistical Physics”. In: *Computational and Structural Biotechnology Journal* 20 (Jan. 1, 2022), pp. 3257–3267. ISSN: 2001-0370. DOI: 10.1016/j.csbj.2022.06.022. URL: <https://www.sciencedirect.com/science/article/pii/S2001037022002367> (visited on 08/27/2024).
- [115] Andrew Leroux et al. “Organizing and Analyzing the Activity Data in NHANES”. In: *Statistics in Biosciences* 11.2 (July 2019), pp. 262–287. ISSN: 1867-1764, 1867-1772. DOI: 10.1007/s12561-018-09229-9. URL: <http://link.springer.com/10.1007/s12561-018-09229-9> (visited on 09/18/2023).
- [116] Marcin Straczekiewicz, Emily J. Huang, and Jukka-Pekka Onnela. “A “One-Size-Fits-Most” Walking Recognition Method for Smartphones, Smartwatches, and Wearable Accelerometers”. In: *npj Digital Medicine* 6.1 (Feb. 23, 2023), pp. 1–16. ISSN: 2398-6352. DOI: 10.1038/s41746-022-00745-z. URL: <https://www.nature.com/articles/s41746-022-00745-z> (visited on 08/27/2024).
- [117] Lily Koffman and John Muschelli. “Evaluating Step Counting Algorithms on Subsecond Wrist-Worn Accelerometry: A Comparison Using Publicly Available Data Sets”. In: (July 30, 2024). DOI: 10.1123/jmpb.2024-0008. URL: <https://journals.humankinetics.com/view/journals/jmpb/7/1/article-jmpb.2024-0008.xml> (visited on 08/27/2024).
- [118] Peter I. Frazier. *A Tutorial on Bayesian Optimization*. July 8, 2018. DOI: 10.48550/arXiv.1807.02811. arXiv: 1807.02811[stat]. URL: <http://arxiv.org/abs/1807.02811> (visited on 03/26/2025).
- [119] Anqi Mao, Mehryar Mohri, and Yutao Zhong. *Cross-Entropy Loss Functions: Theoretical Analysis and Applications*. June 20, 2023. DOI: 10.48550/arXiv.2304.07288. arXiv: 2304.07288[cs]. URL: <http://arxiv.org/abs/2304.07288> (visited on 03/26/2025).
- [120] Anat Mirelman et al. “Gait Impairments in Parkinson’s Disease”. In: *The Lancet Neurology* 18.7 (July 1, 2019), pp. 697–708. ISSN: 1474-4422. DOI: 10.1016/S1474-4422(19)30044-4. URL: [https://doi.org/10.1016/S1474-4422\(19\)30044-4](https://doi.org/10.1016/S1474-4422(19)30044-4) (visited on 05/10/2022).
- [121] Aidan Acquah et al. “Daily Steps Are a Predictor of, but Perhaps Not a Modifiable Risk Factor for Parkinson’s Disease: Findings from the UK Biobank”. In: *medRxiv* (Jan. 1, 2024), p. 2024.08.13.24311539. DOI: 10.1101/2024.08.13.24311539. URL: <http://medrxiv.org/content/early/2024/08/14/2024.08.13.24311539.abstract>.
- [122] Sara C. LaHue, Cynthia L. Comella, and Caroline M. Tanner. “The Best Medicine? The Influence of Physical Activity and Inactivity on Parkinson’s Disease.” In: *Movement disorders : official journal of the Movement Disorder Society* 31.10 (Oct. 2016), pp. 1444–1454. ISSN: 1531-8257 0885-3185. DOI: 10.1002/mds.26728. pmid: 27477046.

- [123] James F. Sallis and Brian E. Saelens. “Assessment of Physical Activity by Self-Report: Status, Limitations, and Future Directions”. In: *Research Quarterly for Exercise and Sport* 71 (sup2 June 1, 2000), pp. 1–14. ISSN: 0270-1367. DOI: 10.1080/02701367.2000.11082780. URL: <https://doi.org/10.1080/02701367.2000.11082780>.
- [124] R J Shephard. “Limits to the Measurement of Habitual Physical Activity by Questionnaires”. In: *British Journal of Sports Medicine* 37.3 (June 1, 2003), p. 197. DOI: 10.1136/bjsm.37.3.197. URL: <http://bjsm.bmj.com/content/37/3/197.abstract>.
- [125] David R. Bassett et al. “Step Counting: A Review of Measurement Considerations and Health-Related Applications”. In: *Sports Medicine (Auckland, N.z.)* 47.7 (2017), pp. 1303–1315. ISSN: 0112-1642. DOI: 10.1007/s40279-016-0663-1. pmid: 28005190. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5488109/> (visited on 07/10/2024).
- [126] Navid Shokouhi et al. “Accuracy of Step Count Estimations in Parkinson’s Disease Can Be Predicted Using Ambulatory Monitoring”. In: *Frontiers in Aging Neuroscience* 14 (June 16, 2022). ISSN: 1663-4365. DOI: 10.3389/fnagi.2022.904895. URL: <https://www.frontiersin.org/journals/aging-neuroscience/articles/10.3389/fnagi.2022.904895/full> (visited on 07/10/2024).
- [127] Edoardo Bianchini et al. “Four Days Are Enough to Provide a Reliable Daily Step Count in Mild to Moderate Parkinson’s Disease through a Commercial Smartwatch”. In: *Sensors* 23.21 (21 Jan. 2023), p. 8971. ISSN: 1424-8220. DOI: 10.3390/s23218971. URL: <https://www.mdpi.com/1424-8220/23/21/8971> (visited on 07/10/2024).
- [128] Aiden Doherty et al. “GWAS Identifies 14 Loci for Device-Measured Physical Activity and Sleep Duration”. In: *Nature Communications* 9.1 (Dec. 10, 2018), p. 5257. ISSN: 2041-1723. DOI: 10.1038/s41467-018-07743-4. URL: <https://doi.org/10.1038/s41467-018-07743-4>.
- [129] D. R. Cox. “Regression Models and Life-Tables”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 34.2 (1972), pp. 187–220. ISSN: 00359246. JSTOR: 2985181. URL: <http://www.jstor.org/stable/2985181> (visited on 11/09/2023).
- [130] D. F. Easton, J. Peto, and A. G. Babiker. “Floating Absolute Risk: An Alternative to Relative Risk in Survival and Case-Control Analysis Avoiding an Arbitrary Reference Group.” In: *Statistics in medicine* 10.7 (July 1991), pp. 1025–1035. ISSN: 0277-6715. DOI: 10.1002/sim.4780100703. pmid: 1652152.
- [131] Betty R. Kirkwood and Jonathan A. C. Sterne. *Essential Medical Statistics*. Hoboken: John Wiley & Sons, Incorporated, 2003.
- [132] David Schoenfeld. “Partial Residuals for the Proportional Hazards Regression Model”. In: *Biometrika* 69.1 (Apr. 1982), pp. 239–241. ISSN: 0006-3444. DOI: 10.1093/biomet/69.1.239. eprint: <https://academic.oup.com/biomet/article-pdf/69/1/239/669606/69-1-239.pdf>. URL: <https://doi.org/10.1093/biomet/69.1.239>.

- [133] Connor B. Weir and Arif Jan. “BMI Classification Percentile And Cut Off Points”. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2023. pmid: 31082114. URL: <http://www.ncbi.nlm.nih.gov/books/NBK541070/> (visited on 01/16/2024).
- [134] Martyn Plummer and Bendix Carstensen. “Lexis: An R Class for Epidemiological Studies with Long-Term Follow-Up”. In: *Journal of Statistical Software* 38 (Jan. 4, 2011), pp. 1–12. ISSN: 1548-7660. DOI: 10.18637/jss.v038.i05. URL: <https://doi.org/10.18637/jss.v038.i05> (visited on 03/27/2025).
- [135] Lai San Hong and Sarah Lewington. “Lexis Expansion- Age-at-Risk Adjustment for Survival Analysis”. In: (2013).
- [136] Sarah Floud et al. “Body Mass Index, Diet, Physical Inactivity, and the Incidence of Dementia in 1 Million UK Women”. In: *Neurology* 94.2 (Jan. 14, 2020), e123–e132. DOI: 10.1212/WNL.00000000000008779. URL: <https://www.neurology.org/doi/10.1212/WNL.00000000000008779> (visited on 06/28/2024).
- [137] Erik von Elm et al. “The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies”. In: *The Lancet* 370.9596 (Oct. 20, 2007), pp. 1453–1457. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(07)61602-X. URL: [https://doi.org/10.1016/S0140-6736\(07\)61602-X](https://doi.org/10.1016/S0140-6736(07)61602-X) (visited on 05/15/2024).
- [138] Q. Xu et al. “Physical Activities and Future Risk of Parkinson Disease”. In: *Neurology* 75.4 (July 27, 2010), pp. 341–348. ISSN: 0028-3878, 1526-632X. DOI: 10.1212/WNL.0b013e3181ea1597. URL: <https://www.neurology.org/doi/10.1212/WNL.0b013e3181ea1597> (visited on 01/16/2024).
- [139] Séverine Sabia et al. “Physical Activity, Cognitive Decline, and Risk of Dementia: 28 Year Follow-up of Whitehall II Cohort Study”. In: *BMJ* (June 22, 2017), j2709. ISSN: 0959-8138, 1756-1833. DOI: 10.1136/bmj.j2709. URL: <https://www.bmj.com/lookup/doi/10.1136/bmj.j2709> (visited on 06/28/2024).
- [140] Francesca Magrinelli et al. “Pathophysiology of Motor Dysfunction in Parkinson’s Disease as the Rationale for Drug Treatment and Rehabilitation”. In: *Parkinson’s Disease* 2016 (2016), p. 9832839. ISSN: 2090-8083. DOI: 10.1155/2016/9832839. pmid: 27366343. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4913065/> (visited on 07/31/2024).
- [141] E. Ray Dorsey et al. “Global, Regional, and National Burden of Parkinson’s Disease, 1990–2016: A Systematic Analysis for the Global Burden of Disease Study 2016”. In: *The Lancet Neurology* 17.11 (2018), pp. 939–953. ISSN: 1474-4422. DOI: 10.1016/S1474-4422(18)30295-3. URL: [https://doi.org/10.1016/S1474-4422\(18\)30295-3](https://doi.org/10.1016/S1474-4422(18)30295-3).
- [142] Roger A. Barker. “Parkinson’s Disease as a Preventable Pandemic”. In: *The Lancet Neurology* 19.10 (Oct. 1, 2020), p. 813. ISSN: 1474-4422, 1474-4465. DOI: 10.1016/S1474-4422(20)30302-1. URL: [https://www.thelancet.com/journals/lanneur/article/PIIS1474-4422\(20\)30302-1/fulltext](https://www.thelancet.com/journals/lanneur/article/PIIS1474-4422(20)30302-1/fulltext) (visited on 07/31/2024).

- [143] Srikanth Vasudevan et al. “Digital Biomarkers: Convergence of Digital Health Technologies and Biomarkers”. In: *npj Digital Medicine* 5.1 (Mar. 25, 2022), pp. 1–3. ISSN: 2398-6352. DOI: 10.1038/s41746-022-00583-z. URL: <https://www.nature.com/articles/s41746-022-00583-z> (visited on 07/31/2024).
- [144] Luis Sigcha et al. “Deep Learning and Wearable Sensors for the Diagnosis and Monitoring of Parkinson’s Disease: A Systematic Review”. In: *Expert Systems with Applications* 229 (Nov. 1, 2023), p. 120541. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2023.120541. URL: <https://www.sciencedirect.com/science/article/pii/S0957417423010436> (visited on 07/31/2024).
- [145] Roberto Erro et al. “What Do Patients with Scans without Evidence of Dopaminergic Deficit (SWEDD) Have? New Evidence and Continuing Controversies”. In: *Journal of Neurology, Neurosurgery & Psychiatry* 87.3 (Mar. 1, 2016), pp. 319–323. ISSN: 0022-3050, 1468-330X. DOI: 10.1136/jnnp-2014-310256. pmid: 25991401. URL: <https://jnnp.bmj.com/content/87/3/319> (visited on 07/31/2024).
- [146] Patrick Hilden et al. “How Many Days Are Needed? Measurement Reliability of Wearable Device Data to Assess Physical Activity”. In: *PLOS ONE* 18.2 (Feb. 24, 2023), e0282162. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0282162. URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0282162> (visited on 07/31/2024).
- [147] Laurens van der Maaten and Geoffrey Hinton. “Visualizing Data Using T-SNE”. In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605. ISSN: 1533-7928. URL: <http://jmlr.org/papers/v9/vandermaaten08a.html> (visited on 08/01/2024).
- [148] Hanrui Zhang et al. “Deep Learning Identifies Digital Biomarkers for Self-Reported Parkinson’s Disease”. In: *Patterns* 1.3 (2020), p. 100042. ISSN: 2666-3899. DOI: 10.1016/j.patter.2020.100042. URL: <https://www.sciencedirect.com/science/article/pii/S2666389920300477>.
- [149] Mícheál Ó. Breasail et al. “Wearable GPS and Accelerometer Technologies for Monitoring Mobility and Physical Activity in Neurodegenerative Disorders: A Systematic Review”. In: *Sensors (Basel, Switzerland)* 21.24 (Dec. 10, 2021), p. 8261. ISSN: 1424-8220. DOI: 10.3390/s21248261. pmid: 34960353. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8705556/> (visited on 05/30/2022).
- [150] Kallirhoe Kalinderi, Vasileios Papaliagkas, and Liana Fidani. “GLP-1 Receptor Agonists: A New Treatment in Parkinson’s Disease”. In: *International Journal of Molecular Sciences* 25.7 (Mar. 29, 2024), p. 3812. ISSN: 1422-0067. DOI: 10.3390/ijms25073812. pmid: 38612620. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11011817/> (visited on 09/05/2024).

- [151] Alastair J. Noyce et al. “PREDICT-PD: Identifying Risk of Parkinson’s Disease in the Community: Methods and Baseline Results”. In: *Journal of Neurology, Neurosurgery & Psychiatry* 85.1 (Jan. 1, 2014), pp. 31–37. ISSN: 0022-3050, 1468-330X. DOI: 10.1136/jnnp-2013-305420. pmid: 23828833. URL: <https://jnnp.bmj.com/content/85/1/31> (visited on 06/01/2022).
- [152] Stef van Buuren and Karin Groothuis-Oudshoorn. “Mice: Multivariate Imputation by Chained Equations in R”. In: *Journal of Statistical Software* 45.3 (Dec. 2011), pp. 1–67. DOI: 10.18637/jss.v045.i03. (Visited on 10/30/2024).
- [153] Melissa J. Azur et al. “Multiple Imputation by Chained Equations: What Is It and How Does It Work?” In: *International Journal of Methods in Psychiatric Research* 20.1 (Feb. 2011), p. 40. DOI: 10.1002/mpr.329. (Visited on 10/30/2024).
- [154] Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Statistics. June 1987. ISBN: 978-0-470-31669-6. DOI: 10.1002/9780470316696. (Visited on 09/19/2024).
- [155] Frank E. Harrell, Kerry L. Lee, and Daniel B. Mark. “MULTIVARIABLE PROGNOSTIC MODELS: ISSUES IN DEVELOPING MODELS, EVALUATING ASSUMPTIONS AND ADEQUACY, AND MEASURING AND REDUCING ERRORS”. In: *Statistics in Medicine* 15.4 (Feb. 1996), pp. 361–387. ISSN: 1097-0258. DOI: 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4. (Visited on 10/30/2024).
- [156] Michael J. Pencina, Ewout W. Steyerberg, and Sr Ralph B D’Agostino. “Extensions of Net Reclassification Improvement Calculations to Measure Usefulness of New Biomarkers”. In: *Statistics in Medicine* 30.1 (Nov. 2010), p. 11. DOI: 10.1002/sim.4085. (Visited on 10/30/2024).
- [157] N. J. D. NAGELKERKE. “A Note on a General Definition of the Coefficient of Determination”. In: *Biometrika* 78.3 (Sept. 1991), pp. 691–692. ISSN: 0006-3444. DOI: 10.1093/biomet/78.3.691. (Visited on 10/30/2024).
- [158] D. R. Cox. “Two Further Applications of a Model for Binary Regression”. In: *Biometrika* 45.3/4 (1958), pp. 562–565. ISSN: 00063444, 14643510. DOI: 10.2307/2333203. JSTOR: 2333203. (Visited on 10/30/2024).
- [159] Peter C. Austin and Ewout W. Steyerberg. “The Integrated Calibration Index (ICI) and Related Metrics for Quantifying the Calibration of Logistic Regression Models”. In: *Statistics in Medicine* 38.21 (2019), pp. 4051–4065. ISSN: 1097-0258. DOI: 10.1002/sim.8281. (Visited on 10/30/2024).
- [160] Gary S. Collins et al. “TRIPOD+AI Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods”. In: *BMJ* 385 (Apr. 16, 2024), e078378. ISSN: 1756-1833. DOI: 10.1136/bmj-2023-078378. pmid: 38626948. URL: <https://www.bmj.com/content/385/bmj-2023-078378> (visited on 09/06/2024).
- [161] Jonathan Bartlett. *Adjusting for Optimism/Overfitting in Measures of Predictive Ability Using Bootstrapping*. Oct. 2014. (Visited on 10/30/2024).

- [162] Jonathan P. Bestwick et al. “Improving Estimation of Parkinson’s Disease Risk—the Enhanced PREDICT-PD Algorithm”. In: *npj Parkinson’s Disease* 7.1 (1 Apr. 1, 2021), pp. 1–7. ISSN: 2373-8057. DOI: 10.1038/s41531-021-00176-9. URL: <https://www.nature.com/articles/s41531-021-00176-9> (visited on 05/30/2022).
- [163] Hang Yuan et al. “Self-supervised learning of accelerometer data provides new insights for sleep and its association with mortality”. In: *NPJ digital medicine* 7.1 (2024), p. 86. (Visited on 10/31/2024).
- [164] Ajit Avasthi et al. “Ethics in Medical Research: General Principles with Special Reference to Psychiatry Research”. In: *Indian Journal of Psychiatry* 55.1 (2013), pp. 86–91. ISSN: 0019-5545. DOI: 10.4103/0019-5545.105525. (Visited on 09/20/2024).
- [165] Sharyl J. Nass et al. “The Value and Importance of Health Information Privacy”. In: *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research*. National Academies Press (US), 2009. (Visited on 09/20/2024).
- [166] Karim Abouelmehdi, Abderrahim Beni-Hessane, and Hayat Khaloufi. “Big Healthcare Data: Preserving Security and Privacy”. In: *Journal of Big Data* 5.1 (Jan. 2018), p. 1. ISSN: 2196-1115. DOI: 10.1186/s40537-017-0110-7. (Visited on 09/20/2024).
- [167] Dariush D. Farhud and Shaghayegh Zokaei. “Ethical Issues of Artificial Intelligence in Medicine and Healthcare”. In: *Iranian Journal of Public Health* 50.11 (Nov. 2021), pp. i–v. ISSN: 2251-6085. DOI: 10.18502/ijph.v50i11.7600. (Visited on 09/20/2024).
- [168] Irene Dankwa-Mullan. “Health Equity and Ethical Considerations in Using Artificial Intelligence in Public Health and Medicine”. In: *Preventing Chronic Disease* 21 (2024). ISSN: 1545-1151. DOI: 10.5888/pcd21.240245. (Visited on 09/20/2024).
- [169] *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 6 September 2024 Laying Down Harmonised Rules on Artificial Intelligence (AI Act) and Amending Certain Union Legislative Acts*. June 13, 2024. URL: <https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A32024R1689> (visited on 10/07/2024).
- [170] *High-Level Summary of the AI Act / EU Artificial Intelligence Act*. Feb. 27, 2024. URL: <https://artificialintelligenceact.eu/high-level-summary/> (visited on 10/07/2024).
- [171] Adrian Gropper. “Open-Source Health Care Software”. In: *AMA Journal of Ethics* 13.9 (Sept. 1, 2011), pp. 632–636. ISSN: 2376-6980. DOI: 10.1001/virtualmentor.2011.13.9.stas1-1109. URL: <https://journalofethics.ama-assn.org/article/open-source-health-care-software/2011-09> (visited on 10/07/2024).
- [172] *Responsible Open Science: An Ethics and Integrity Perspective / Programme / H2020*. CORDIS | European Commission. Apr. 12, 2024. URL: https://cordis.europa.eu/programme/id/H2020_SwafS-30-2020 (visited on 10/07/2024).

- [173] Effy Vayena, Alessandro Blasimme, and I. Glenn Cohen. “Machine Learning in Medicine: Addressing Ethical Challenges”. In: *PLOS Medicine* 15.11 (Nov. 2018), e1002689. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1002689. (Visited on 09/20/2024).
- [174] David Patterson et al. *Carbon Emissions and Large Neural Network Training*. Apr. 2021. DOI: 10.48550/arXiv.2104.10350. arXiv: 2104.10350 [cs]. (Visited on 09/20/2024).
- [175] Nick Paul Taylor. *FDA Rejects Verily Filing for Wrist-Worn Parkinson’s Clinical Trial Device*. MedTech Dive. June 14, 2021. URL: <https://www.medtechdive.com/news/fda-rejects-verily-filing-for-wrist-worn-parkinsons-clinical-trial-device-google/601724/> (visited on 09/19/2024).
- [176] Yuanyuan Chen et al. “Device-Measured Movement Behaviours in over 20,000 China Kadoorie Biobank Participants”. In: *International Journal of Behavioral Nutrition and Physical Activity* 20.1 (Nov. 2023), p. 138. ISSN: 1479-5868. DOI: 10.1186/s12966-023-01537-8. (Visited on 09/21/2024).
- [177] Laurent Servais et al. “Stride Velocity 95th Centile: Insights into Gaining Regulatory Qualification of the First Wearable-Derived Digital Endpoint for Use in Duchenne Muscular Dystrophy Trials”. In: *Journal of Neuromuscular Diseases* 9.2 (2022), pp. 335–346. ISSN: 2214-3599. DOI: 10.3233/JND-210743. (Visited on 09/19/2024).
- [178] Weiliang Qiu et al. *powerSurvEpi: Power and Sample Size Calculation for Survival Analysis of Epidemiological Studies*. Version 0.1.3. Mar. 1, 2021. URL: <https://cran.r-project.org/web/packages/powerSurvEpi/index.html> (visited on 03/23/2025).