Identification of predictors of gait speeds in The Irish Longitudinal Study on Ageing using automated feature selection and explainable machine learning

by

James Davis

A thesis submitted to the University of Dublin, Trinity College Dublin in fulfilment of the requirements for the degree of
Master of Science (Research)
Medical Gerontology

Discipline of Medical Gerontology
School of Medicine
Trinity College Dublin
University of Dublin
2023

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work. I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement. I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

ACKNOWLEDGEMENTS

I would firstly like to give my sincere thanks and gratitude my supervisor Prof. Román Romero-Ortuño and all the FRAILMatics team: Dr. Silvin Knight, Dr. Rossella Rizzo, Dr. Eoin Duggan, and Dr. Feng Xue. Roman's enthusiasm and determination has been unwavering, and I am incredibly grateful for his support both academically and personally. I would also like to specially thank Silvin for all the wonderful chats.

Special thanks to Dr. Ann Monaghan, Dr. Belinda Hernández, Dr. Orna Donohue, and Dr. Raquel Gutiérrez Zúñiga for all your help, advice, and support throughout my journey.

Big thanks to Sean McGrath and Trinity's High Performance Computing team for training and facilitation of the FRAILMatics' High Performance Computing cluster, Tinney.

Thank you to the TILDA team and most notably Prof. Rose Ann Kenny, Lucy Doogan, and Matt McElheron. Thank you also to the lunch time crowd in MISA, particularly Laura Denia-Perez for the wonderful painting sessions.

Of course, a huge thanks to all the TILDA participants. Without them this kind of research would not be possible.

I would like to thank Science Foundation Ireland for funding me as part of the FRAILMatics project.

I would not be where I am today if not for my parents, Helen and Adrian, who have always believed in me. Thank you for all your love and support. Thank you to my partners parents, David and Pamela for their support and letting me live in their house.

Most of all I want to thank my best friend and partner Amy, and our amazing little man, Finn.

This research was funded by Science Foundation Ireland (SFI), grant number 18/FRL/6188. The Irish Longitudinal Study on Ageing (TILDA) is funded by Atlantic Philanthropies, the Irish Department of Health, and Irish Life. Where stated, calculations were performed on the Tinney cluster maintained by the Trinity Centre for High Performance Computing (Research IT) (https://www.tchpc.tcd.ie/hosted#tinney). This cluster is funded form a Grant from Science Foundation Ireland under Grant number 18/FRL/6188.

Ethical approval was granted from the Health Sciences Research Ethics Committee at Trinity College Dublin (granted 9 June 2014 for Wave 3; approval reference "Main Wave 3 Tilda Study") and all participants provided written informed consent. All research was performed in accordance with the Declaration of Helsinki.

PUBLICATIONS AND PRESENTATIONS RELATED TO THIS THESIS

First Author Papers

Comparison of gait speed reserve, usual gait speed, and maximum gait speed of adults aged 50+ in Ireland using explainable machine learning

- James RC Davis, Silvin P Knight, Orna A Donoghue, Belinda Hernández, Rossella Rizzo, Rose Anne Kenny, Roman Romero-Ortuno
- Frontiers in Network Physiology
- 2021
- https://doi.org/10.3389/fnetp.2021.754477

A linear regression-based machine learning pipeline for the discovery of clinically relevant correlates of gait speed reserve from multiple physiological systems

- J Davis, SP Knight, R Rizzo, OA Donoghue, RA Kenny, R Romero-Ortuno
- 2021 29th European Signal Processing Conference (EUSIPCO), 1266-1270
- https://doi.org/10.23919/EUSIPCO54536.2021.9616187

First Author Abstracts

ASSOCIATIONS OF FRAILTY WITH CONSECUTIVE GAIT SPEED TRAILS MEASURED BOTH WITH AND WITHOUT ADDITIONAL STRESSORS

- J Davis, SP Knight, R Rizzo, OA Donoghue, RA Kenny, R Romero-Ortuno
- Age and Ageing 50 (Supplement_3), afab219. 136
- https://doi.org/10.1093/ageing/afab219.136

Posters and Presentations

Irish Gerontological Society Annual Scientific Meeting 2021.

- Presented poster for "ASSOCIATIONS OF FRAILTY WITH CONSECUTIVE GAIT SPEED TRAILS MEASURED BOTH WITH AND WITHOUT ADDITIONAL STRESSORS"

29th European Signal Processing Conference (EUSIPCO) 2021

- submitted poster and video presentation for short paper entitled: "A linear regression-based machine learning pipeline for the discovery of clinically relevant correlates of gait speed reserve from multiple physiological systems" and took part in live Q&A session.

The Irish Longitudinal Study in Ageing All Teams Meeting (03/03/2022, TILDA Offices, Trinity Central, Trinity College Dublin)

- Oral presentation: Explainable machine learning for the investigation of predictors of gait speeds

Discipline of Medical Gerontology Postgraduate Research Day 2022 (01/04/2022, MISA 1st Floor Seminar Room)

- Oral presentation: An Investigation of Physiological Reserve via Performance in Physical Tasks using Explainable Machine Learning

SUMMARY

This work reports the novel application of automated feature selection and explainable machine learning to identify and compare, in participants aged 50 years or over from wave 3 of The Irish Longitudinal Study on Ageing (TILDA), predictors of three gait speed modalities: usual gait speed (UGS), maximum gait speed (MGS), and gait speed reserve (GSR = MGS - UGS). The principal aim of the investigation was to identify which factors were associated with each gait modality, with a comparative focus on GSR. Stepwise feature selection was applied to shortlists of input features covering multiple domains, including demographics, anthropometrics, medical history, cognition, cardiovascular system, physical strength, and sensory and psychological domains.

In a first experiment using data from 2397 participants, a stepwise linear regression-based feature selection algorithm was applied to a shortlist of 34 input features in the prediction of GSR. A mean R_{adj}^2 (SD) 5-fold cross-validation score of 0.16 (0.03) was achieved with 14 variables (with 80% training and 20% test R_{adj}^2 scores of 0.18 and 0.16, respectively). Of the 14 selected features, 11 had statistically significant ($p<0.05$) effects in the model: sex, Montreal Cognitive Assessment (MOCA) score, third level education, chair stands time, age, body mass index (BMI), grip strength, cardiac output, number of medications, fear of falling (FOF), and mean cognitive reaction time (CRT).

In a second experiment, explainable machine learning was applied to an expanded set of 88 input features. Using data from 3925 participants, features were selected by a histogram gradient boosting regression-based stepwise feature selection algorithm. Feature importance and input-output relationships were explored using TreeExplainer from the Shapely Additive Explanations (SHAP) explainable machine learning package. The mean R_{adj}^2 (SD) from 5-fold cross-validation score on training data and the R_{adj}^2 score on test data were: 0.38 (0.04) and 0.41 for UGS; 0.45 (0.04) and 0.46 for MGS; and 0.19 (0.02) and 0.21 for GSR, respectively. Selected features by decreasing SHAP values were education, grip strength, mean

CRT motor reaction time, MOCA errors, age, chair stands time, height, sex, accuracy proportion in the sound-induced flash illusion, FOF, orthostatic intolerance, Mini-Mental State Examination (MMSE) errors, and number of cardiovascular conditions.

Both models selected features across multiple input domains, underscoring the nature of GSR as a measure of individual reserve across multiple physiological systems. In the prediction of GSR, both algorithms identified the importance of prospectively non-modifiable factors such as advancing age, female sex, lower educational attainment, and existing morbidities; but also highlighted potentially modifiable factors such as reduced upper and lower body strength (lower grip strength and longer chair stands time, respectively), lower cognitive (MOCA) and psychomotor performance (CRT), and lower self-efficacy in the psychological domain (fear of falling).

R^2_{adj} scores were marginally higher with machine learning; yet the main advantage of this algorithm over the linear regression-based pipeline is that it allowed for the identification of clinically meaningful non-linearities in the visualised relationship between selected features and GSR. Potential clinical cut-offs and regions of interest for certain features were identifiable, making the models highly interpretable for clinicians.

Although the linear modelling was faster and simpler to use, results suggest that the tree-based explainable machine learning methodology is preferable due to its non-parametric nature and a model explainer such as SHAP that allows for visualisation of input-output relationships. In older adults, the demonstration of GSR is necessary on a daily basis to maintain independent living (e.g., for being able to complete a road crossing or catch a means of public transport). Overall, findings support a network physiology approach to the study of physiological reserve and could help policy makers and clinicians design strategies to promote resilience and functional independence in community-dwelling older adults.

ABBREVIATIONS, ACRONYMS, AND BRIEF EXPLANATION OF FEATURE NAMES

Feature Names

ADLs	Impairments to activities of daily living
Age	Age (years)
Antidepressants	if any antidepressants were being taken
Antihypertensives	if any antihypertensives were being taken
BMI	body mass index (kg/m^2)
CAGE	problematic alcohol use scale: Cutting down, Annoyance by criticism, Guilty feeling, and Eye-openers
CardiacOutput_RS	cardiac output at derived by Finometer during resting state (RS) of active stand test (L/min)
CESD	Centre for Epidemiological Studies Depression scale
ChairStandsTime	Time taken to stand from chair, walk 3 m, turn around, walk 3 m back, and sit down again (s)
CRT_correct	number of correct trails in the choice reaction test
CRT_mean	mean cognitive reaction time in the choice reaction test (ms)
CRT_SD	standard deviation of cognitive reaction times in the choice reaction test (ms)
cs_score_a	contrast sensitivity score at 1.5 cycles per degree line spacing
cs_score_b	contrast sensitivity score at 3 cycles per degree line spacing
cs_score_c	contrast sensitivity score at 6 cycles per degree line spacing
cs_score_d	contrast sensitivity score at 12 cycles per degree line spacing
cs_score_e	contrast sensitivity score at 18 cycles per degree line spacing
dBP_RS	resting state diastolic blood pressure (mmHg)
dBP_RS_SampEn	sample entropy of resting state diastolic blood pressure signal (unitless)
dBP_Seated	seated diastolic blood pressure (mmHg)
dBP_SeatStandDiff	difference in standing and seated blood pressure (mmHg)
dBP_Standing	standing diastolic blood pressure (mmHg)
Edu3	level of educational attainment (3 levels: none/primary, secondary, or third level/higher)
FOF	fear of falling
GripStrength	maximum grip strength from 4 grip strength trails, two on each hand (kg)
HADSA	Hospital Anxiety and Depression Scale: Anxiety subscale
Hearing_SR	self-rated hearing
Height	height (cm)
HHb_RS	mean resting state deoxygenated haemoglobin concentration (µMol/L)
HHb_RS_SampEn	entropy of resting state deoxygenated haemoglobin concentration signal
HR_RS	mean resting state heart rate (bpm)
HR_Mean_Free	mean heart rate during free breathing (bpm)
HR_Mean_Paced	mean heart rate during paced breathing (bpm)

HR_rMSSD_Free	root-mean-square of successive differences between RR heartbeat intervals during free breathing (ms)
HR_rMSSD_Paced	root-mean-square of successive differences between RR heartbeat intervals during paced breathing (ms)
HR_SDNN_Paced	standard deviation of NN heartbeat intervals (ms)
HR_Span_Free	span of heart rate (max HR - min heart rate) during free breathing (bpm)
HR_Span_Paced	span of heart rate (max HR - min heart rate) during paced breathing (bpm)
HR_TotalPower_Free	total spectral power of heart rate during free breathing (ms^2)
HR_TotalPower_Paced	total spectral power of heart rate during paced breathing (ms^2)
HR_rMSSD_PacedFreeDiff	difference between HR_rMSSD_Paced and HR_rMSSD_Free (ms)
HR_RS_SampEn	sample entropy of resting state heart rate signal (unitless)
HR_Seated	seated heart rate (bpm)
HR_SeatStandDiff	difference between standing and seated heart rate (bpm)
HR_Standing	standing heart rate (bpm)
IADLs	Impairments to instrumental activities of daily living
LVET_RS	resting state left ventricular ejection time (ms)
MAP_RS	resting state mean arterial pressure (mmHg)
MAP_RS_SampEn	sample entropy of resting state mean arterial pressure signal
Maxslope_RS	maximum slope of blood pressure vs time graph during resting state (mmHg/s)
Meds	number of medications (excluding supplements)
MMSE_errors	number of errors in Mini-Mental State Examination test
MOCA_errors	number of errors in Montreal Cognitive Assessment
MRT_mean	mean motor response time in choice reaction test (ms)
MRT_SD	standard deviation in motor response times in choice reaction test (ms)
NumCVD	number of cardiovascular diseases
O2_RS_SampEn	sample entropy of resting state oxygenated haemoglobin concentration signal
O2Hb_RS	resting state oxygenated haemoglobin concentration (µMol/L)
PhasicDizziness	whether the participant experienced dizziness upon standing in the active stand test
PulseInterval_RS	mean resting state pulse interval (ms)
PulseWaveVelocity	pulse wave velocity, measure of arterial stiffness (m/s)
SART_Errors	number of errors in the sustained attention to response task
SART_mean	mean reaction time in the sustained attention to response task (ms)
SART_SD	standard deviation in sustained attention to response task reaction times (ms)
sBP_RS	mean resting state systolic blood pressure (mmHg)
sBP_RS_SampEn	sample entropy of resting state systolic blood pressure signal
sBP_Seated	seated systolic blood pressure (mmHg)
sBP_SeatStandDiff	difference between standing and seated systolic blood pressure (mmHg)
sBP_Standing	standing systolic blood pressure (mmHg)
Sex	sex (female = 1)

Shams_2B1F_150	proportion of correct answers in the 2 beep 1 flash sound flash illusion test with the flash-beep pair 150 ms before the single beep
Shams_2B1F_230	proportion of correct answers in the 2 beep 1 flash sound flash illusion test with the flash-beep pair 230 ms before the single beep
Shams_2B1F_70	proportion of correct answers in the 2 beep 1 flash sound flash illusion test with the flash-beep pair 70 ms before the single beep
Shams_2B1F_m150	proportion of correct answers in the 2 beep 1 flash sound flash illusion test with the flash-beep pair 150 ms after the single beep
Shams_2B1F_m230	proportion of correct answers in the 2 beep 1 flash sound flash illusion test with the flash-beep pair 230 ms after the single beep
Shams_2B1F_m70	proportion of correct answers in the 2 beep 1 flash sound flash illusion test with the flash-beep pair 70 ms after the single beep
Smoker	smoking status: never, past, current
StrokeVolume_RS	mean resting state stroke volume (mL)
TPR_RS	mean resting state total peripheral resistance (dyn·s·cm^{-5})
TSI_RS	mean resting state tissue saturation index (%)
TSI_RS_SampEn	sample entropy of resting state tissue saturation index signal
UCLA	University of California Los Angeles Anxiety scale
VisualAcuity	visual acuity of best eye
VisualAcuityLeft	visual acuity of left eye
VisualAcuityRight	visual acuity of right eye
WaistHipRatio	ratio of waist circumference to hip circumference (cm)
Weight	weight (kg)

Other Acronyms and Abbreviations

AS	active stand
BPM	beats per minute
CAPI	computer-assisted personal interview
CI	confidence intervals
cpd	cycles per degree
CRT	cognitive reaction time (in the choice reaction time test) (ms)
CV	cross-validation
GBT	gradient boosting trees
GSR	gait speed reserve (cm/s)
HGBR	Histogram gradient boosting regression
HR	heart rate
HRV	heart rate variability
ICE	individual conditional expectation
IQR	Interquartile range
LIME	Local Interpretable Model-agnostic Explanations
MGS	maximum gait speed (cm/s)
MMSE	Mini-Mental State Examination
MOCA	Montreal Cognitive Assessment

MRT	motor reaction time (in the choice reaction time test)
NIRS	near infrared spectroscopy
PDP	partial dependence plot
RANSAM	random sampling
RS	resting state
SampEn	sample entropy
SART	sustained attention to reaction test
SCQ	self-completion questionnaire
SHAP	Shapely Additive Explanations
SIFI	sound induced flash illusion test
TILDA	The Irish Longitudinal Study on Ageing
TUG	timed up and go
UGS	usual gait speed (cm/s)

TABLE OF CONTENTS

1 INTRODUCTION

Gait speed is a measure of general fitness [1]; faster gait speed is associated with the ability to meet occupational demands in younger adults [2], while slower gait speed is associated with functional decline and morbidity in older adults [3, 4]. Even though usual (or comfortable) gait speed (UGS) and maximum gait speed (MGS) are significantly intercorrelated [5], changing from comfortable to maximum speed requires a general effort across many body systems. The difference between these two gait speeds has been referred to as walking speed reserve or gait speed reserve (GSR) [6].

UGS is a commonly measured gait characteristic in clinical practice and has well established associations with age [7], physical function [8], and frailty [9]. On the other hand, MGS has been associated with physical and cognitive function [2, 10]. Gait speed reserve (GSR) may be a useful proxy measure of physiological reserve in humans. For example, some studies have suggested that in community-dwelling older adults, the simultaneous consideration of both usual and maximum gait speed could increase the specificity of the identification of frailty [11, 12]. The health associations of these three modalities of gait speed (UGS, MGS, GSR) are somewhat different but there appears to have been no systematic attempts previously to model predictors of GSR in a large representative sample of community-dwelling older adults where many demographic, anthropometric and clinical features are measured across multiple physiological systems. In older adults, the demonstration of GSR is necessary on a daily basis to maintain independent living (e.g., for being able to complete a road crossing or catch a means of public transport). Therefore, the identification of potentially modifiable factors impacting on GSR could be important for strategies to promote resilience and functional independence in community-dwelling older adults.

The aim in using the three chosen gait parameters is to cover a range of aspects of an older adults physical, cognitive, and psychological condition. Usual gait speed generally assesses how one walks normally, i.e., hopefully without too much effort. This encompasses both routine physical ability and one's psychological idea of what a normal comfortable walking speed is. Maximum gait speed on the other hand seeks to investigate the limit of one's ability. This limit may be more physically or mentally constructed. Given the link between mind and body even if one limit comes before the other, this may eventually result in the other factor atrophying due to lack of use. Gait speed reserve then acts as a potential proxy for overall physiological reserve. Does an individual have much extra to give on top of their normal performance?

The principal aim of this study was to investigate, in community-dwelling participants aged 50 years or over from wave 3 of The Irish Longitudinal Study on Ageing (TILDA), if UGS, MGS, and/or GSR were predicted by factors from multiple domains pertaining to individuals (e.g., physical, cognitive, psychological, socio-demographic) and if so, whether differences existed between the three gait modalities in terms of what factors predicted each one.

Two experiments were conducted. Experiment 1 consisted of a linear regression-based stepwise feature selection procedure that predicted GSR using a first iteration of manually selected input features across domains. Experiment 2 expanded the set of input features and employed a more sophisticated stepwise feature selection and explainable machine learning methodology to predict not only GSR, but also UGS and MGS. Experiment 2 also further compared UGS, MGS, and GSR in terms of their statistical relationships to the clinically relevant variables of past and future falls and faints. In experiment 2, machine learning was used to first, identify the set of features that, from the initial shortlist of input features, best described UGS, MGS, and GSR; the shortlist of features was theory-driven and not purely exploratory, that is, features were selected that might have physiological plausibility. Then, using explainable machine learning methods, the selected models for UGS, MGS, and GSR were investigated to observe how each feature in the model was associated with the output in a non-parametric manner. In the context of the selected features and visualisations of the input-output relationships, clinical interpretations were then discussed with respect to the cohort studied and the hypothesis that UGS, MGS and GSR are multisystem-driven phenomena.

A comparison of strengths and weaknesses of using a more sophisticated explainable machine learning methodology over a simpler linear regression-based analysis in the context of exploring predictors of GSR was another main aspect of this work.

Explainable machine learning is an ever-growing field [13]. At the most basic level, the analysis of model coefficients from linear or polynomial models is widely implemented; in them, it is necessary to know the strength and direction of the associations between inputs and output. Forest plots visually express model coefficients and confidence intervals.

The growth of more complex machine learning methods has come with the problem of less transparency and explainability. Global feature importance can be assessed via permutation testing, in which each feature is assessed by randomly shuffling its values across samples and measuring the change in performance; if there is no drop (or an increase) in performance, then the order in which the values are distributed across the samples is not important and therefore neither is the feature. On the other hand, shuffling the values of an important feature should decrease the performance of the model. This approach can however be very misleading in the presence of collinearities.

Tree-based models have an inherent explainability in that they are rule-based; a single decision tree can be totally explained by following the tree from top to bottom. The method quickly becomes impractical with increasing size of trees and the use of tree ensembles (e.g., random forests, gradient boosting). On single decision trees, caution is required in interpreting them as they are prone to overfitting and high variance across datasets. Other tree-specific methods assess feature importance by assessing how each feature contributes to making splits in the data. There are various methods that determine the number of split points, the improvement resulting from the split, the number of samples associated with splits, and the distance of the split from the tree root (how early in the tree the split was made). Such metrics are quick to calculate but have drawbacks in that they only consider training data, and they are prone to bias towards high cardinality features (features with many values) as they have greater potential for split points.

Partial dependence plots (PDP) are a powerful model-agnostic visual tool first presented by Friedman alongside his introduction of gradient boosting machines [14]. PDPs use a partial dependence function to calculate the marginal contribution feature (i.e., the difference in model output with and without the feature in question) of a feature with all the other features held at their means. This allows for a visual estimation of the input-output relationships. Two-dimensional contour PDPs are also very popular and allow for the interaction between two features to be explored.

In a similar manner, Individual Conditional Expectance (ICE) plots [15] visualise how a range of values for a feature impacts the output for a given sample, i.e., with the other features held constant at the values that the sample holds. ICE plots offer more granularity than PDPs, and ICE plots from a subset of samples are often displayed together. Where PDPs display the mean effect of a feature, ICE shows the marginal effect of a feature on a specific sample.

Another recent approach to explainability of is Local Interpretable Model-agnostic Explanations (LIME) [16]. Being model agnostic is only requiring the model itself, and the input and output values. LIME attempts to build simple-interpretable surrogate models at the local level that can be used to explain a more complex model. To explain why a model made a prediction for a given sample, LIME will take a subset of samples surrounding the sample of interest and build a linear model to predict the outcome. The complexity of the surrogate model can be increased to give a higher fidelity or decreased to reduced fidelity and increase simplicity. Hyperparameters such as the number of features used in the surrogate model, the type of model (linear regression, RIDGE, LASSO), and the number of samples surrounding the sample of interest, are all hyperparameters that can be adjusted to aim for different levels of explanation. However, herein lies the cons of this approach; changing these hyperparameters can lead to different explanations of what the model is supposedly doing, which leads to a lack of confidence in the approach. The use of a linear surrogate model may also not be appropriate for some models.

The approach utilised in this work is SHapely Additive exPlanations (SHAP) [17, 18]. SHAP is based on Lloyd Shapley's work during the 1950s in the field of cooperative game theory Shapley [19]. Shapley derived a method of fairly attributing worth to players in a game based on their marginal contributions. When applied to machine learning, the game is the model, and the players are the features in the model. The values quantifying the worth are called Shapley values and are a unique solution to a method that will simultaneously satisfy conditions of fairness.

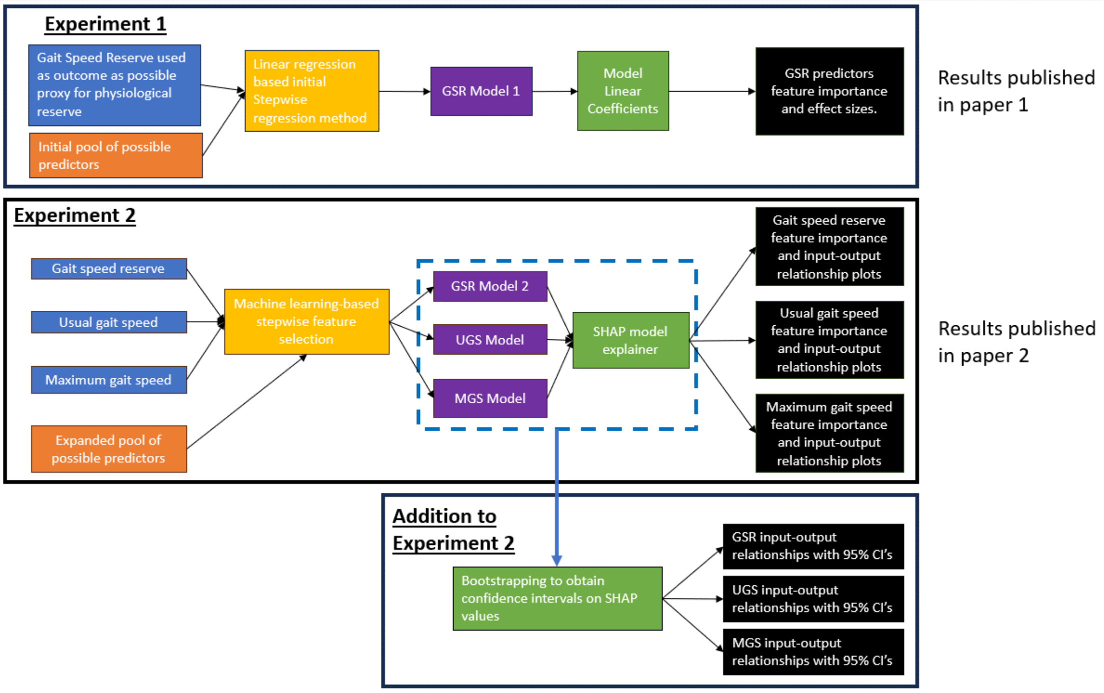
Other analytical tools such as network analysis are available to explore datasets such as the one used in this work. They could of course be used to probe different aspects of the data. Network analysis for example would be less suited to the specific task of exploring how the features relate directly to the gait variables. Instead, it would reveal how, and to what degree, each feature in the dataset relates to all the others. From this, clusters or groups of more related feature may emerge, and conversely some features might not be well connected to others. As with the methods used in this work, there are numerical and visual ways to present those results.

The thesis begins with an introduction and is followed by a section that describes the materials and methods for the two main experiments. Next, the results and discussion for experiment 1, which are taken from my first published paper [20], are presented. The following section presents the results and discussion from experiment 2, which is based on my second published paper [21]. After this, the results from work that adds bootstrapped confidence intervals onto the models from experiment 2 are presented. A discussion focusing on the two GSR models from the two methodologies follows. Finally, conclusions covering all experiments and methodologies are provided. Some supplementary materials relating to experiment 2 are given in an appendix.

2 MATERIALS AND METHODS

2.1 OVERVIEW OF METHODS, ORDER OF EXPERIMENTS, AND THESIS STRUCTURE

The order of experiments used in this thesis is show in Figure 1. This is meant to be used as an aid in following the thesis and not necessarily meant to be fully understood at this point. Experiment 2 builds upon the dataset and methodology used in Experiment 1. Then after the results of paper 2 were published, an additional analytical step was taken.



Experiment 1
Gait Speed Reserve used as outcome as possible proxy for physiological reserve
Initial pool of possible predictors
Linear regression based initial Stepwise regression method
GSR Model 1
Model Linear Coefficients
GSR predictors feature importance and effect sizes.
Results published in paper 1
Experiment 2
Gait speed reserve
Usual gait speed
Maximum gait speed
Expanded pool of possible predictors
Machine learning-based stepwise feature selection
GSR Model 2
UGS Model
MGS Model
SHAP model explainer
Gait speed reserve feature importance and input-output relationship plots
Usual gait speed feature importance and input-output relationship plots
Maximum gait speed feature importance and input-output relationship plots
Results published in paper 2
Addition to Experiment 2
Bootstrapping to obtain confidence intervals on SHAP values
GSR input-output relationships with 95% CI's
UGS input-output relationships with 95% CI's
MGS input-output relationships with 95% CI's

Figure 1. Overview of thesis experiments and results.

The thesis is structured such that for each section (i.e., features used in experiments, analytical methods, and results) experiment one is presented first, with experiment two following such that instead of repeating anything from experiment one, there is a focus on what is new or different. This is intended to create a better flow when describing how the data and methods evolved throughout my studies.

2.2 Data

2.2.1 The Irish Longitudinal Study on Ageing

The data for the experiments conducted in this thesis were leveraged from The Irish Longitudinal Study on Ageing (TILDA). TILDA is a large-scale, nationally representative longitudinal study on ageing in Ireland. The aspiration of TILDA is to make Ireland the best place in the world to grow old by conducting research that can positively influence health, medicine, society, and policy. Wave 1 of the study took place between October 2009 and February 2011 with the recruitment of 8175 community-dwelling individuals aged 50 years or over. Additional 329 individuals who were younger than 50 years were also included as they were the partner or spouse of a participant. Participants were chosen randomly from a database of all Irish residential addresses using the RANSAM sampling procedure [22]. So far, five waves and an additional "COVID wave" have been completed. Of the five main waves, waves 1 and 3 included a health centre assessment. Wave 6 is currently in recruitment phase and also includes a health centre assessment. In every wave, all participants underwent a computer-assisted personal interview (CAPI) and a paper-based self-completion questionnaire (SCQ) [22]. These assessed demographics, medical history, psychology, social, familial, and financial aspects, some cognitive and physical performance, and simple cardiovascular health. In waves 1 and 3 (and currently in wave 6), participants were also invited to attend a more extensive health centre assessment in either Cork or Dublin (wave 1) or Dublin (wave 3). The health centre assessment included objective tests of gait, physical performance and strength, cognition, psychology, senses, bloods, cardiovascular health, and bone density. The present work is based on data collected at wave 3, in which 4309 of a total of 6694 who took part in the wave also attended the health centre assessment. Whilst UGS was measured at both waves 1 and 3, MGS (which is required for the calculation of GSR) was only measured at wave 3.

Ethical approval was granted from the Health Sciences Research Ethics Committee at Trinity College Dublin (granted 9 June 2014 for Wave 3; approval reference "Main Wave 3 Tilda Study") and all participants provided written informed consent. All research was performed in accordance with the Declaration of Helsinki.

2.2.2 Analytical Sample

For both experiments, the analytical sample was obtained from wave 3 of TILDA, because this is the only wave so far with MGS data.

For experiment 1, the analytical sample consisted of participants aged 50 years or older with GSR data and no missing values in the shortlist of input features.

For experiment 2 the analytical sample consists of those aged 50 years or older with data for UGS and MGS (and therefore also for GSR since GSR=MGS-UGS).

2.2.3 Gait Speed Measures

At wave 3 of TILDA, gait speed was measured as part of a health centre assessment. Measurements in units of cm/s were made using a 4.88 m computerised walkway (GAITRite, CIR Systems, NY, USA). A two-meter space before and after the walkway was used for acceleration and deceleration. Participants were first asked to walk at their normal (usual) pace: UGS; and then as fast as they safely could: MGS. Two walking trials were obtained in each condition and the mean value for each was used in this analysis. GSR was defined as MGS – UGS.

2.2.4 Shortlisted Features

2.2.4.1 Socio-demographics, Anthropometrics, Lifestyle, Disability and Medical History

Table 1. Socio-demographics, Anthropometrics, Lifestyle, Disability and Medical History input features included in experiment 1 and experiment 2. (Brief explanations of feature names can be found on page ix).

Experiment 1	Experiment 2
Age	Age
Sex	Sex
Edu3	Edu3
BMI	BMI
	Height
	Weight
	WaistHipRatio
Smoker	Smoker
	CAGE
	ADL
IADL	IADL
nMeds	nMeds
Antidepressants	Antidepressants
Antihypertensives	Antihypertensives
nCVD	nCVD

2.2.4.1.1 Experiment 1

Socio-demographic information included age in years, sex (male=0; female=1), and level of educational attainment (Edu3): either primary/none (Edu3=1), secondary (Edu3=2), or tertiary/higher (Edu3=3).

Body mass index (BMI, in kg/m^2) was included as an anthropometric measure.

Lifestyle included smoking status (Smoker), which was coded as: never=0, past=1, or current=2. Disability included the number of impairments to instrumental activities of daily living (IADL): preparing a hot meal; doing household chores (laundry, cleaning); shopping for groceries; making telephone calls; taking medications; and managing money such as paying bills and keeping track of expenses.

Self-reported medication use was included via the number of medications (excluding supplements) currently being taken (nMeds), and whether antidepressants or antihypertensives were being taken.

Self-reported number of cardiovascular diseases (nCVD) was also included. The cardiovascular diseases included for counting are high BP, angina, heart attack, congestive heart failure, high cholesterol, heart murmur, and abnormal heart rhythm.

2.2.4.1.2 Experiment 2

The features in the experiment 2 shortlist were the same as in experiment 1 with the addition of height, weight, waist-hip ratio, the CAGE scale for assessing problematic alcohol use (Cutting down, Annoyance by criticism, Guilty feeling, and Eye-openers) [23], and impairments with respect to personal activities of daily living (ADLs): dressing, including putting on shoes and socks; walk across a room; bathing or showering; eating, such as cutting up food; getting in or out of bed; and using the toilet, including getting up or down.

2.2.4.2 Cardiovascular System

Table 2. Cardiovascular input features included in experiment 1 and experiment 2. Under experiment 2, the cardiovascular features are sub-categorised into Finometer, NIRS (near infra-red spectroscopy), HRV (heart-rate variability), sphygmomanometer, and pulse wave velocity. See page ix for Abbreviations, Acronyms.

Experiment 1	Experiment 2				
	Other	Sphyhmomanometer	Heart Rate Variability (HRV)	NIRS Derived	Finometer
PulseWaveVelocity	PulseWaveVelocity	sBP_Seated	HR_Mean_Free	O2Hb_RS	sBP_RS
sBP_Seated	PhasicDizziness	dBP_Seated	HR_rMSSD_Free	HHb_RS	dBP_RS
HR_Seated		HR_Seated	HR_TotalPower_Free	TSI_RS	MAP_RS
sBP_SeatStandDiff		sBP_Standing	HR_Mean_Paced		HR_RS
MAP_RS		dBP_Standing	HR_SDNN_Paced	TSI_RS_SampEn	StrokeVolume_RS
sBP_AS_NadirDelta		HR_Standing	HR_rMSSD_Paced	O2_RS_SampEn	LVET_RS
Lvet_RS		sBP_SeatStandDiff	HR_TotalPower_Paced	HHb_RS_SampEn	PulseInterval_RS
CardiacOutput_RS		dBP_SeatStandDiff	HR_rMSSD_PacedFreeDiff		Maxslope_RS
PhasicDizziness		HR_SeatStandDiff	HR_Span_Free		CardiacOutput_RS
sBP_RS_SampEn			HR_Span_Paced		TPR_RS
HR_RS_SampEn					
					sBP_RS_SampEn
					dBP_RS_SampEn
					MAP_RS_SampEn
					HR_RS_SampEn

2.2.4.2.1 Experiment 1

The cardiovascular features included in experiment 1 were all included in experiment 2 and will just be stated here. An explanation is given in the following section detailing the experiment 2 features.

The cardiovascular features included in experiment 1 were seated systolic blood pressure (sBP_Seated), seated heart rate (HR_Seated), and the difference between seated and standing systolic blood pressure (sBP_SeatStandDiff), all measured using an oscillometric sphygmomanometer; as well as mean arterial pressure (MAP_RS), left ventricular ejection time (LVET_RS), and cardiac output (CardiacOutput_RS), all measured using a Finometer during the baseline resting state of the active stand test in addition, input features included the change in systolic blood pressure between supine rest and the lowest point post-stand as measured with the Finometer (sBP_AS_NadirDelta), sample entropy of the systolic blood pressure and heart rate signals during the baseline resting state of the active stand test (sBP_RS_SampEn and HR_RS_SampEn) [24], pulse wave velocity, and whether or not the participant experienced dizziness during standing as self-reported after the active stand test (PhasicDizziness).

2.2.4.2.2 Experiment 2

During the health centre assessment, a cardiovascular assessment was conducted. An overview of the assessment is as follows. Participants were asked to lie supine on a bed and rest for 10 mins. After those 10 minutes, the heart rate variability (HRV) tests were conducted. First, HRV data were recorded for 5 minutes of free breathing (i.e., spontaneous breathing or breathing at whatever rate feels comfortable) followed by 5 minutes of paced breathing guided by an audio track delivered via headphones. After the HRV tests, pulse wave velocity was measured non-invasively. Following this, the active stand phase began with another 10 minutes of supine rest. After 10 minutes, the participant was asked to stand up as quickly as it was safely possible and remain standing still for three minutes, after which the test finished. The participant was then asked if they experienced any dizziness or light-headedness.

The resting state (RS) cardiovascular measurements in the feature shortlist were made during an approximate 10-minute window in which the participant was laying supine in a comfortably lit room at an ambient temperature of between 21 °C and 23 °C. The full TILDA active stand protocol in which the RS window takes place has been detailed elsewhere [24-26]. Throughout the RS stage, participants underwent non-invasive continuous haemodynamic monitoring, recorded at 200Hz using a Finometer MIDI device (Finapres Medical Systems BV, Amsterdam, the Netherlands). All RS parameters selected for the shortlist were mean values from the last minute of supine rest [24]. Hemodynamic parameters were: systolic blood pressure (sBP_RS), diastolic blood pressure (dBP_RS), and mean arterial pressure (MAP_RS), all in units of mmHg; heart rate (HR_RS) in bpm; stroke volume (StrokeVolume_RS) in mL; left ventricular ejection time (LVET_RS) in ms; pulse interval (PulseInterval_RS) in ms; maximum slope (Maxslope_RS) in $\mathrm{mmHg/s}$; cardiac output (CardiacOutput_RS) in $\mathrm{L/min}$; and total peripheral resistance (TPR_RS) in $\mathrm{dyn} \cdot \mathrm{s} \cdot \mathrm{cm}^{-5}$. A near-infrared spectroscopy (NIRS) device (Portalite; Artinis Medical Systems, Zetten, the Netherlands), attached over the participants' left frontal lobe area, was also employed during the RS and the following cerebral oxygenation features were extracted, again as the mean values from the final minute of rest: oxygenated haemoglobin concentration (O2Hb_RS) and deoxygenated haemoglobin concentration (HHb_RS), both in units of $\mu\mathrm{Mol/L}$; and tissue saturation index (TSI_RS), as a percentage [24]. Previously derived sample entropy values (a measure of the amount of disorder in the signals) for resting sBP (sBP_RS_SampEn), dBP (dBP_RS_SampEn), MAP (MAP_RS_SampEn), heart rate (HR_RS_SampEn), O2Hb (O2Hb_RS_SampEn), HHb (HHb_RS_SampEn), and tissue saturation index (TSI_RS_SampEn) were also shortlisted [24]. In addition, participants were asked if they experienced dizziness upon standing (PhasicDizziness: yes or no), and this feature was also included in the shortlist.

Resting heart rate variability measures were also shortlisted; these were obtained in two five-minute blocks as detailed elsewhere [27]. In short, for each block, participants were laying supine. In the first block, participants were asked to breath spontaneously (free breathing), and in the second block, they were asked to breathe according to a pre-recorded set of auditory instructions (paced breathing at a frequency of 0.2 Hz). Measurements were obtained using a 3-lead electrocardiogram (Medilog Darwin, Oxford Instruments Medical Ltd, UK). The data were subject to a 0.01 – 1000 Hz band-pass filtering before R peak detection was performed with a proprietary software [28]. The data collection and processing are described in detail elsewhere [27]. Time domain features were: mean heart rate in bpm; root-mean-square of successive differences between RR intervals in ms; standard deviation of NN intervals in ms; and difference between maximum and minimum heart rate in bpm, derived for both free (HR_Mean_Free, HR_rMSSD_Free, HR_SDNN_Free, HR_Span_Free) and paced breathing (HR_Mean_Paced, HR_rMSSD_Paced, HR_SDNN_Paced, HR_Span_Paced). The difference between free and paced breathing values was calculated for rMSSD (HR_rMSSD_PacedFreeDiff). In the frequency domain, total spectral power in the 0 - 0.4 Hz frequency band was measured for both free (HR_TotalPower_Free) and paced breathing (HR_TotalPower_Paced) in units of milliseconds squared, ms^2.

sBP, dBP and HR were also determined in a more conventional manner using a sphygmomanometer in seated (sBP_Seated, dBP_Seated, and HR_Seated) and standing (sBP_Standing, dBP_Standing, and HR_Standing) positions; all with units of mmHg. The difference between seated and standing values were calculated for each of the measures (sBP_SeatStandDiff, dBP_SeatStandDiff, and HR_SeatStandDiff).

Pulse wave velocity (PulseWaveVelocity), a non-invasive measure of arterial stiffness with units of m/s, was also included as a cardiovascular feature. In TILDA, the average of two measurements between the carotid and femoral arteries (in m/s) was obtained using a Vicorder® (SMT medical GmbH & Co. Wuerzburg, Germany). Full details have been described elsewhere [29, 30].

2.2.4.3 Physical Strength

Table 3. Physical strength features included in experiment 1 and experiment 2.

Both Experiments
Grip Strength
Chair Stands Time

Upper and lower body strength were assessed via grip strength and chair stands time. Grip strength was measured in kg using a hydraulic hand dynamometer (Baseline®, Fabrication Enterprises, Inc., White Plains, NY, USA). The value for grip strength (GripStrength) was taken as the maximum value from a total of eight measurements with four made on each hand. Lower body strength was assessed using the chair stands test, in which the time (in seconds) was recorded for the participants to complete five chair stands as quickly as possible, keeping the arms folded across their chest (ChairStandsTime). Chair height was 46 cm.

2.2.4.4 Cognitive and Psychological

Table 4. Cognitive and psychological features included in experiment 1 and experiment 2. (Brief explanations of feature names can be found on page ix).

Experiment 1	Experiment 2
MOCA	MOCA_errors
MMSE	MMSE_errors
CRT_mean	CRT_mean
	CRT_SD
CRT_correct	CRT_correct
	MRT_mean
	MRT_SD
SART_mean	SART_mean
SART_SD	SART_SD
	SART_errors
	CESD
HADSA	HADSA
UCLA	UCLA
FOF	FOF

2.2.4.4.1 Experiment 1

Global cognition was assessed using two paper-based assessments: the Montreal Cognitive Assessment (MOCA) [31] and the Mini-Mental State Examination (MMSE) [32]. Concentration and cognitive processing time were assessed using two computer assisted tasks: the choice reaction task (CRT) [33] and the sustained attention to response task (SART) [34]. The CRT required participants to hold down a central button until an on-screen stimulus (either the word “YES” or “NO”) appeared, at which time they had to press the corresponding button on a keyboard. After pressing either button, participants were then required to return to the central button to continue. This was repeated approximately 100 times. In the SART test, participants watched a screen that displayed the numbers 1 – 9 sequentially a total of 23 times. A number appeared for 300 ms with an interval of 800 ms between numbers: the entire trial lasts approximately four minutes. Participants were instructed to press a button at the appearance of every number except for a specific number (i.e., 3). From this the mean cognitive reaction time (CRT_mean) and the number of correct CRT presses (CRT_correct) were extracted. CRT is the time taken to release the central button in response to the stimulus. From the SART, mean and standard deviation of reaction time (SART_mean, SART_SD), and number of trails in which the participant pressed the button when the number 3 appeared (SART_errors) were extracted. CRT and SART times are both measured in milliseconds.

The psychological domains of anxiety, and loneliness were assessed the Hospital Anxiety and Depression Scale – Anxiety subscale (HADSA), and the UCLA Loneliness Scale (UCLA), respectively. Fear of falling (FOF) was determined with a yes or no question [30].

2.2.4.4.2 Experiment 2

For experiment 2, a variation on the MOCA and MMSE features was used that instead of the number of correct answers (i.e., total score) indicated the number of errors made: MOCA_errors and MMSE_errors.

Another aspect of the CRT is the motor reaction time. Where the cognitive reaction time is the time between stimulus and releasing the central button, the motor reaction time (MRT) is the time between releasing the central button and pressing the required button. The mean and standard deviation of the motor reaction time (MRT_mean and MRT_SD) across all trails were included. Other additions were the standard deviation in CRT (CRT_SD) and number of errors in the SART (SART_errors).

Depressive symptoms were also added in the form of the Center for Epidemiologic Studies Depression Scale (CESD) [35].

2.2.4.5 Sensory

Table 5. Sensory features included in experiment 1 and experiment 2. In experiment 2, the sensory features are sub-categorised into Hearing, Visual Acuity, Contrast Sensitivity, and Shams SIFI. (Brief explanations of feature names can be found on page ix).

Experiment 1	**Experiment 2**			
	Hearing	**VisualAcuity**	**Contrast Sensitivity**	**Shams SIFI**
Hearing_SR	Hearing_SR	VisualAcuityLeft	cs_score_a	Shams_2B1F_m70
VisualAcuity		VisualAcuityRight	cs_score_b	Shams_2B1F_m150
		VisualAcuityBest	cs_score_c	Shams_2B1F_m230
			cs_score_d	Shams_2B1F_70
			cs_score_e	Shams_2B1F_150
				Shams_2B1F_230

2.2.4.5.1 Experiment 1

In experiment 1, only two basic sensory features were included: self-rated hearing (Hearing_SR) and visual acuity. Hearing_SR was ascertained by the question: “Is your hearing (with or without a hearing aid): 1. Excellent, 2. Very good, 3. Good, 4. Fair, or, 5. Poor?”. Visual acuity was measured for both eyes using a LogMar chart and the visual acuity for the best eye defined as $100 - (\min([VisualAcuityLeft, VisualAcuityRight]) \times 50$.

2.2.4.5.2 Experiment 2

For experiment 2, the range of sensory features was expanded to include additional visual acuity for each eye separately, contrast sensitivity scores, and multi-sensory integration assessment.

Contrast sensitivity (CS) was measured at five spatial frequencies; in cycles per degree (cpd) they were: 1.5 cpd (cs_score_a), 3 cpd (cs_score_b), 6 cpd (cs_score_c), 12 cpd (cs_score_d), and 18 cpd (cs_score_e). The procedures for visual acuity and contrast sensitivity measurements are described in detail elsewhere [36]

Multisensory integration was measured using the Shams sound-induced flash illusion (SIFI) test [37]. The procedure used in TILDA is described in more detail elsewhere [38]; but in short, participants were subjected to a set of beeps and flashes and asked to report how many flashes they perceived. Five general flash-beep combinations were presented to the participants: 2 beeps + two flashes; 1 beep + 1 flash; 0 beeps + 1 flash; 0 beeps + 2 flashes; and 2 beeps + 1 flash. The flash-beep configurations used in this analysis are the so-called 'illusory' 2 beep 1 flash (2B1F) trials. In 2B1F trials, the flash is synchronous with one of the beeps; the other beep occurred either 70 ms, 150 ms, or 230 ms before (SIFI_2B1F_70, SIFI_2B1F_150, SIFI_2B1F_230) or after (SIFI_2B1F_m70, SIFI_2B1F_m150, SIFI_2B1F_m230) the flash-beep pair. SIFI susceptibility represented accuracy for judging how many flashes were presented when one flash was presented with two beeps (2B1F). Lower accuracy, judging one flash as two, thus indicates higher SIFI susceptibility and stronger integration. SIFI susceptibility was expressed as proportion correct. As there were two trials per condition, these variables were considered discrete (i.e., participants scored 0, .5 or 1 proportion correct) [39].

2.2.5 Falls and Faints Features – Experiment 2

To further compare UGS, MGS and GSR in a clinical outcome context, their correlations with both historical and future falls and syncope were assessed.

Historical fallers/fainters were defined as participants who reported at least one fall/faint in the year prior to wave 3.

Future fallers/fainters were defined as those who reported at least one fall/faint between wave 3 and wave 5 (approximately four years later).

Each of those four variables were binary categorical with occurrence of falls/faints coded as ‘1’ and absence coded as ‘0’.

2.3 METHODS

The feature selection process used was built upon across the two experiments. In experiment 1, the stepwise feature selection framework was implemented using a simple linear regression as the core regression technique. Experiment 2 expanded on this mainly by changing the core regression method to a machine learning regression. The following sections will describe the feature selection for both experiments, the statistical methods used to compare past and future falls and faints using the three gait variables, histogram gradient boosting regression, the Shapely Additive Explanations (SHAP), and bootstrapping confidence intervals.

All operations were performed using Python 3. The feature selection was executed on the Tinney High Performance Computing Cluster at Trinity College Dublin (https://www.tchpc.tcd.ie/hosted#tinney). The Tinney cluster is maintained by the Trinity Centre for High Performance Computing (Research IT). This cluster is funded form a Grant from Science Foundation Ireland under Grant number 18/FRL/6188.

2.3.1 Feature selection – experiment 1

The data were divided via an 80/20 train/test split. From there, using the training data, an automatic feature selection was employed that utilised a stepwise approach in which features were added to a linear regression model one at a time such that they maximised the mean adjusted r-squared , $\overline{R^2_{adj}}$, which was calculated as the mean R^2_{adj} value obtained from a 5-fold cross-validation (CV). The 5-fold CV was introduced to reduce overfitting to the entire dataset and help identify the features that performed best across multiple subsets of the data. On each iteration of the CV, a pipeline consisting of a standard scaler and a linear regression was employed. The features selected for the final model were those corresponding to the peak of a $\overline{R_{adJ}^2}$ vs. added features plot. The selected features were used to train a model on the training data. The model was tested on the remaining 20% of the data. This process was implemented using the Scikit-learn package (v0.19.1) for Python (v3.8.3).

2.3.2 Feature selection – experiment 2

All operations were performed using Python 3. The feature selection was executed on the Tinney High Performance Computing Cluster at Trinity College Dublin.

The main process of the stepwise feature selection is the same as in the previous section but this time, a histogram gradient boosting regressor was used instead of a simple linear regression and at the cross validation step a 100-iteration hyperparameter search was performed, whereby each iteration was subject to a 5-fold CV. Standard scaling was not applied as it was not necessary. The hyperparameter tuning aims to find better performing hyperparameters for the HGBR model and is in the form of a 100-iteration randomised search over a set of predefined hyperparameter distributions:

```
{'max_iter': [2000],

'loss': ['least_squares'],

'random_state': [42],

'early_stopping': [True],

'learning_rate': loguniform(0.005, 0.1),

'max_leaf_nodes': randint(2, 10),

'min_samples_leaf': randint(100,200)}
```

The evaluation metric employed was adjusted-R^2 (R^2_{adj}). This metric is used to avoid the continual increase in R^2 that occurs with the addition of new features regardless of whether they significantly increase the variance explained by the model. For each temporary model, the best parameters are chosen based on the mean R^2_{adj} from CV ($\overline{R^2_{adj}}$). After selecting a new feature, a process then checks to see if any of the previously selected feature have been made redundant, and if so, they are removed.

For the purpose of performance monitoring, on each iteration of the loop, the current best model is fit to the entire training dataset and evaluated on both the training and test sets to give training and test R^2_{adj} scores. These scores do not influence the feature selection.

2.3.3 Statistical Associations Between Gait Speed Modalities and Faller/Fainter Status

The normality of the distribution of the three gait speed variables was determined using the 1-sample Kolmogorov-Smirnoff test. All three gait speed variables resulted to be non-normally distributed. Hence, to examine the bivariate associations between UGS, MGS, and GSR and historical and future occurrence of falls and faints, the non-parametric two-sided independent samples Mann-Whitney U-test was utilised.

2.3.4 Overview of Machine Learning Steps – experiment 2

A general overview of the machine learning steps are as follows. The machine learning regression model employed is called Histogram Gradient Boosting Regression. In a stepwise fashion, features are tried out one by one in this model and the best one is selected and added; this step is repeated over and over until the model does not get any better. The final model is trained on the set of features that give the best performance. This final model is then passed through an explainable machine learning step whereby a method from the SHAP package called TreeExplainer is used to observe the relationships between each of the features in the model and the output of the model.

2.3.5 Histogram Gradient Boosting Regression

The regression model employed for this analysis was the histogram gradient boosting regressor (HGBR) from Scikit-learn [40] version 0.24. The Scikit-learn implementation is based on Microsoft's light gradient boosting machines [41].

Introduced by Friedman in 2001, gradient boosting machines [14] have emerged as the one of the most effective and popular machine learning techniques for use with tabular data. These methods have emerged to be at the top of leader boards in terms of performance in modelling tabular data (outperforming or equalling neural networks with reduced complexity[18]) but are also versatile, having been used for time-series and survival analysis [42]. Gradient boosting is a machine learning technique in which models are built sequentially to predict the residuals (actual – predicted values) of the previous model. Usually, a baseline prediction is made using the expected model output (i.e., the mean output value), and then the residuals from that prediction (actual – predicted) become the output values for the next boosting regressor; the residuals from that regressor then become the output values for the following regressor, as so on until either no further improvement in performance metric is achieved or a maximum number of boosts is reached. Gradient-boosted trees (GBT) are a gradient boosting method in which the regressor technique employed is tree-based. Most gradient boosting methods are tree-based. HGBR is based on Light Gradient Boosting Machines (LightGBM) from Microsoft [43]. Trees have benefits in that they do not require normalisation or scaling of input values and are robust to outliers [14]. Simple decision trees have advantages of being interpretable by following the decision path, but this advantage quickly becomes impractical as the size of the tree grows, or ensembles of trees are used in prediction. Single decision tree models are prone to overfitting, but ensembles of trees address this problem by producing many trees each learning different rules about the data to help increase the generalisability of the model. GBT's are ensemble methods. LightGBM and HGBR employ histograms techniques to bin input feature values, which allows for faster computation of tree split points and also provides native support for categorical features and missing data. Essentially, the model learns an association between missing data for a particular feature and the output feature. The support for missing data helps to avoid the need for data imputation or removal of features. The categorical data support avoids the need for dummy variables and one-hot encoding which can drastically increase the dimensionality of the input feature space.

Unlike ordinary linear regression, more advanced machine learning techniques such as gradient boosting have several *hyperparameters* that control aspects of the model's functionality. Some hyperparameters are quite robust and the default settings are usually sufficient, but others need to be tuned to suit the application. The term 'hyperparameter' is used instead of just 'parameter' because in machine learning vocabulary, a 'parameter' refers to an input feature or variable. The histogram gradient boosting hyperparameters that were tuned in this work were: max_iter, loss, random_state, early_stopping, learning_rate, max_leaf_nodes, and min_samples_leaf.

The *max_iter* hyperparameter refers to the maximum number of iterations (i.e., decision trees) that the algorithm will be allowed to complete. The argument to *early_stopping* is either true or false and tells the machine that if there is no improvement in performance above a certain tolerance then the iterations can stop early and save time and energy. There are stochastic processes in the HGBR algorithm and the *random_state* hyperparameter sets the kernel from which pseudo-random numbers are generated; this allows for exact reproducibility of the stochastic processes. The loss function used by the regressor is set using *loss*. A loss function determines how a "best fit" line is drawn through the data. The choices are 'least_squares', which minimises the sum of squared errors between data and fit line; 'absolute_error', which minimises the sum of absolute errors; and 'poisson', which minimises a Poisson distributed error, which is common when modelling count data. Single decision trees are very prone to overfitting which is why ensembles of trees are used instead. When the ensemble is built sequentially however, the danger of overfitting remains and so the output from each tree is scaled down by the *learning_rate* to prevent a single tree from overfitting to the data and ensures that multiple trees contribute to the overall prediction. The "max_leaf_nodes" hyperparameter limits the number of leaf nodes in a given tree. Leaf nodes are nodes that do not undergo any further splitting. Loosely speaking, more leaf nodes can tend towards more overfitting for that tree. Finally, *min_samples_leaf* puts a lower limit to the number of samples that must be considered for a given leaf mode; more samples can help to reduce overfitting, but too many can produce underfitting.

2.3.6 Explainable machine learning

The explainable machine learning method used in this work is call Shapley Additive Explanations (SHAP), which is based on Lloyd Shapley's 1950s work on cooperative game theory where he derived a method of fairly attributing worth to players in a game based on their marginal contributions. When applied to machine learning, the game is the model, and the players are the features in the model. The values quantifying the worth are called Shapley values and are a unique solution to a method that will simultaneously satisfy conditions of fairness.

1. Local accuracy/efficiency

For a given sample, the sum of contributions from each feature plus the mean model output produce the prediction for that sample.

2. Missingness/dummy player

If there is no difference in prediction with or without a given feature, that feature's worth is equal to zero.

3. Symmetry

If two features alter the prediction in the same way, they are both attributed equal worth.

4. Consistency/monotonicity

If a model changes such that the contribution of a feature has a greater impact on the output, the worth of that feature cannot decrease.

The use of SHAP for explainability allows for model interpretations that are built from the local level (the level of each sample) upwards from which global interpretability emerges. SHAP values are computed by the SHAP TreeExplainer package for each feature and sample, i.e., the SHAP values are computed locally. The SHAP package provides graphical methods for observing how a model arrived at a prediction by presenting how the contributions from each feature add up to the predicted value. A SHAP value represents how much of a positive or negative impact (in units of the outcome variable) a feature had for a specific sample. From this, one can derive global metrics of contribution or importance for each feature that reflect the impact of that feature across many samples. The default manner for assessing global feature importance is through mean absolute SHAP values, $\overline{|SHAP|}$.

The nature of SHAP values being true to local impacts of features means that low-frequency, high-impact effects do not go unnoticed. For example, a particular feature might, for most samples, have a low impact; however, for some small subset of samples the feature might have a very large impact. SHAP interaction values are also readily available that explain the impact of interactions between two features. SHAP values are presented as having a positive or negative impact on the output of the model with respect to the expected model output i.e., the mean output of the model. So, for an individual sample, the SHAP value for a particular feature might be for example, -2.5; this should be interpreted as: the value of that feature for that sample is associated with a model output that is -2.5 units less than the model's mean output.

SHAP also has its limitations. SHAP values are not magical representations of reality and much like a regression coefficient need to be understood in the context of the model and the question asked by the model. In the context of tree-based models, it is important to remember that a kind of feature importance has already been implemented during the building of the model. The tree split points are made by seeing which splits along which features give the best gain in performance. This is important when considering collinearities. If given two highly collinear features, the tree will pick one based on either a slight additional gain in performance from one of them, or a stochastic process. In this case, the unselected feature will have no importance in the model whereas in reality the feature is practically the same as the other. In a different scenario, we have two moderately/highly correlated features that are both selected by the tree because despite their correlation, even after chosen the better of the two, the second feature still adds a benefit on top of the first. In this case, SHAP can overcome the collinearity and attribute worth to each feature. Another limitation of SHAP (and LIME) is that it has been shown that mechanisms exist by which dishonest users can fool feature perturbation-based explainers [44].

All SHAP values shown in the results are for the test data.

2.3.6.1 Bootstrapped Confidence Intervals for SHAP values – addition to experiment 2

In order to assess the uncertainty surrounding the SHAP value results, 95% confidence intervals (CI) on the main contribution from each feature were constructed using a bootstrapping method. Bootstrapping is an iterative process whereby sampling *with replacement* is repeatedly performed on a dataset and a statistic of some kind is obtained on each iteration. The distribution of the bootstrapped statistics is then used to obtain an aggregate statistic e.g., perform 1000 iterations of sampling with replacement, on each iteration calculate the mean of the bootstrap sample, and after the 1000 iterations are complete calculate the mean and 95% confidence intervals of the 1000 bootstrap mean values. Sampling with replacement means that after an individual sample has been drawn, it is placed back into the pool so that it is possible to draw it again. In the world of machine learning, producing bootstrapped confidence intervals generally entails a process whereby the model is trained and test using a bootstrap sample on each iteration. The confidence intervals for SHAP values were calculated as the mean and 95% CI of SHAP values at each unique feature value e.g., age=50, age=51, age=52, etc. The CIs were calculated on the SHAP values representing the contribution of a feature without the influence of interactions with other features. In this work, 1000 bootstrap iterations were performed in which the bootstrapped training datasets contain the same number of samples as the original training set.

3 PUBLISHED PAPERS

Section 4 Experiment 1 – was published as a short paper entitled "A Linear Regression-Based Machine Learning Pipeline for the Discovery of Clinically Relevant Correlates of Gait Speed Reserve from Multiple Physiological Systems" [20], published as part of the IEEE's European Signal Processing Conference 2021 (EUSPICO 2021):

J. Davis, S. P. Knight, R. Rizzo, O. A. Donoghue, R. A. Kenny and R. Romero-Ortuno, "A linear regression-based machine learning pipeline for the discovery of clinically relevant correlates of gait speed reserve from multiple physiological systems," 2021 29th European Signal Processing Conference (EUSIPCO), 2021, pp. 1266-1270, doi: 10.23919/EUSIPCO54536.2021.9616187.

Section 5 Experiment 2 – was published as a full paper entitled "Comparison of gait speed reserve, usual gait speed, and maximum gait speed of adults aged 50+ in Ireland using explainable machine learning" [21] published in Frontiers of Network Physiology.

J. R. C. Davis et al., "Comparison of Gait Speed Reserve, Usual Gait Speed, and Maximum Gait Speed of Adults Aged 50+ in Ireland Using Explainable Machine Learning", Frontiers in Network Physiology, Original Research vol. 1, 2021-November-05 2021, doi: 10.3389/fnetp.2021.754477.

In this thesis, a small addition has been made to the content of the latter paper, namely the figure containing histograms of three gait variables.

4 EXPERIMENT 1 – A LINEAR REGRESSION-BASED STEPWISE FEATURE SELECTION FOR THE PREDICTION OF GAIT SPEED RESERVE

4.1 RESULTS

4.1.1 Analytical Sample

Of the 4309 participants who took part in the TILDA wave 3 health centre assessment, 3925 aged 50+ completed both the usual and maximum walking tests required to generate the GSR data (Figure 2). Of those, 2397 (61%) were included in the linear regression analysis as they had no missing values for any of the features. Female sex made up 52.9% of the 2397 participants.

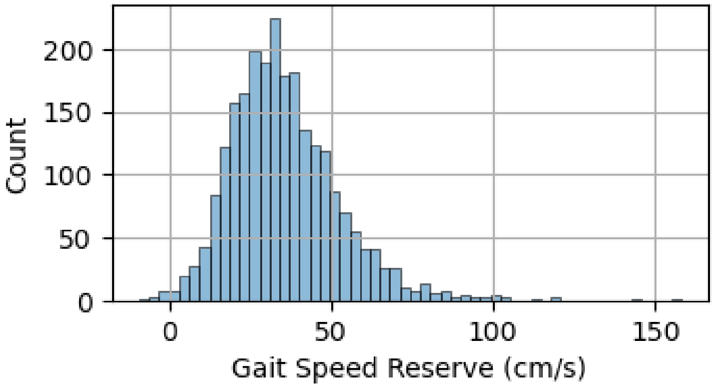


Figure 2. Histogram of gait speed reserve (cm/s) in 2397 TILDA wave 3 participants aged 50 years or over included in experiment 1.

4.2 Selected Features

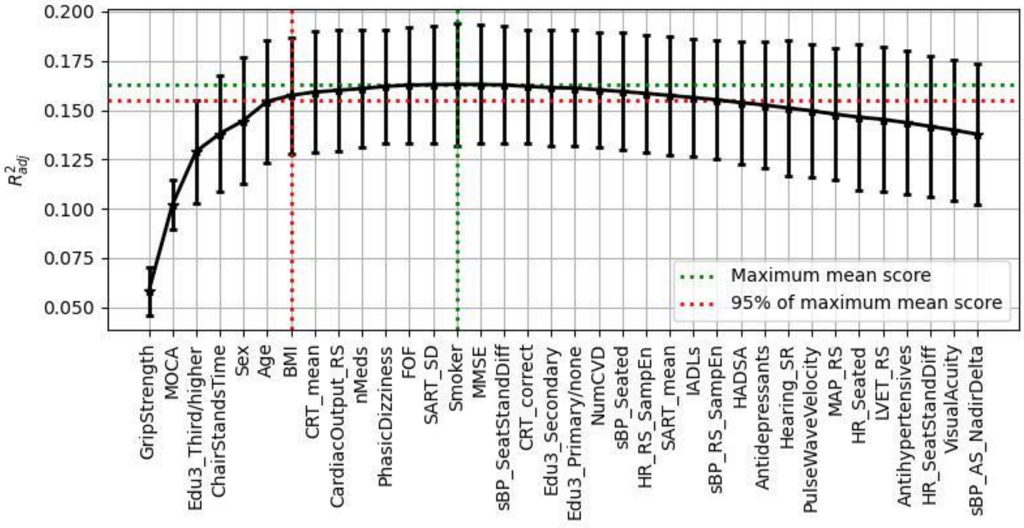


Figure 3. Automated feature selection curve. The x-axis shows the names of the features in the order they are selected for the model. All features to the left of a given feature are included in the model with that feature. The y-axis represents $\boldsymbol{R^2_{adj}}$. Mean and standard deviation (SD) of $\boldsymbol{R^2_{adj}}$ values from the 5-fold CV are shown. The green lines indicate the peak mean score and the feature at which it occurred. The red lines indicate the 95% on the peak mean score and the feature at which it is achieved. (Brief explanations of feature names can be found on page ix).

Figure 3 shows the $\overline{R^2_{adj}} \pm SD$ as features are added to the model. The features appear on the x-axis in a cumulative manner, i.e., the $\overline{R^2_{adj}}$ value at a given x-coordinate corresponds to that of a model containing the feature at that coordinate plus all the other features to its left. The maximum $\overline{R^2_{adj}}$ of 0.16 ± 0.03 was achieved in the feature selection with 14 features. A linear regression model containing these 14 features returned an R^2_{adj} of 0.18 and 0.16 on the training and test data, respectively. However, just the first seven predictors built 95% of the peak $\overline{R^2_{adj}}$ score. In order of addition to the model, these seven features were: grip strength, MOCA score, third level education, chair stands time, sex, age, and BMI.

4.2.1 Model Coefficients

The effect sizes with 95% confidence intervals of the regression coefficients from a model trained using all 14 selected features are shown in Figure 4. The input features are standardised, and the effect size (shown along the x-axis) is in terms of GSR (units of cm/s). The y-axis orders the model features in order of decreasing coefficient magnitude from top to bottom. The 11 statistically significant coefficients (shown in green in Figure 4) are: sex, third level education, MOCA, chair stands time, age, BMI, grip strength, cardiac output at resting state, number of medications, fear of falling, and mean choice reaction time. Female sex, longer chair stands time, older age, higher BMI, higher number of medications, fear of falling, and longer mean cognitive reaction time were associated with a decrease in GSR; whilst third level education, higher MOCA score, greater grip strength, and higher baseline cardiac output were associated with increased GSR.

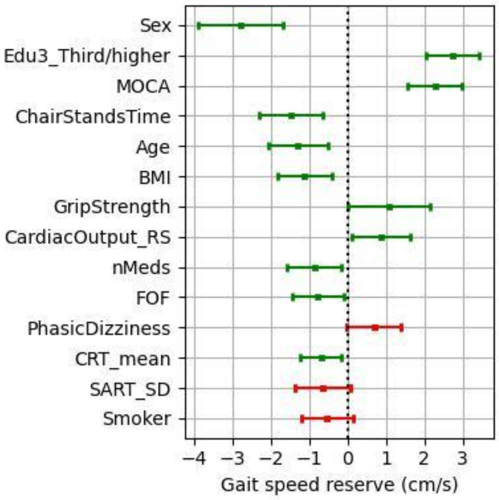


Figure 4. Visual summary of regression coefficients for the standardised input features. The y-axis presents the final model features in order of descending coefficient magnitude from top to bottom. The x-axis shows the coefficients effect size with 95% confidence interval in terms of absolute GSR with units of cm/s. Green markers represent statistically significant (p<0.05) effects. Sex is coded as male = 0 and female =1. (Brief explanations of feature names can be found on page ix).

4.3 Discussion of Experiment 1

A linear regression-based machine learning pipeline was developed for the discovery of clinically relevant predictors of GSR across multiple physiological systems in TILDA wave 3 participants. The first 7 of the 14 selected predictors (grip strength, MOCA score, third level education, chair stands time, sex, age, and BMI) explained 95% of the maximum $\overline{R_{adj}^2}$ achieved (0.16). When examining the regression coefficients, it was found that 11 variables were statistically significant: sex, third level education, MOCA, chair stands time, age, BMI, grip strength, cardiac output at resting state, number of medications, fear of falling, and mean choice reaction time.

Results show that there were significant associations between GSR and features from multiple physiological systems (e.g., cognitive, psychological, musculoskeletal, cardiovascular), which supports that GSR is driven by multiple systems and hence could be useful as an indicator of overall physiological reserve.

Results are consistent with previous MGS research showing significant differences in maximum walking speed for different ages and between men and women [45, 46]. They are also consistent with previous findings that higher BMI and physical workload among those with lower education contributed most to the educational disparities in age-related decline in MGS [47]. Cognitive performance has also been cited as a significant predictor of MGS [10] and this study suggests that the MOCA may be more predictive than the MMSE test in this regard. Recent work has suggested that longer choice reaction time may be associated with longitudinal mobility decline [33]. Furthermore, results underscore that even though MGS can be expected to be reduced in individuals with weaker lower extremity muscle strength [48, 49], upper limb strength assessment also needs to be considered for GSR prediction. Indeed, previous studies have shown that the movement velocity of the upper limbs is a significant determinant of MGS, suggesting that the ability to move any region of the body rapidly might be a critical factor in MGS [50]. The importance of both upper and lower limb muscles as predictors of GSR offers clinical opportunities for strengthening exercises as a way to improve physiological reserve [51]. Fear of falling being predictive of decreased GSR is also clinically plausible, which offers opportunities for psychological interventions to improve self-efficacy [52]. These results also suggest that there is clinical scope for cardiovascular health optimisation in the context of GSR improvement. Indeed, there is evidence in the literature that evaluating MGS in conjunction with UGS is useful for risk stratification in cardiovascular disease patients [53].

A limitation of experiment 1 is that due to the inclusion of participants with only complete data for the 34 initial features, the analytical sample size was reduced to 61% of the original sample. In addition to reducing the sample size it could also bias towards healthier participants as it is typically unhealthier participants with more missing data. In future work, this limitation could be addressed by exclusion of features (where clinically acceptable) with excessive proportion of missing data, or by performing multiple imputation of missing data. In addition, even though GSR had significant predictors in the study, the total variance explained (judged by R^2_{adj}) was low to moderate [54], suggesting that there is further scope for consideration of additional variables. Future feature discovery should attempt to increase the amount of variance explained with additional predictors and/or implementation of the study in external cohorts. However, the level of explained variance is in keeping with previous observations that MGS as a single-item tool is limited to fully predict future falls in community-dwelling older persons [55]. Another limitation to consider is that measurement of GSR restricts analysis to TILDA participants who attended the health assessment centre. Those who did not attend are likely to be frailer than those who did [56].

Given the observed association between GSR and sex, a stratification by sex may have revealed more nuanced differences between the two groups. Furthermore, an analysis by age groups could be considered to explore whether associations with GSR might vary with age. In addition, in future work one could compare current results with those from non-linear and non-parametric machine learning models (as in experiment 2).

According to previous research, an older persons' probability of being frail (by Fried's physical frailty phenotype) with an insufficient GSR could be around 40% [12], which further bolsters the clinical relevance of the present results, since many of the identified associations are potentially modifiable on a prospective basis. Indeed, obesity prevention, cardiovascular risk reduction, cognitive training, appropriate prescribing and monitoring of medication, neuro-psychological interventions against fear of falling, and muscle strengthening could all potentially improve GSR in older populations. GSR could also have safety implications during daily activities that require a sudden increase in pace such as crossing the road, running for the bus, reacting to hazards, etc., and hence be important to maintain older people's functional independence. These results demonstrate the importance of a network physiology approach for the understanding of frailty and resilience in ageing, where systems work together towards the generation of physiological reserve [57].

5 Experiment 2 – Comparison of the predictors of gait speed reserve, usual gait speed, and maximum gait speed using stepwise feature selection and explainable machine learning

Having completed experiment 1 with a linear model at its core, an interest arose to use non-parametric machine learning to explore non-linearities that might be present in some of the relationships between input features and GSR. A further aim was to also model the parent features of GSR: UGS and MGS. The question as to whether the latter two are also multisystem phenomena was of interest as well as a comparison of the domains and features that predicted each gait modality.

Of note regarding the presentation of the results, the method for feature selection describes a situation whereby features can be removed from the model if they are made redundant by the addition of new features; this did not occur in any of the models and as such, all features named henceforth with regard to feature selection are to be understood as features added to the model.

5.1 Results

5.1.1 Analytical Cohort

In TILDA wave 3, 4309 participants completed the health centre assessment [24], where the gait speed tests were conducted. After exclusion of participants aged less than 50 years or with missing data for either UGS or MGS, there were 3925 participants, with 2156 (55%) being female. A flowchart of the included sample can be seen in Figure 5.

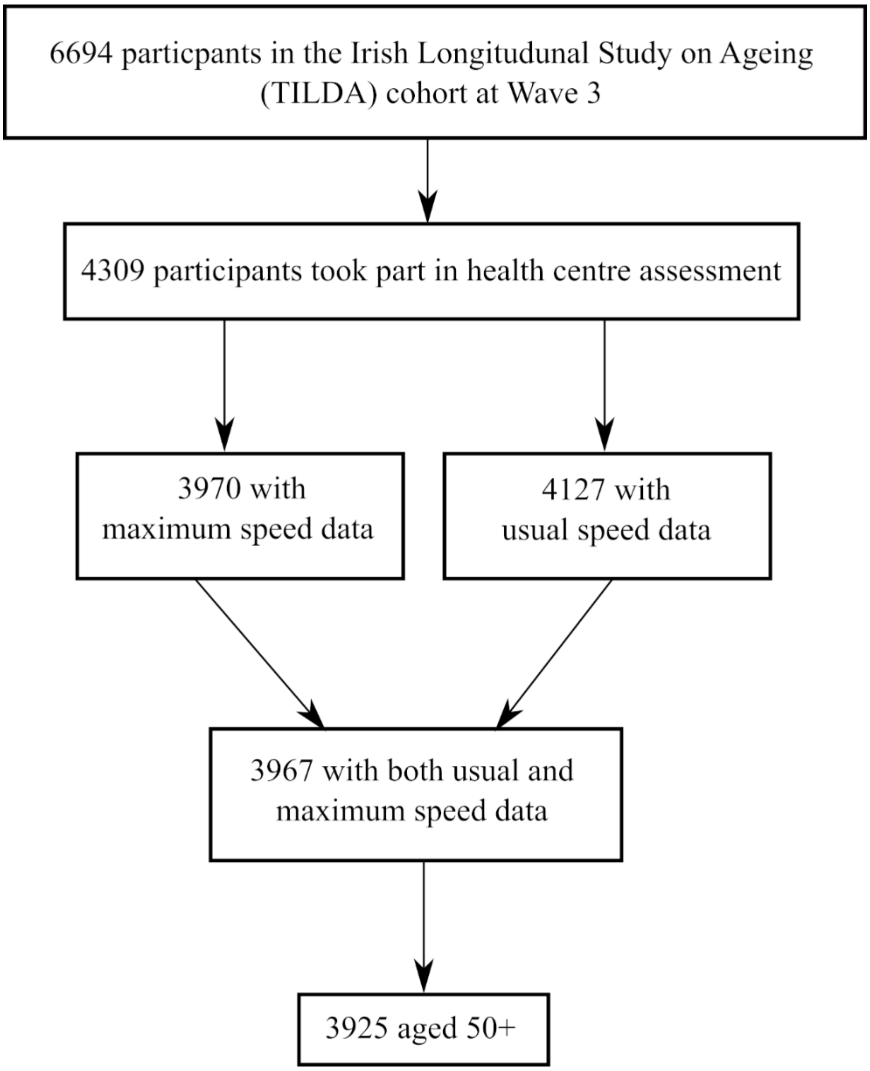

6694 particpants in the Irish Longitudunal Study on Ageing (TILDA) cohort at Wave 3
4309 participants took part in health centre assessment
3970 with maximum speed data
4127 with usual speed data
3967 with both usual and maximum speed data
3925 aged 50+

Figure 5. Analytical sample inclusion flowchart for experiment 2.

The educational attainment breakdown was as follows: third/higher: 1685 (43%), secondary: 1571 (40%), and primary/none: 669 (17%). The analytical cohort had a mean (SD) age of 64.5 (7.8) years, UGS of 136.7 (19.2) cm/s, MGS of 171.0 (26.9) cm/s, and GSR of 34.3 (16.6) cm/s. The histograms of these measures are shown in Figure 6.

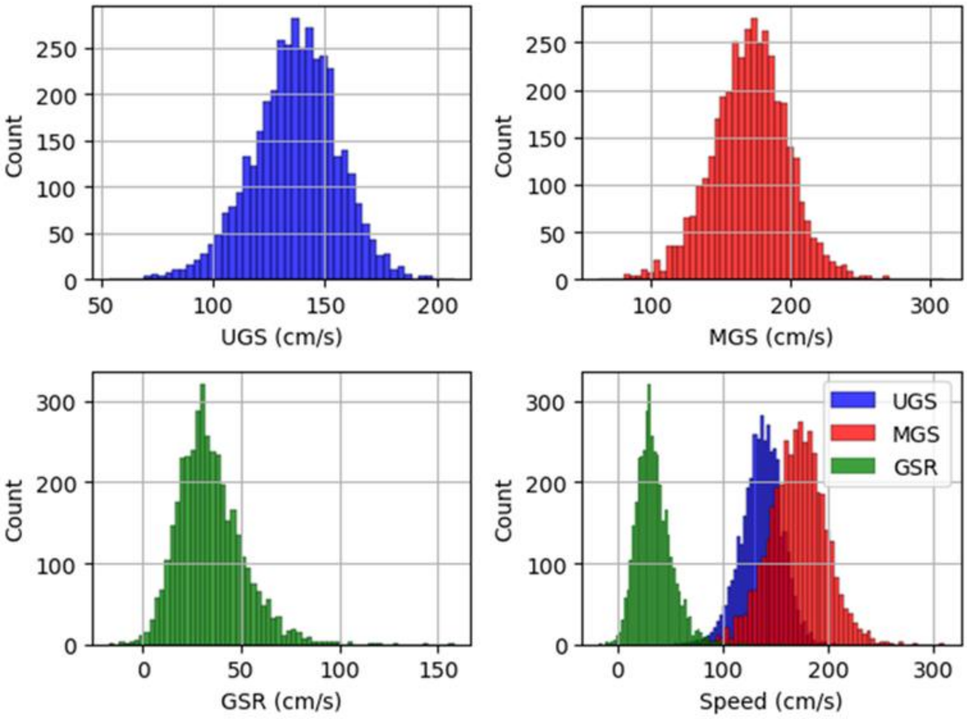


Figure 6 . Histograms of usual gait speed (UGS), maximum gait speed (MGS), and gait speed reserve (GSR) in 3925 participants included in experiment 2.

5.1.2 Group differences in faller and fainter status

21.3% of wave 3 participants were historical fallers, 3.8% historical fainters, 31.9% future fallers, and 5.4% future fainters. Table 6 shows the results of the association between UGS, MGS, GSR, and these clinical variables. Differences between historical fallers were all statistically significant, with a largest median difference for MGS. Statistical significance of $p<0.05$ was demonstrated in historical fainters for UGS and MGS only, with the largest difference also for MGS. A similar pattern emerged for future fallers and fainters. A previous study that found that combining UGS and MGS to calculate an individual's GSR did not provide additional insight into fall status [58]. However, results herein suggest that GSR was also useful to capture falls in this TILDA wave 3 sample.

Table 6. Group statistics and results of independent samples Mann-Whitney U-test for historical and future falls and faints occurrence. (IQR: interquartile range)

Historical falls and faints

	Non-fallers (median (IQR))	Fallers (median (IQR))	Difference in group median	Mann-Whitney p-value	Non-fainters (median (IQR))	Fainters (median (IQR))	Difference in group median	Mann-Whitney p-value
UGS (cm/s)	139.3 (24)	133.1 (26)	6.2	<0.001	138.3 (25)	133.0 (19)	5.3	0.005
MGS (cm/s)	174.5 (33)	166.2 (36)	8.3	<0.001	172.55 (34)	165.95 (34)	6.6	0.010
GSR (cm/s)	33.0 (20)	30.4 (22)	2.6	<0.001	32.3 (20)	28.95 (23)	3.35	0.118
Future falls and faints								
	Non-fallers (median (IQR))	Fallers (median (IQR))	Difference in group median	Mann-Whitney p-value	Non-fainters (median (IQR))	Fainters (median (IQR))	Difference in group median	Mann-Whitney p-value
UGS (cm/s)	139.7 (23)	134.8 (27)	4.95	<0.001	138.4 (24)	134.1 (30)	3.7	<0.001
MGS (cm/s)	175.1 (33)	168.0 (36)	7.05	<0.001	173.2 (33)	165.7 (39)	6.0	<0.001
GSR (cm/s)	33.20 (21)	31.4 (21)	1.85	<0.001	32.7 (21)	31.1 (21)	1.6	0.209

5.1.3 Usual Gait Speed

The peak $\overline{R^2_{adj}}(SD)$ achieved for the UGS model was 0.38 (0.04), with training and test scores of 0.43 and 0.41, respectively. The expected model output was 136.6 cm/s. The features chosen for the model, in order of selection as per Figure 7 were age, chair stands time, BMI, grip strength, number of medications, resting state pulse interval, mean motor reaction time, height, depression score, sit-to-stand difference in diastolic blood pressure, and left visual acuity.

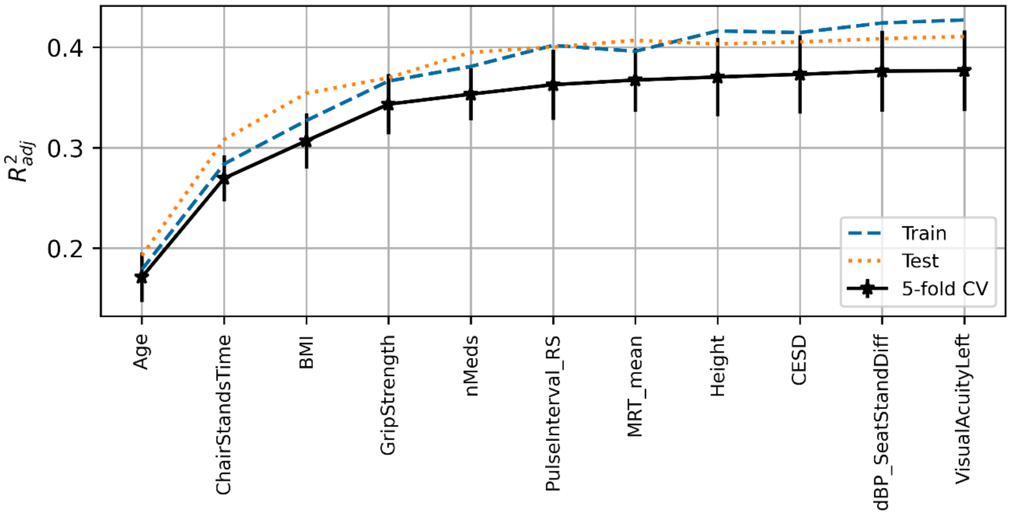


Figure 7. Visualisation of the feature selection process for the usual gait speed model. From left to right on the x-axis, the features are in order of addition to the model. The y-axis shows the dimensionless $\boldsymbol{R^2_{adj}}$ *metric. Mean 5-fold cross-validation scores with error bars showing ± SD are shown in black, train scores in dashed blue, and test scores in dotted red. (Brief explanations of feature names can be found on page ix).*

A SHAP summary plot is shown in Figure 8; each point on the x-coordinate represents a samples SHAP value, and its colour signifies the value of the feature for that sample, with light brown being high, black being low, and nan (missing) values appearing grey. On the y-axis, features are arranged from top to bottom in order of decreasing mean absolute SHAP value: chair stands time, age, body mass index, number of medications, grip strength, resting state pulse interval, height, mean motor reaction time, CESD depressive symptoms score, difference in seated and standing diastolic blood pressure, and visual acuity in the left eye. The figure suggests that upper limits (light brown) of certain variables (e.g., chair stands time, age, body mass index, number of medications) are more negatively impactful than their lower limits, which are positively impactful. The opposite is the case for upper limits of grip strength, for example.

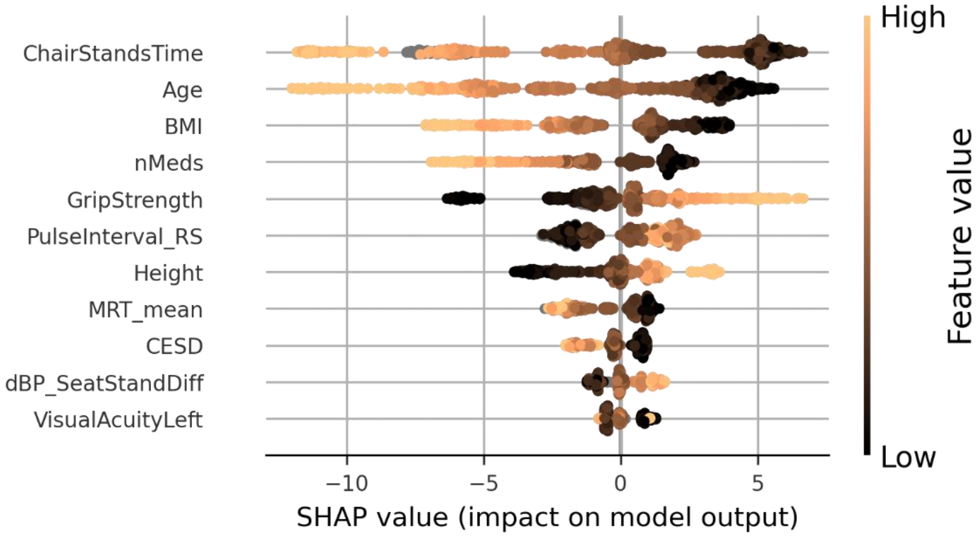


Figure 8. SHAP summary plot for the final usual gait speed model. Features are ordered from top to bottom by decreasing mean absolute SHAP value. For each feature, each point represents a single sample in the test data. A sample's x-coordinate displays the SHAP value for that sample with respect to a given feature. The colour of a sample indicates the value of the feature, with light brown being high, black low, and grey missing. (Brief explanations of feature names can be found on page ix).

Scatter plots of SHAP value vs. feature can be seen for all features in Figure 9. SHAP values (left y-axis) vs. input feature value (x- axis) with underlaid histogram (right y-axis shows histogram counts) are shown for each feature in the UGS model. Features are arranged top-to-bottom and left-to-right in order of decreasing mean absolute SHAP value. At the zero point on the left y-axis (SHAP value = 0), the corresponding x-coordinate values for that feature are associated with having no impact on the model (i.e., they are associated with the mean model output). The vertical spread observed in the SHAP values vs. input feature plots indicates the presence of interaction effects. Although not chosen for the model the data points are coloured by sex.

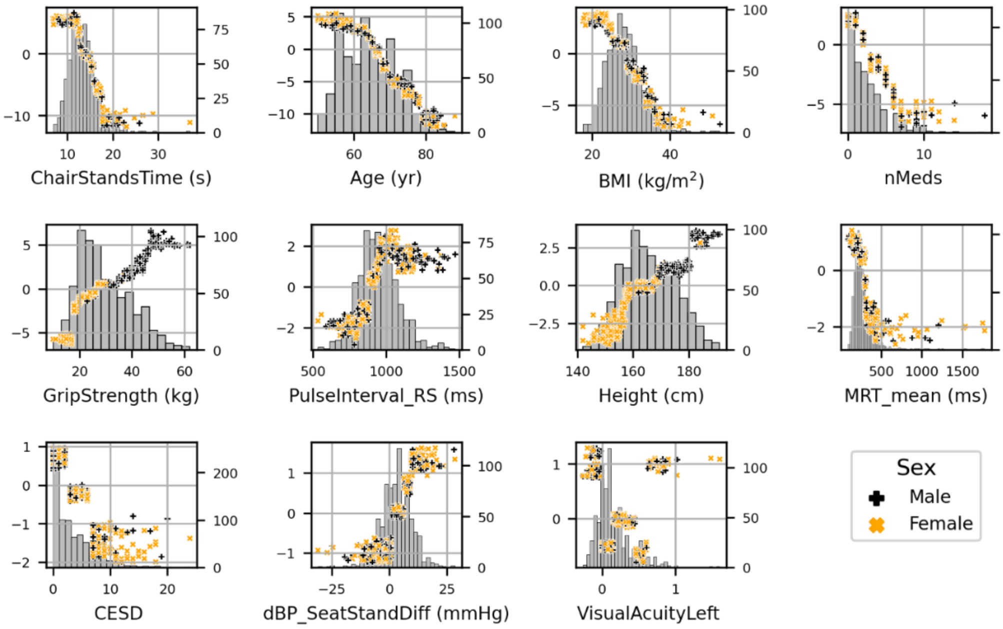


Figure 9. SHAP values (left y-axis) vs. input feature value (x- axis) with underlaid histogram (right y-axis showing histogram counts) for each feature in the usual gait speed model. Features are arranged top to bottom and left to right in order of decreasing mean absolute SHAP value. Points are coloured by sex: male is black '+' and female is orange 'x'. (Brief explanations of feature names can be found on page ix).

To further investigate the interaction effects suggested by vertical spreading in Figure 9, a plot (Figure 10) of features ordered by decreasing mean absolute SHAP interaction value was produced; in it, features are ranked from left to right in order of decreasing mean absolute SHAP interaction values (orange dotted line). Also shown in dashed blue are the mean maximum absolute SHAP interaction values, which can highlight the effects of outliers.

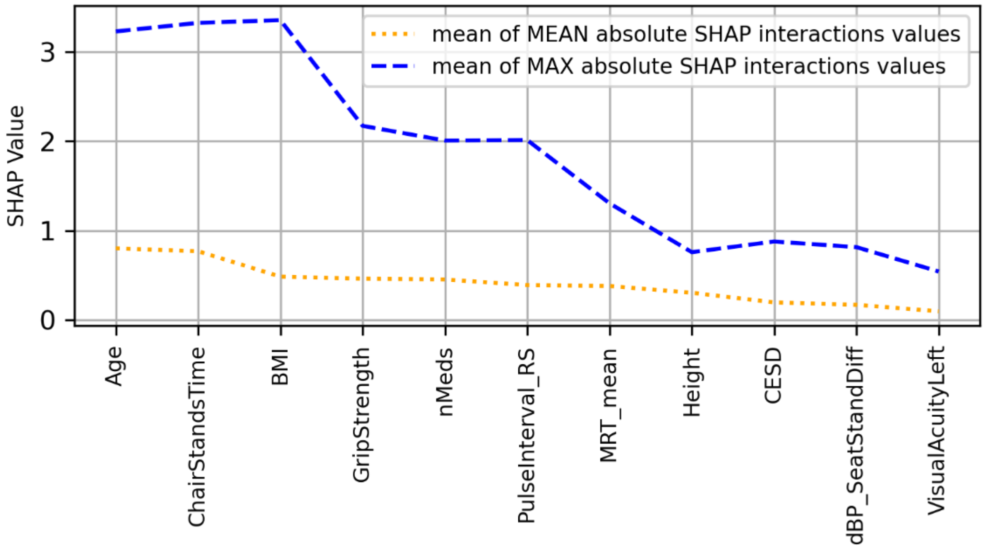


Figure 10. Features ranked from left to right in order of decreasing mean absolute SHAP interaction values (orange dotted line) for the usual gait speed model. Also shown by blue dashed line are the mean maximum absolute SHAP interaction values, which can highlight the effects of outliers. (Brief explanations of feature names can be found on page ix).

The scatter plots of the top four interaction effects in the model (i.e., age, chair stands time, body mass index, and grip strength) are shown in Appendix A. In the scatter plots, the points are coloured according to the value of the main interaction feature. The interactions are computed for the features in whatever numerical form they exist in, but for ease of visualisation, continuous features are coloured according to what quartile a particular samples value falls in; black indicates the value is in the lowest quartile and light brown the highest quartile. In each figure, the subplots are ordered from top-left to bottom-right by decreasing mean absolute SHAP interaction value.

5.1.4 Maximum Gait Speed

The peak $\overline{R_{adj}^2}(SD)$ achieved for the MGS model was 0.45 (0.04), with training and test scores of 0.54 and 0.46, respectively. The expected model output was 170.9 cm/s.

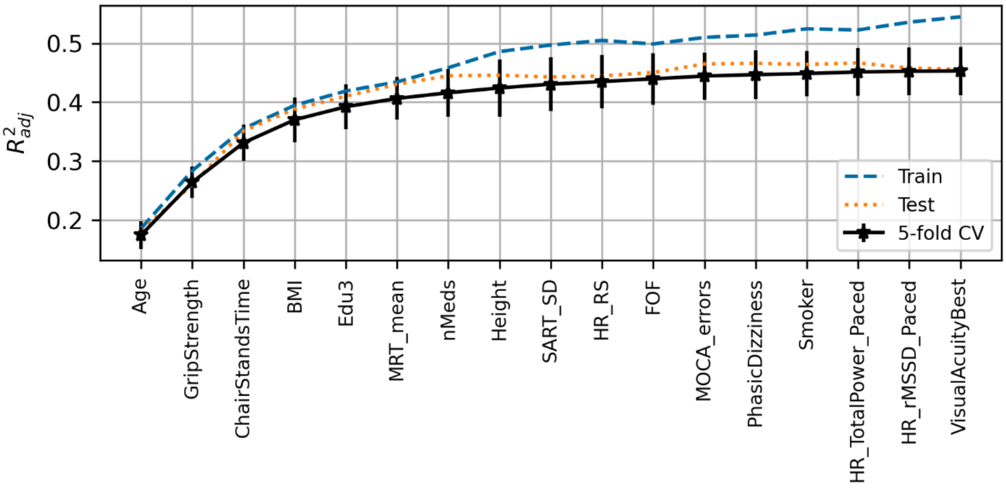


Figure 11. Visualisation of the feature selection process for the maximum gait speed model. From left to right on the x-axis, the features are in order of addition to the model. The y-axis shows the dimensionless $\boldsymbol{R_{adj}^2}$ *metric. Mean 5-fold cross-validation scores with error bars showing ± SD are shown in black, train scores in dashed blue, and test scores in dotted red. (Brief explanations of feature names can be found on page ix).*

Features chosen for the model, in order of selection were: age, grip strength, chair stands time, body mass index, education, mean motor reaction time in the choice reaction time test, number of medications, height, the standard deviation of the mean reaction time in the sustained attention to response task, resting state heart rate, fear of falling, MOCA errors, orthostatic intolerance during active stand, smoking status, total power of the heart rate during paced breathing, the root mean square of successive differences between heartbeats during paced breathing, and best visual acuity. Figure 11 shows the visualisation of the feature selection process for this model.

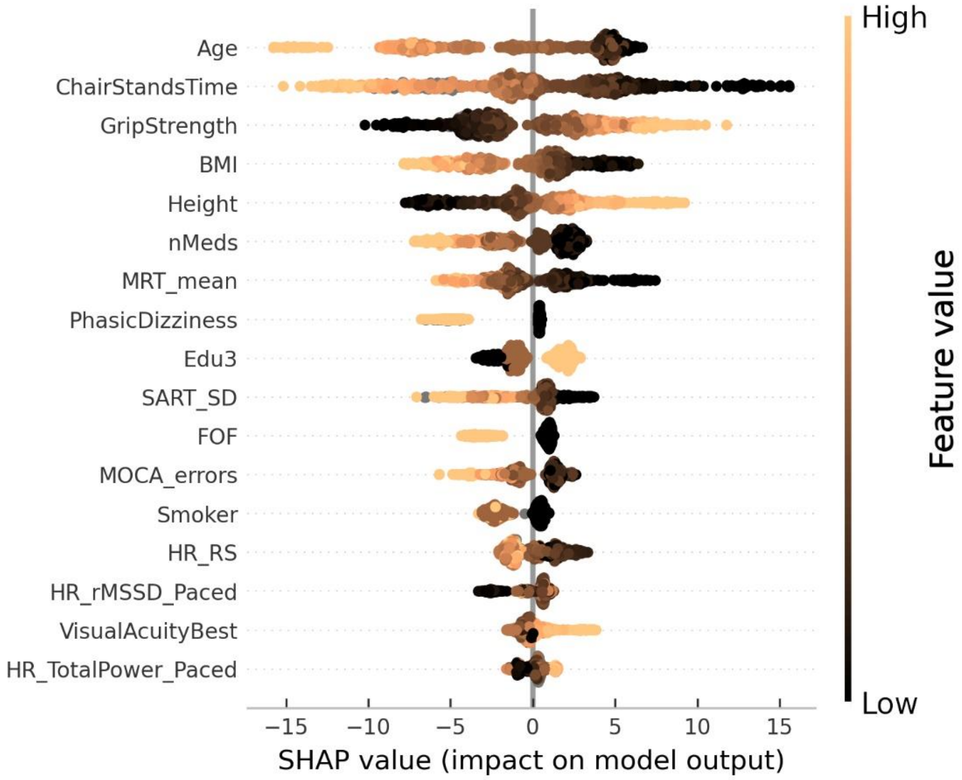


Figure 12. SHAP summary plot for the maximum gait speed model. Features are ordered from top to bottom by decreasing mean absolute SHAP value. For each feature, each point represents a single sample in the test data. A sample's x-coordinate displays the SHAP value for that sample with respect to the given feature. The colour of a sample indicates the value of the feature, with light brown being high, black low, and grey missing. (Brief explanations of feature names can be found on page ix).

In the SHAP summary plot for the MGS model shown in Figure 12, the feature importance ranked in order of decreasing mean absolute SHAP values was: age, chair stands time, grip strength, body mass index, height, number of medications, mean motor reaction time in the choice reaction time test, orthostatic intolerance during active stand, education, the standard deviation of the mean reaction time in the sustained attention to response task, fear of falling, MOCA errors, smoking, mean heart rate pre-active stand, the root mean square of successive differences between heartbeats during paced breathing, visual acuity, and total power of the heart rate during paced breathing.

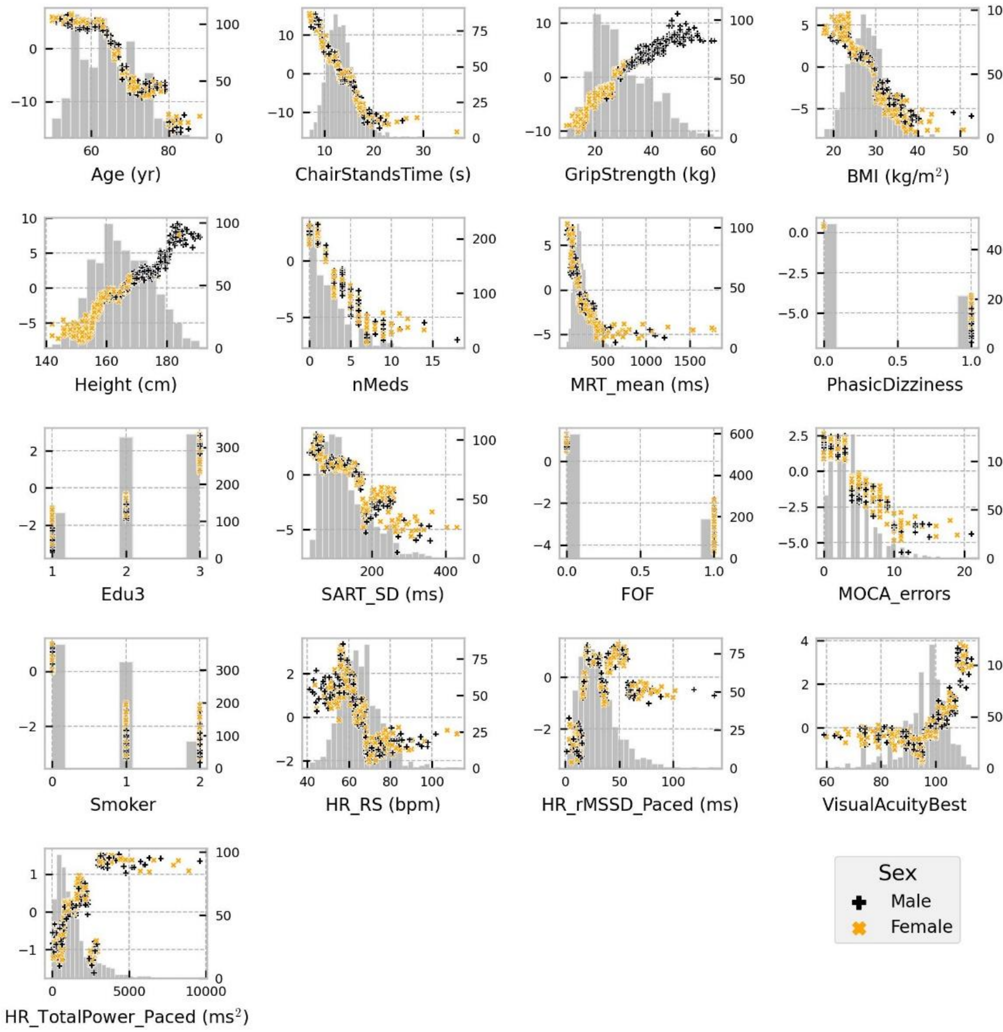


Figure 13. SHAP values (left y-axis) vs input feature value (x- axis) with underlaid histogram (right y-axis shows histogram counts) for each feature in the maximum gait speed model. Features are arranged top to bottom and left to right in order of decreasing mean absolute SHAP value. Points are coloured by sex: male is black '+' and female is orange 'x'. (Brief explanations of feature names can be found on page ix).

Figure 13 shows the SHAP values versus input feature values with underlaid histogram for each feature in the MGS model. Figure *14* shows a plot of features ordered by decreasing mean absolute SHAP interaction value, and Appendix B contains the scatter plots of the top four interaction effects in the model.

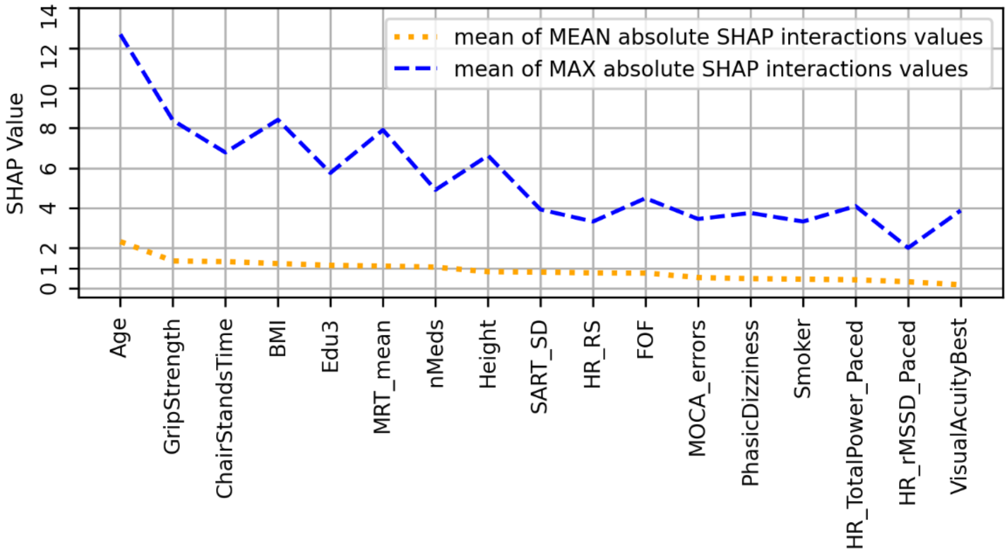


Figure 14. Features ranked from left to right in order of decreasing mean absolute SHAP interaction values (orange dotted line) for the maximum gait speed model. Also shown by blue dashed line are the mean maximum absolute SHAP interaction values, which can highlight the effects of outliers. (Brief explanations of feature names can be found on page ix).

5.1.5 Gait Speed Reserve

The peak $\overline{R^2_{adj}}(SD)$ achieved for the GSR model was 0.19 (0.02), with training and test scores of 0.22 and 0.21, respectively. The model expected output was 34.2 cm/s.

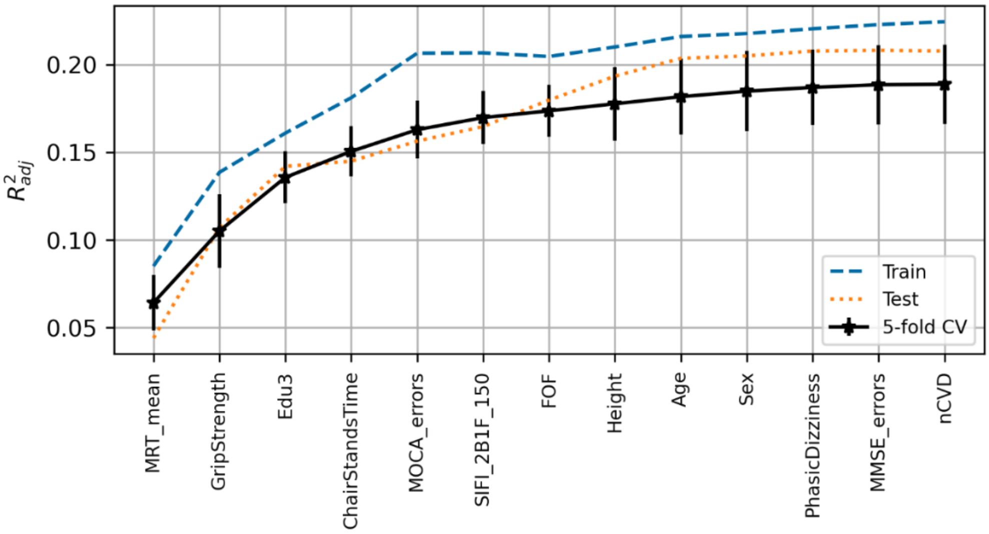


Figure 15. Visualisation of feature selection process for gait speed reserve. From left to right on the x-axis, the features are in order of addition to the model. The y-axis shows the dimensionless $\boldsymbol{R^2_{adj}}$ metric. Mean 5-fold cross-validation scores with error bars showing ± SD are shown in black, train scores in dashed blue, and test scores in dotted red. (Brief explanations of feature names can be found on page ix).

Figure 15 shows the visualisation of the feature selection process. In order of selection, the features chosen were mean motor reaction time in the choice reaction time test, grip strength, education, chair stands time, MOCA errors, accuracy proportion in the sound induced flash illusion (two beeps and one flash with stimulus-onset asynchrony of +150 ms), fear of falling, height, age, sex (0 = male; 1 = female), orthostatic intolerance in the active stand test, MMSE errors, and number of cardiovascular conditions.

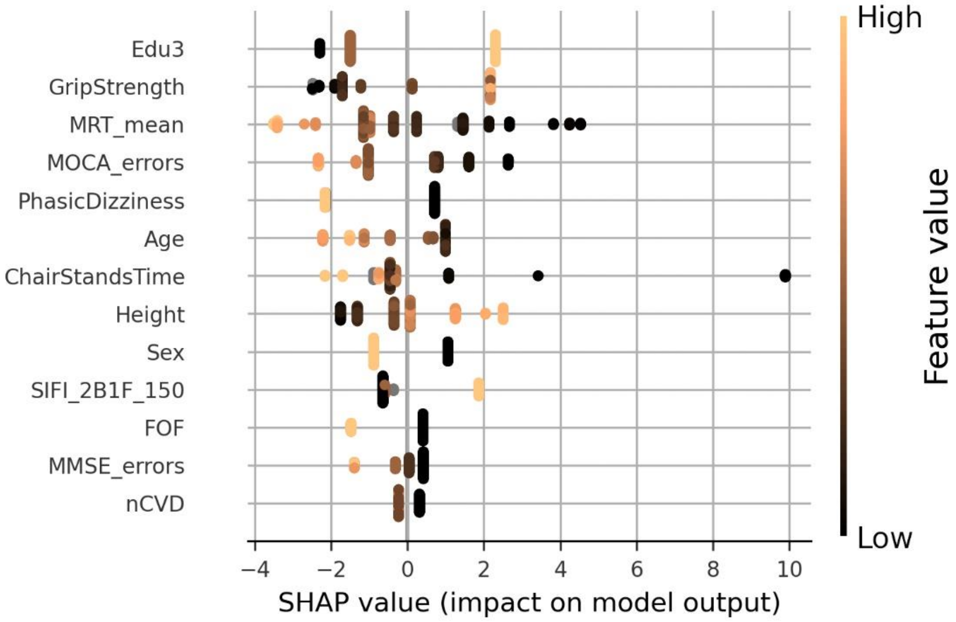


Figure 16. SHAP summary plot for the final gait speed reserve model. Features are ordered from top to bottom by decreasing mean absolute SHAP value. For each feature, each point represents a single sample in the test data. A sample's x-coordinate displays the SHAP value for that sample with respect to the given feature. The colour of a sample indicates the value of the feature, with light brown being high, black low, and grey missing. (Brief explanations of feature names can be found on page ix).

In the SHAP summary plot for the GSR model shown in Figure 16, the feature importance ranked in order of decreasing mean absolute SHAP values was level of educational attainment, grip strength, mean MRT, MOCA errors, age, chair stands time, height, sex, accuracy proportion in the sound induced flash illusion, fear of falling, orthostatic intolerance, MMSE errors, and number of cardiovascular conditions.

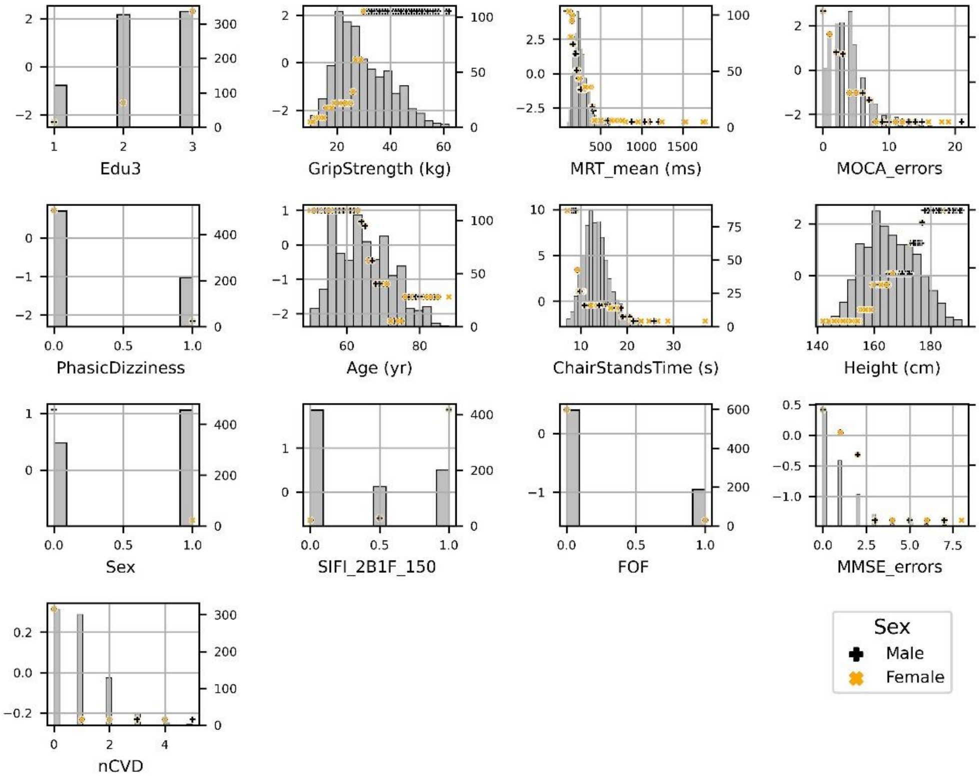


Figure 17. SHAP values (left y-axis) vs input feature value (x- axis) with underlayed histogram (right y-axis shows histogram counts) for each feature in the gait speed reserve model. Points are coloured by sex: male is black '+' and female is orange 'x'. (Brief explanations of feature names can be found on page ix).

Figure 17 shows the SHAP values versus input feature values with underlaid histogram for each feature in the GSR model. The absence of vertical spread in the SHAP vs. feature scatter plots is due to the maximum leaf nodes hyperparameter being set equal to two for the histogram gradient boosting model. This results in there being no interaction terms since the predictions made by each tree only considered features independently (i.e., a maximum leaf node limit of two means that for a given tree only a single split is made along a single feature).

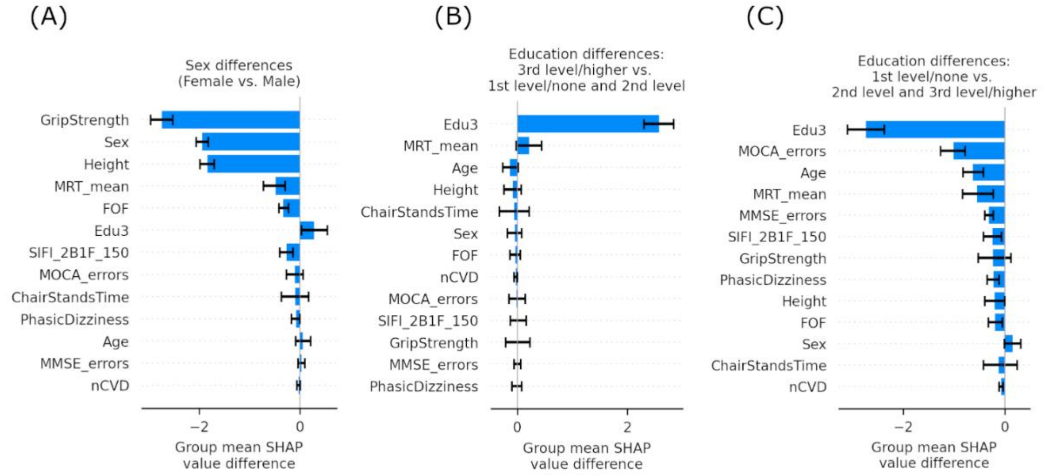


Figure 18. Bar graphs showing the group mean differences in SHAP values between subgroups with 95% confidence intervals for each feature in the gait speed reserve model. (A) shows the differences in sex, (B) shows the differences between participants with third/higher level of educational attainment and all others, and (C) shows the differences between participants with first level/no education and all others. (Brief explanations of feature names can be found on page ix).

The group mean differences in SHAP values for each feature along with 95% confidence intervals can be seen in Figure 18 for: (A) sex, (B) third level education vs. all others, and (C) first/no education vs. all others. For sex, the grip strength feature produced a larger difference in means than sex itself with grip strength having a less positive impact for women. Height, mean MRT, fear of falling, and SIFI accuracy, were all significant and all exhibited a negative mean impact difference. On the other hand, for education there was a positive group mean difference for women in comparison to men. When comparing third/higher educational attainment to the rest, education itself seemed to make the only significant difference. However, when comparing primary/no educational attainment to secondary and tertiary educational attainment in Figure 18 (C), several other significant differences other than education were observed: MOCA errors, age, mean MRT, MMSE errors, illusion accuracy, orthostatic intolerance, fear of falling, and number of cardiovascular diseases.

5.1.6 Summary of experiment 2 results

The distribution of features between the three models is displayed in Figure 19 in the form of a venn diagram. The top left set in blue contains the UGS features, the top right set in red contains the MGS features, and the bottom centre set in green contains the GSR feautres. Firstly, there are features unique to each model: PulseInterval_RS, CESD, dBP_SeatStandDiff, VisualAcuityLeft for the UGS model; SART_SD, HR_RS, Smoker, VisualAcuity, and HR_rMSSD_Paced, HR_TotalPower_Paced for MGS; and Shams_2B1F_150, Sex, MMSE_errors, and NumCVD for GSR. In the centre, common to all three models are Age, GripStrength, ChairStandsTime, MRT_mean, and Height. Features common only to UGS and MGS are BMI and Meds, and those common only to MGS and GSR are Edu3, MOCA_errors, FOF, and PhasicDizziness. There are no features that are common only to UGS and GSR.

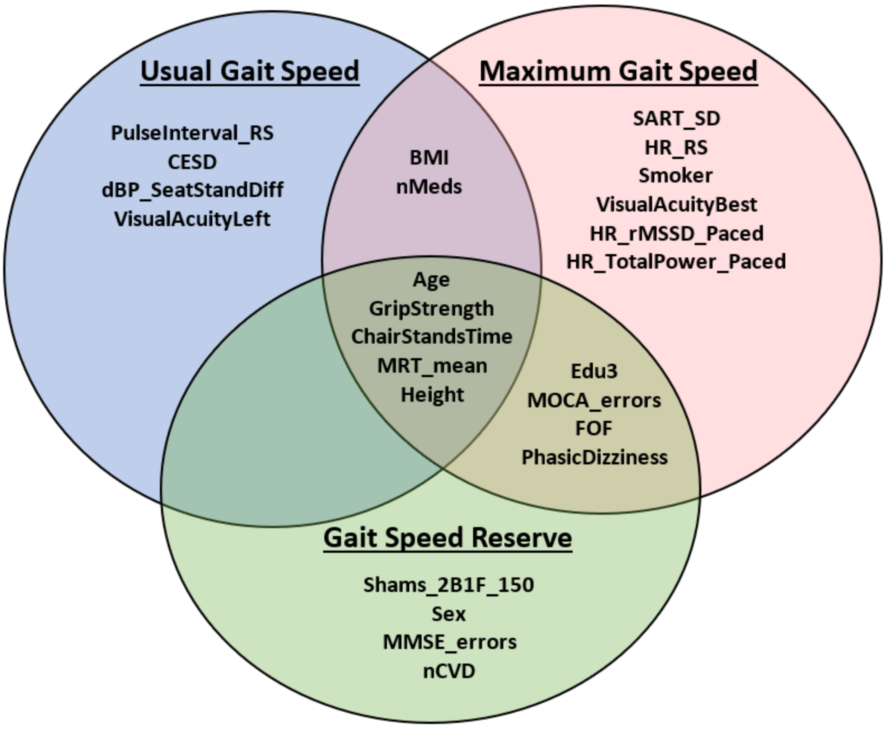


Figure 19. Venn diagram of the features selected across the three models: UGS (usual gait speed), MGS (maximum gait speed), GSR (gait speed reserve). (Brief explanations of feature names can be found on page ix).

5.2 Discussion of Experiment 2

5.2.1 Overall summary of findings

In experiment 2, using data from wave 3 of TILDA, a gradient boosting trees-based stepwise feature selection pipeline was employed for the discovery of clinically relevant predictors of UGS, MGS, and GSR using a shortlist of 88 features across 5 domains. The features selected for the respective models explained MGS and UGS to a greater extent than GSR. As shown in Figure 19 there were common features, but also some features unique to each of the three models.

5.2.2 Model Prediction

Based on model R_{adj}^2 values, GSR (19%) was less predictable than MGS (45%) and UGS (38%). While not aware of previous published data for comparison with GSR prediction, a previous study by Bohannon reported linear regression R^2 values of 13% for UGS and 41% for MGS [48]. Experiment 2 results agree in that the MGS model yielded a larger prediction score than that of UGS.

5.2.3 Common Features

Across the three models, there were five common selected features: age, grip strength, chair stands time, mean motor reaction time in the choice reaction time test, and height. The top four features with the most impactful interactions (by mean absolute SHAP interaction value) were the same for the UGS and MGS models: age, chair stands time, grip strength, and BMI.

Our results agree with Bohannon's previous findings that UGS and MGS decline with increasing age [48]. Other authors have also shown similar findings for UGS [7, 59, 60]. As per SHAP value vs. feature plots, increasing age was negatively associated with UGS, MGS, and GSR at ≥68, ≥68, and ≥66 years, respectively. Height is also unsurprising as a common predictor; indeed, taller people have longer legs and can achieve longer strides and higher velocity in any gait modality. Consequently, gait speed is often normalized by height [48, 61, 62].

It is also clinically plausible that higher grip strength (as a marker of upper limb strength) and shorter chair stands time (more representative of lower limb endurance) were common determinants of all three performance metrics. Indeed, sarcopenia (low muscle mass and/or strength), of which both grip strength and the five chair stands test are indicative measures [63], has been associated with reduced gait speed and poor functional outcomes in older people [64-66]. In experiment 2 models, slower chair stands time was associated with a decline in UGS, MGS, and GSR once time increased beyond 14.2 s, 13 s, and 10.6 s, respectively; whilst increases in UGS, MGS, and GSR began at values of 13.4 s, 13 s, and 10.6 s, respectively. Grip strength of ≤26 kg was associated with slower UGS, MGS and GSR while grip strengths of ≥35 kg, 27 kg, and 27 kg, respectively, were associated with faster performance. These values for grip strength, while interesting from an absolute point of view, have a reduced clinical significance given the large differences in grip strength between men and women. Except for height, the other features relationships to the model output appear quite homogeneous with respect to sex.

Higher mean motor reaction time in the choice reaction time test was associated with lower speed in all three models. In previous research, shorter CRT has been associated with faster gait speed after adjusting for potential confounders and suggests that in older adults, engaging more frequently in cognitively stimulating activities may improve neuromotor performance and mobility [67]. In addition, experiment 2 results resonate with previous TILDA work utilising traditional linear statistics showing that participants in the slower MRT group (<250 ms) at wave 1 seemed to have faster mobility decline as assessed by the timed up and go at wave 3, approximately four years later [33]. Of note, in the latter study, the MRT cutoff was set arbitrarily, but in the present study the negative/positive impact thresholds for UGS, MGS and GSR were 299 ms, 231 ms, and 229 ms, respectively. The less physically demanding UGS model was only negatively influenced above a relatively slower MRT threshold.

The counter-intuitive results of higher grip strength, quicker chair stands time, and quicker MRT being associated with an increase in UGS when compared to MGS may be revealing of the underlying determining mechanisms of both acts; MGS may be a more physically determined act than UGS and easier to improve on than UGS.

Common between UGS and MGS models were BMI and number of medications, in the clinically expected directions, i.e., obesity and number of medications had a negative impact on gait speed. As regards obesity, research has suggested that obese adults may select their walking speed to minimise pendular energy transduction, energy cost, and perceived exertion during walking [68]. In the UGS and MGS models, a BMI ≥29 kg/m^2 had negative impact association. Hypothetically, it is possible that in TILDA, obese individuals equally reduced their UGS and MGS, which could possibly explain why BMI was not a feature in the GSR model. As regards number of medications, a similar mechanism could apply. In any case, findings are in keeping with previous research showing that drug interactions may increase the likelihood of gait speed decline among older adults [69]. In the UGS and MGS models, more than two medications had a negative impact association. This is below the usual polypharmacy definition of 5+ regular medications and the negative impact association with medications could be related to the underlying health conditions rather than due to the medications themselves. Of note, visual acuity featured in both UGS (left) and MGS (best), but not in GSR, which could have a similar underlying reason (i.e., both UGS and MGS equally limited).

There were no features exclusively shared by UGS and GSR, but there were four features in the intersection of MGS and GSR: education, MOCA errors, fear of falling, and orthostatic intolerance. As regards the former two, tertiary education was associated with increased gait speed, and primary and secondary levels with a decrease. Greater than three MOCA errors negatively impacted both models. Interestingly, better MOCA performance is associated with higher education [70] and places greater emphasis on frontal executive function and attention tasks than the MMSE [71]. Planning for the MGS task may require greater attention and executive function than performing the UGS task [10], and this may explain MOCA being related to GSR and MGS. Two or more MMSE errors were associated with GSR decrease.

Analogously, orthostatic intolerance and fear of falling may selectively limit the more demanding MGS task, but not the more comfortable UGS task. Orthostatic intolerance can be caused by orthostatic hypotension, which in some studies has been associated with reduced gait speed [72]. In addition, orthostatic intolerance can be a feature of vestibular disorders such as benign paroxysmal positional vertigo (BPPV) [73], and research has suggested that the gait characteristics of BPPV can be attributed to an inadequate, cautious gait control [74], which may preferentially manifest in the MGS task. Fear of falling can also become stronger when facing the MGS task, compared to walking at UGS [75].

5.2.4 Unique features

Features exclusive to UGS were depression, diastolic blood pressure change from sitting to standing, and resting state pulse interval. Higher levels of depressive symptoms have been associated with worse performance in specific quantitative gait variables in community-residing older adults, including lower velocity [76]. In the model, CESD negatively impacted UGS when CESD>2 points.

Similarly, TILDA work showed that slower recovery of BP after standing (systolic and/or diastolic) was independently associated with poorer gait performance [72]. On the other hand, a higher pulse interval indicates a higher heart rate variability and a more parasympathetic-driven autonomic cardiac control, which has been associated with healthier states [77] and mirrors the fact that for the UGS model, higher pulse intervals had positive influence. In the model, a baseline pulse interval of 799 ms or less had a negative impact on UGS (this is roughly 75.1 bpm: 60 seconds per minute / 0.799 seconds per beat).

Exclusive to the MGS model were the standard deviation of the mean reaction time in the sustained attention to response task, smoking, the mean heart rate pre-active stand, the total power of the heart rate during paced breathing, and the root mean square of successive differences between heartbeats during paced breathing. In a previous study, community-dwelling participants who displayed poorer sustained attention walked more slowly during both single and dual gait tasks [78]. In the model, standard deviation of the mean SART reaction time <157.7 ms was associated with slower MGS. Research has shown that in habitual smokers, smoking acutely reduces baseline levels of vagal-cardiac nerve activity and completely resets vagally mediated arterial baroreceptor-cardiac reflex responses [79], which could be in keeping with heart rate and heart rate variability features being selected in this model. A baseline heart rate of 67.9 bpm or more had a negative impact on MGS in the model. Comparing this to the pulse interval of 799 ms (equivalent to 75.1 bpm) associated with the beginning of negative impact association in the UGS model, we see that in terms of an increasing heart rate MGS begins to decline earlier than UGS.

Finally, features exclusive to GSR were accuracy proportion in the sound induced flash illusion (two beeps and one flash with stimulus-onset asynchrony of +150 ms), sex, MMSE errors, and number of cardiovascular diseases. Male sex was associated with increased GSR. Alternatively, this may also be because the variance explained by the GSR model was relatively low and the effect of sex might disappear when additional features are selected as in other models. One or more cardiovascular diseases was negatively associated with GSR, which is in keeping with the possibility that this type of disease may limit MGS more than UGS. As noted by a previous study [49], the difference between UGS and MGS is predominantly dictated by the latter. A notable exclusive associate of GSR was the proportion of accuracy in the sound induced flash illusion. This can be interpreted in the context that worse visual–somatosensory integration is associated with worse balance in older people [80], and that an increase in susceptibility to the sound-induced flash illusion during standing relative to sitting was present in older adults prone to falling [81].

5.2.5 Strengths of the Methodology and Study

A main strength of the methodology is the use of the Histogram Gradient Boosting Regressor machine learning model that: bins values for faster computation; offers native support for categorical features without the need for one-hot encoding (dummy variables); has native support for missing values not requiring removal of features/samples or imputation procedures; obviates the need to scale features as it is based on decision trees; allows for non-linear relationships, making no assumptions about underlying structure; and is capable modelling feature interactions. The native support for both categorical features and missing data, together with not needing to perform scaling, reduces the time and effort required during data pre-processing. This is especially useful in the feature selection stage of a study where many features that do not end up in the model would otherwise still have to undergo those pre-processing steps.

The use of a tree-based machine learning model such as HGBR leads to another strength in that it allowed for exploration of the input-output relationships by way of the TreeExplainer explainable machine learning method from SHAP. So far, TreeExplainer is the only SHAP method that allows for exact computation of Shapley values, which with theoretical grounding in game theory, are used to assess the contributions of features to the model output. SHAP values allow for visualisations of input-output relationships and of the contributions of feature interactions. They can also be used to derive feature importance metrics that are built up from the contributions from each individual sample in the test data.

With the SHAP value versus feature plots, one can recognise the presence of what could be considered as ‘floor’ and ‘ceiling’ effects in the features. This highlights the importance of using non-linear models in this type of research, as even if the relationship observed within the ‘active’ region of the feature is indeed linear, a linear model cannot detect the plateau regions and would instead return a model coefficient that underestimates the effect size in the ‘active’ region. Potential clinical cut-offs and regions of interest for certain features are identifiable, as detailed above, making the models highly interpretable for clinicians. Beyond the technical aspects, the visualisations made possible by the explainable machine learning methods are also a strength for the more clinical reader. Having run a complex machine learning model, not only the associations captured between features and model output can be observed, but also the relationships between feature interactions and the output. Cut-offs, regions of interest, clusters, trends are all on show which can allow for better insight and hypothesis generation.

Another strength of the study is the comparison of UGS, MGS, and GSR in terms of features selected to describe them from a range of 88 features across multiple domains. The TILDA data leveraged allowed for a large and granularly characterised sample size of 3925 participants, which represents a worthwhile contribution to the existing literature.

5.2.6 Limitations of Methodology and Study

However, while these cut-offs and regions of interest may be able to inform the clinician, it is possible that they may vary between populations. In terms of the analytical sample, this only included TILDA wave 3 participants who underwent the health centre assessment where gait speed tests were conducted. Even though at wave 1 TILDA was designed as a nationally representative cohort of people aged 50 or more years living in Ireland [30], the analytical sample at wave 3 may no longer be population-representative, and therefore results are not necessarily generalisable to the Irish population. Indeed, TILDA work showed that participants attending the health assessment centre were generally fitter than those having a health assessment in their homes [82], which means that other features may have been selected in the models if frailer people had been included in the analytical sample.

Despite having many advantages, the machine learning methodology also has limitations. The features selected need to be considered in terms of the 'package' of features chosen for the final model. Furthermore, it cannot be assumed that features not chosen for a model are necessarily non-predictive of the outcome variable.

Even though measures were put in place to help reduce overfitting (cross-validation on training data used in choosing features and hyperparameters, and models evaluated on a held-out test dataset), in the absence of an external validation sample, the risk of overfitting still exists. Despite using held-out test data, the absence of an external validation test means that the generalisability of the results is unknown. The confidence intervals of the effects and associations are also not known in this work; however, application of bootstrapping methods may be used in future work to address this limitation. A rigorous time complexity analysis was not performed but given its stepwise nature, the computation time of the feature-selection step scales with the square of the number of features considered. The number of hyperparameter iterations and the k-fold cross-validation in place also scale up the computation time, however, parallelising the code in future work could greatly reduce computation time. The computation time can also be reduced by the early stopping function that halts the feature selection if there is no improvement or a decline for two consecutive attempts. However, when (or if) this criterion is met depends on the data.

Furthermore, the models are dependent on the predictors that were entered. Even though the ‘shortlist’ of predictors was quite comprehensive (i.e., 88 features across 5 domains in experiment 2), consideration may not have been given to other potentially relevant predictors that were either not measured or realised. In view of GSR being less predictable than UGS and MGS, it is possible that including additional features in the GSR model (perhaps personality/social/lifestyle factors) would improve the model prediction. Height-normalised gait speed could have been considered in the models, but this is not something that I chose to consider a priori given the intended data-driven approach.

Another limitation is regarding sex differences in grip strength and height. Height may not be too much of an issue as it is non-modifiable and is a common choice for gait speed normalisation; but the thresholds observed in grip strength with respect to positive or negative deviation from the mean in UGS, MGS, or GSR are heavily distorted by sex. A sex-stratified investigation of grip strength in this context may be of clinical benefit in the future given its modifiable nature and its high importance in all three models.

Finally, it must be made clear that despite the use of word ‘impact’ when explaining the relationship of input features to the output, all results are associations and causal relationships cannot be assumed.

5.2.7 Potential Clinical Relevance

The five features selected for all three models (age, grip strength, chair stands time, mean motor response time, and height) showed common factors affecting UGS, MGS, and GSR.

Whilst there are similarities between the three gait speed models, the differences in features chosen for each model suggest that there are physiological differences in the nature of the three gait variables. This was also suggested in the different clinical associations between the gait speeds and clinical outcomes such as falls and faints. In the domain of psychology and cognition, UGS and MGS differed the most, with UGS being associated with depressive symptoms, whilst MGS was associated with cognitive performance in the SART and MOCA tests (MOCA was also associated with GSR). Education was associated with MGS and GSR but not with UGS. Fear of falling being present in MGS and GSR but not in UGS could suggest that the fear may not be in relation to usual day-to-day activity and walking but instead towards moving out of one's "comfort zone." The unique presence of MMSE and a sound-induced flash illusion variable in the GSR model could suggest that GSR is comparatively more related to the cognitive and sensory domains. The sound-induced flash illusion test assesses multi-sensory integration. It may be possible that UGS is more reflective of baseline health and perhaps is more sensitive to negative health outcomes, leaning more towards the frailty end of the frailty-fitness spectrum [83]. MGS and GSR, on the other hand, may reflect more of the fitness end of the spectrum; the ability to go beyond baseline towards better fitness and more reserve but not necessarily less frailty. A potential clinical take away from this work is that modifiable associates could be targeted for a particular gait characteristic with a view to improving the higher-level aspects of health such as frailty or fitness that is more linked to that variable. Given that each of the gait speed variables was predictive of potentially different health outcomes, this work shows avenues for ultimately targeting modifiable predictors of clinically meaningful outcomes for older people's functional independence.

6 BOOTSTRAPPED CONFIDENCE INTERVALS ON SHAP VALUES.

These analyses were produced after the completion of the paper that the previous chapter are based on. A better idea of the uncertainty surrounding the effects of each feature was deemed necessary. Although the results presented in experiment 2 had some level of uncertainty measurement surrounding the performance of the model (through 5-fold cross validation on the training data, and evaluation on training and test data) it did not contain any information on the robustness of the input-output relationships themselves. The confidence intervals show how the input-output relationships (as determined by SHAP) for a given feature, vary across the 1000 bootstrapped test data sets.

6.1 USUAL GAIT SPEED

The mean and 95% confidence intervals (CI) surrounding the main, non-interacting, contributions from each feature on the UGS model output are visualised in Figure 20. Of note, one sees that above ~ 1100 ms PulseInterval_RS loses significance, as does dBP_SeatStandDiff above ~15 mmHg. VisualAcuityLeft appears to be almost totally insignificant at the 95% CI level.

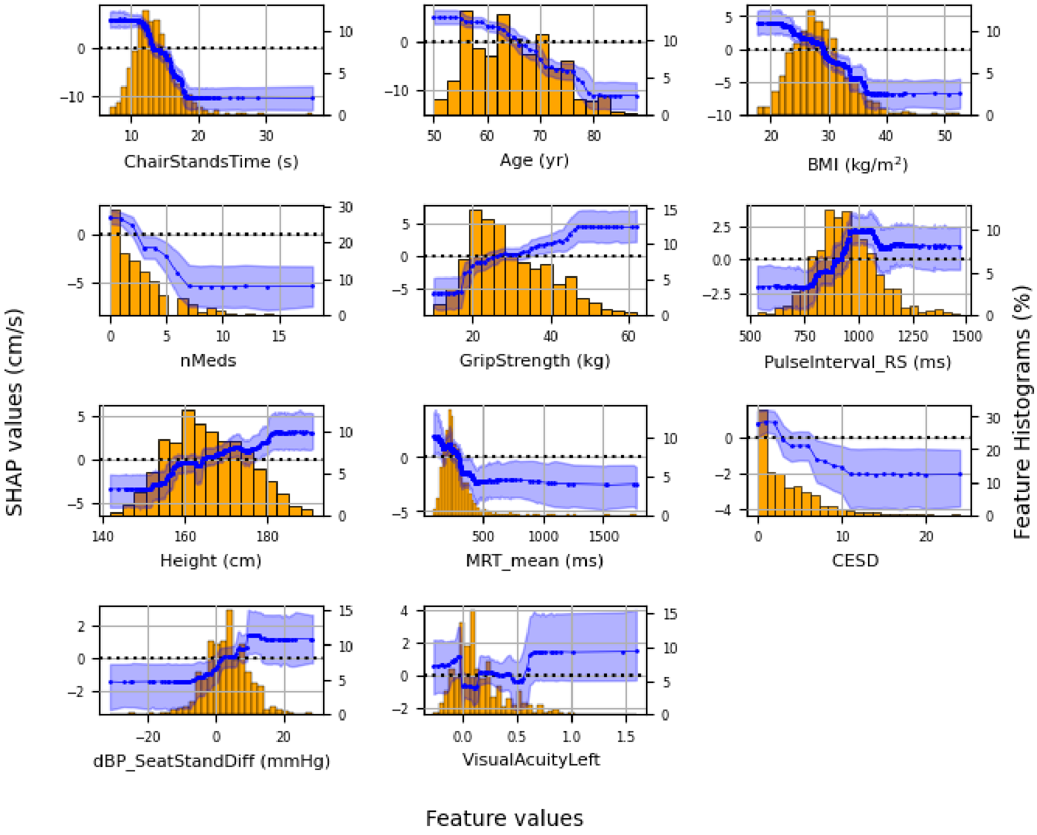


Figure 20. Bootstrapped mean and 95% confidence intervals for the SHAP values of the main (non-interacting) impacts of each feature in the prediction of usual gait speed. Each subplot pertains to an input feature and contains the mean and 95% CI values (left y-axis) and a histogram of the input feature (right y-axis). The SHAP values are in units of cm/s i.e., the same units as maximum gait speed. (Brief explanations of feature names can be found on page ix).

6.2 Maximum Gait Speed

The mean and 95% CI surrounding the main, non-interacting, contributions from each feature on the UGS model output are visualised in Figure 21. The lowest four features in terms of mean absolute SHAP value importance (HR_RS, HR_rMSSD_Paced, VisualAcuityBest, HR_TotalPower_Paced) appear to have an insignificant impact at the 95% CI level across most, if not all, of their range of input values.

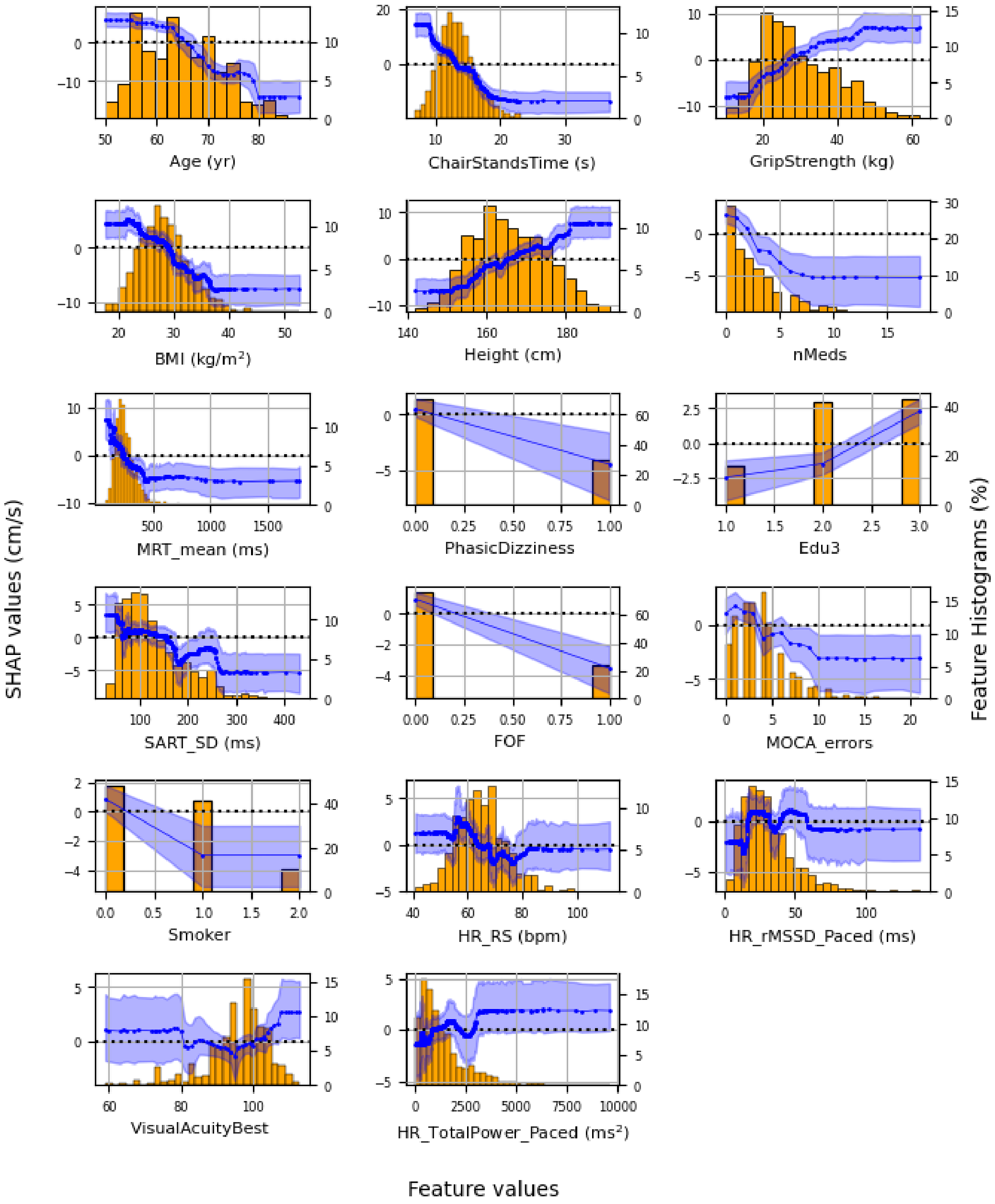


Figure 21. Bootstrapped mean and 95% confidence intervals for the SHAP values of the main (non-interacting) impacts of each feature in the prediction of maximum gait speed. Each subplot pertains to an input feature and contains the mean and 95% CI values (left y-axis) and a histogram of the input feature (right y-axis). The SHAP values are in units of cm/s i.e., the same units as usual gait speed. (Brief explanations of feature names can be found on page ix).

6.3 Gait Speed Reserve

The mean and 95% CI surrounding the main, non-interacting, contributions from each feature on the GSR model output are visualised in Figure 22. The mean and

95% CIs are displayed. Of note is that nCVDs appears non-significant above nCVD = 0. The CI's for nCVD are also asymmetric about the mean (the mean is closer to the upper CI bound) and quite wide.

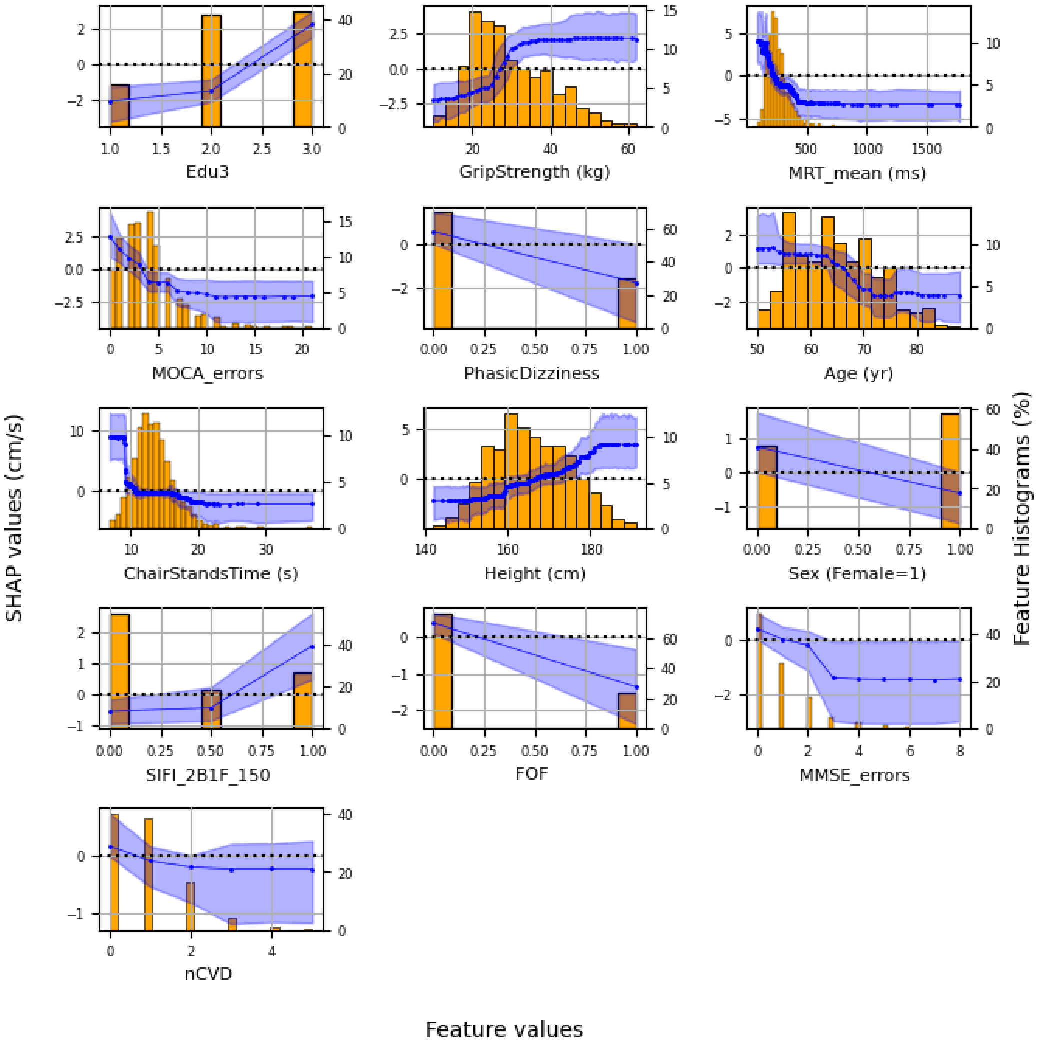


Figure 22. Bootstrapped mean and 95% confidence intervals for the SHAP values of the main (non-interacting) impacts of each feature in the prediction of gait speed reserve. Each subplot pertains to an input feature and contains the mean and 95% CI values (left y-axis) and a histogram of the input feature (right y-axis). The SHAP values are in units of cm/s i.e., the same units as gait speed reserve. (Brief explanations of feature names can be found on page ix).

6.4 Discussion

The bootstrapped confidence intervals mostly show robustness in the main effects of the features in each model and generally support the previous results. There were however some features that lacked confidence at the 95% level. These features are the nCVD for the GSR model; VisualAcuityLeft for the UGS model; and HR_RS, HR_rMSSD_Paced, VisualAcuityBest, and HR_TotalPower_Paced for the

MGS model. The CIs being asymmetric about the mean for nCVD could be due to the fact that a lot of participants probably have a cardiovascular disease of relatively low severity (high cholesterol, high blood pressure), and a smaller number of participants have comparatively more severe cardiovascular diseases such as history of congestive heart failure. The more common, lower severity diseases may have a smaller impact on GSR than that of the less common high severity diseases. The overall effect of nCVD on GSR is however not significant. The CI's also suggest that for some features, such as PulseInterval_RS in the UGS model, there are regions of values of the feature that have a more robust impact on the model output, and other regions that do not have significant effects.

The visual acuity features (VisualAcuityBest and VisualAcuity) were non-significant in the UGS and MGS models, respectively. The MGS model had the most features deemed insignificant; in addition to VisualAcuityBest there were three resting state cardiovascular features: HR_RS, HR_rMSSD_Paced, and HR_TotalPower_Paced.

7 Discussion

Experiment 1 was a first iteration of the concept of using feature selection to explore the multisystem nature of gait speed reserve (if there was indeed one to be found). It used a simple linear regression at its core and employed a 5-fold cross validation when evaluating each feature to avoid overfitting.

Table 7. Summary of scores and features selected for the usual gait speed, maximum gait speed, and both gait speed reserve models. Features are ordered from top to bottom by decreasing mean absolute SHAP value. The gait speed reserve features from experiment 1 are in descending order of linear model coefficient magnitude. Features in bold are deemed significant; underlined features are those selected for both gait speed reserve models. (Brief explanations of feature names can be found on page ix).

	Experiment 1	Experiment 2		
	Gait Speed Reserve	Gait Speed Reserve	Usual Gait Speed	Maximum Gait Speed
CV Mean R^2_{adj} (SD)	0.16 (0.03)	0.189 (0.02)	0.377 (0.04)	0.453 (0.04)
Train R^2_{adj}	0.18	0.224	0.427	0.545
Test R^2_{adj}	0.16	0.208	0.411	0.456
	Sex	**GripStrength**	**ChairStandsTime**	**Age**
	Edu3_Third/higher	**Edu3**	**Age**	**ChairStandsTime**
	MOCA	**ChairStandsTime**	**BMI**	**GripStrength**
	ChairStandsTime	**MOCA_errors**	**nMeds**	**BMI**
	Age	**SIFI_2B1F_150**	**GripStrength**	**Height**
	BMI	**FOF**	**PulseInterval_RS**	**nMeds**
	GripStrength	**Height**	**Height**	**MRT_mean**
	CardiacOutput_RS	**Age**	**MRT_mean**	**PhasicDizziness**
	nMeds	**Sex**	**CESD**	**Edu3**
	FOF	**PhasicDizziness**	**dBP_SeatStandDiff**	**SART_SD**
	PhasicDizziness	**MMSE_errors**	VisualAcuityLeft	**FOF**
	CRT_mean	nCVD		**MOCA_errors**
	SART_SD			**Smoker**
	Smoker			HR_RS
				HR_rMSSD_Paced
				VisualAcuityBest
				HR_TotalPower_Paced

The linear regression yields a simple process in terms of speed, no hyperparameters to tune, and convenience in terms of evaluating and interpreting the model using regression coefficients. The drawbacks of such a method include having to deal with missing data by either dropping samples (which is the approach that was taken) or imputation, the restriction of only being able to model linear associations, and the way categorical features are incorporated using one-hot encoded dummy variables.

Experiment 1 was ultimately superseded by experiment 2 in terms of methodology, number of features considered for selection, and in its inclusion of usual and maximum gait speed. The methodology allowed for a larger sample size due to the ability to handle missing data in the input features, a more sophisticated modelling of relationships between inputs and outputs, and a way to visualise and explore the learned relationships. The results of the two experiments are not directly comparable due to differences in sample and considered features but since there are similarities between the two, it suggests a robustness regarding them.

All four models selected features from multiple domains with several features appearing in every model and some that were unique to one or two of the gait variables. A summary of selected and significant features and metrics for all models is presented in Table 7.

7.1 Gait Speed Reserve Models

The linear regression-based feature selection applied to a shortlist of 34 features in experiment 1 yielded a $\overline{R_{adj}^2} \pm SD$ score from 5-fold CV of 0.16 ± 0.03, and R_{adj}^2 scores of 0.18 and 0.16 on the training and test data, respectively.

The histogram gradient boosting trees-based feature selection performed on 88 features achieved an $\overline{R_{adj}^2} \pm SD$ from 5-fold CV of 0.19 $\pm$ 0.02 and with training and test R_{adj}^2 scores of 0.22 and 0.21, respectively.

There was only a very modest improvement in performance for the experiment 2 model. Perhaps there were other features that were not included in these experiment that could better explain GSR, but the issue could be rooted in the variability of the UGS and MGS used to derive GSR. There is some amount of randomness or variability surrounding exactly how fast an individual will walk at a given time, and when deriving a feature like GSR, those random effects and variabilities may combine to produce a greater amount of variability around the derived feature.

Across the two methodologies, the selected features were mostly similar and covered multiple domains.

In experiment 1, 14 features were selected and 3 of them (phasic dizziness, standard deviation in SART, and smoking status) were not statistically significant in the final linear regression model. In experiment 2, 13 features were selected and from inspection of confidence intervals around SHAP values, one feature (number of cardiovascular diseases) appeared insignificant across the full range of input values.

Although a direct comparison cannot be made between the two experiments' results due to differing sample sizes and input feature shortlists, it seems worth observing what similarities exist between them in the prediction of GSR.

The features selected in both experiments were sex, educational attainment, MOCA, chair stands time, age, grip strength, fear of falling, and phasic dizziness. Other similarities were the selection of BMI in experiment 1 and height in experiment 2, and the selection of CRT_mean in experiment 1 and MRT_mean in experiment 2. With respect to these similarities, neither height, weight, nor MRT_mean were included in the experiment 1 shortlist.

Although causal inference is possible on observational data, it is however not possible to make any causal inferences from the experiments conducted in this work. To make a causal inference on the relationship between a variable and an outcome, a specific model must be constructed, following the rules of causal inference with directed acyclic graphs or some similar method, such that an appropriate set of controls are used. This must be done separately for each variable of interest; and no causal inferences can be made about the relationship between control variables and the outcome. However, with all that being said, we are still free to discuss and hypothesise about the potential causal links between variables and outcomes.

Sex may be generally assumed to be linked to gait speed via factors such as strength and height, but since both grip strength and height/BMI were selected alongside sex, it begs the question of what other sex differences may influence gait speeds, possibly psychological aspects or competitiveness. Grip strength itself is likely not causal to gait speeds but overall strength and vitality (which grip strength may be a marker of) could be. Height and BMI have potential causal links as longer legs are associated with faster gait. Chair stands time has a logical causal relationship to gait performance as better chair stands performance is a marker of better lower body strength and ability. MOCA is a test of global cognition and a possible link to gait performance here is related to one's ability to comprehend the task being given to them and also to assess their surrounding environment enough to comfortably perform the task. Fear of falling more obviously is causal to gait performance as those with such a fear are going to limit their gait speeds in order to ensure that they maintain balance and control. Possibly linked to fear of falling in some cases, those who experience orthostatic dizziness may have greater caution when walking, especially if they were seated beforehand. Aside from this aspect though, phasic dizziness is a complex variable with possible contributions from cardiovascular health, nutrition, sleep, stress, and psychology and so a potential causal link to gait speed is not clear; once again, phasic dizziness may just represent broader physical and mental health. Level of educational attainment again does not likely have a direct causal link to gait speed (i.e., getting a degree probably does not increase one's gait speed) but it is perhaps indicative of one's social situation and ability to pursue goals. A key aspect to keep in mind with educational attainment is that in the TILDA cohort, depending on when it was that certain levels of education were attained, the variable can act as a proxy for socio-economic status. For some older generations, education was not as it is today and for some even completing a second level education was not an option. Therefore, with educational attainment being to some degree a proxy for socio-economic status, it may be representative of factors such as stress, nutrition, employment, and healthcare throughout an individual's lifetime, all of which may impact current overall physical and mental health, which could in turn effect gait performance. Despite grip strength being the first feature selected in experiment 1 and sex being fifth, in the final model the linear model coefficient for sex was the biggest and that of grip strength was seventh (after BMI). In experiment 2, grip strength was selected second (after mean motor response time) and sex tenth. As judged by SHAP importance, those features remained in similar positions (sex being placed ninth, just after height). The explainable machine learning model suggests that even though grip strength is strongly predicted by sex, it is grip strength that has the greater association with GSR and not sex specifically. The differences in mean SHAP value for female vs. male sex (Figure 18 (A)) support this by showing that the difference of the impact of grip strength between men and women is greater than the difference of the impact of sex between sexes. Other statistically significant group mean differences between men and women were height (just below sex in terms of SHAP importance), mean motor response time, fear of falling, and SIFI multisensory integration, all of which had greater positive impacts in men. Education however, showed a significantly greater impact in women. Although sex in the model could act as a proxy for other factors associated with sex or gender, there could be more fundamental sex differences that impact gait speed reserve and further work could explore this. The lack of the linear model's ability to capture more nuanced relationships as in the explainable machine learning method may be why some of the effects of other features such as grip strength and height could have been encapsulated by sex.

Educational attainment was present in both models. The third-level/higher dummy variable was selected in the linear regression model which is consistent with that level of educational attainment showing the biggest impact for education in the explainable machine learning model. Figure 18 (B) and (C) explore education further by displaying the group mean SHAP value differences for third level/higher education vs. first/none and second level (B) and first level/none vs. second and third level/higher education (C). There were no significant differences in the former group comparison, but the latter showed significant group mean SHAP values for (in order of magnitude of difference) MOCA, age, mean motor response time, MMSE, SIFI multisensory integration, phasic dizziness, height, fear of falling, and sex. All but sex showed a negative difference, i.e., the impact of the features was smaller for the first level/none group; higher education was more positively impactful for women than for men. It’s worth reiterating here that for the generations involved in TILDA, educational attainment can often be used as a proxy for socio-economic status; typically, those from higher socio-economic backgrounds would be more likely to attain a higher level of education, and vice versa for those from lower economic backgrounds.

Phasic dizziness (orthostatic intolerance) was initially selected by the linear regression model but was deemed non-significant at the 95% CI level. Consistently, looking at the explainable machine learning model results, it is seen that phasic dizziness 95% CIs are bordering non-significance.

In terms of R^2_{adj}, the machine learning approach produced a slightly better score with less uncertainty. Although the higher score might be expected from the machine learning model, it is, however, not possible to concretely compare them due to differences in the feature shortlist. Moving forward from performance metrics to a perhaps more important aspect, the types of relationships that were captured and then displayed by the explainable machine learning methodology point to a significant advantage of that method over the simpler linear regression method. Taking age as a point of comparison, in experiment 2 there are different regions observed in the relationship between it and GSR: a region of slow decline, followed by a sudden steeper decrease between 63 and 70; followed roughly by a plateau region. In the linear model, the general average slope of the relationship does give a good idea that GSR declines with age but cannot provide the more nuanced multi-region relationship. Another example is chair stands time. The SHAP value plots showed that most of the range of values for chair stands time was quite unimpactful: between ~11 s and 16 s (which covers most of the centre of the distribution) the features had no impact on GSR, the right tail of the distribution showed a small negative impact on GSR. The left tail however showed a large positive impact on the model. Put more simply, across most of the distribution of chairs stand times there was no change in GSR; however, there was a big positive impact for very fast times, and a small negative impact for very slow times. A possible reason for this might be that UGS and MGS decline similarly with slower chair stands times but that those with very fast chair stands have a greater capacity for a very fast MGS.

The explainable machine learning model did not offer any information on interactions between features because for the final model the hyperparameter tuning resulted in a hyperparameter that effectively restricted the depth of the tree such that features could not interact. This could be due to random chance, or because including feature interactions worsened the performance. Future work could manually set the hyperparameters to allow for feature interaction so that this can be studied.

The use of highly visual methods in the inspection and interpretation of models requires or invokes a different philosophy than that of a standard linear regression model and its coefficients. The more complex machine learning models can learn relationships that are not easily boiled down to a single number. They instead require the analyst or expert to take a step back and view the relationship as a whole. Whilst still considering the values associated with features' impact on the model output, a more intuitive sense of the relationship may be gained visually, observing different regions of impact or the shape of the relationship.

In relation to potentially using these results for the improvement of older persons care, the main suggestion would be to routinely measure usual and maximum gait speeds and gait speed reserve, and if declines are observed, be they gradual or sudden, perhaps look to the variables and domains identified in this work as possible avenues for further investigation or rehabilitation.

Future work could include analyses stratified by sex or age group, and the study of different aspects of gait characteristics and dynamics such as cadence and variability. Other high-level indicators of general health, physical strength, and performance such as grip strength and chair stands could be studied in a similar way. The timed-up-and-go (TUG) test, which requires the participant to stand up, walk 3 m, turn around, walk 3 m back, and then sit down could also be of interest to explore as it is a more person-involved task that arguably requires more lower body strength, balance, spatial awareness, and cognitive input. TUG was not included in the present work as a predictor due to its strong correlation and similarity with the gait tasks. Personality is another factor not included in this work that could play a role in one's gait speeds. Personality could impact the way individuals interpret and response to the tasks and could possibly be a factor in walking speeds itself. Another advancement from this work could be to investigate the causal effects of some of the selected features though careful and dedicated causal inference modelling.

8 CONCLUSIONS

In experiment 1, a stepwise linear regression-based machine learning pipeline was used to select the most important gait speed reserve predictors from 34 shortlisted features from multiple domains. Variables were selected one at a time such that they maximised the mean R_{adj}^2 score from a 5-fold cross-validation. A peak score of (0.16 ± 0.03) was achieved with 14 variables (giving R_{adj}^2 of 0.18 and 0.16 on 80% training and 20% test data, respectively). Of the 14 selected features, 11 had statistically significant ($p<0.05$) effects in the model: sex, MOCA, third level education, chair stands time, age, BMI, grip strength, cardiac output, number of medications, fear of falling, and mean choice reaction time. The three selected but statistically insignificant features were phasic dizziness, standard deviation in sustained attention to response task times, and smoking status.

In experiment 2, a non-parametric machine learning based stepwise feature selection was used to find the sets of predictors from an expanded shortlist of 88 input features that resulted in the best explained variance in UGS, MGS, and GSR. Explainable machine learning methods were then used to analyse and explore each of the models. The variables selected with a histogram gradient boosting regressor based machine learning stepwise feature selection explained a greater proportion of variation in MGS and UGS than GSR. There were common features to all three models (i.e., age, grip strength, chair stands time, mean motor reaction time in the choice reaction time test, and height), but also some unique features to each of them. By SHAP feature importance, the top four features were chair stands time, age, BMI, and number of medications to the UGS model; age, chair stands time, grip strength, and height to the MGS model; and level of educational attainment, grip strength, mean motor response time, and MOCA errors to the GSR model. Overall, findings on all three models were clinically plausible and support a network physiology approach [84] to the understanding of predictors of performance-based tasks. Each model contained features from multiple physiological systems, and this supports the hypothesis that GSR as well as UGS and MGS are multisystem-driven phenomena. By employing an explainable machine learning model, observations may help clinicians gain new insights into the possible determinants of physiological reserve in community-dwelling older adults. Of the features selected, some are prospectively non-modifiable (e.g., age, sex, height, educational attainment). Others, however, may be directly modifiable through changes in lifestyle, engaging in physical exercise, or cognitive stimulation (e.g., BMI, weight, smoking, chair stands time, grip strength, MOCA, motor response time, SART). For some variables, it may be useful to focus on ensuring that an older person avoids reaching threshold values that are associated with a rapid decline in gait speed. Conversely, if engaging in rehabilitation, those threshold values may be the targets so as to reach a more stable situation with respect to walking speed. Having explored the predictors of GSR and found multisystem associations, further work could investigate whether GSR is a useful measure in predicting additional adverse health outcomes (other than falls) and if it can contribute to informing on overall physiological reserve.

The machine learning approach allowed for the stepwise selection of the set features that best explained a target variable in a non-parametric manner that can also capture high-order interactions. Explainable machine learning allowed for the selected models to be visualised to observe the input-output relationships, and the relationship between feature interactions and the model output. Using a tree-based machine learning model enabled the use of the TreeSHAP explainable machine learning package, which uses the tree structure to be able to compute exact Shapley values in low-order polynomial time.

The use of bootstrapped confidence intervals on the SHAP values of the main effects from each feature allowed for a better exploration of the relationship between the features and the outputs. Uncertainty can vary across the range of input values for a given feature. A visual inspection of the 95% CI plots informed on the overall confidence of a features effect. Some features displayed non-significance across almost their entire input value range (resting state heart rate, root-mean-sum-of-squared-differences between successive heartbeats during paced breathing, best visual acuity, total power of heartbeat during paced breathing for maximum gait speed, and left visual acuity for usual gait speed, and number of cardiovascular diseases for gait speed reserve), and others were seen to be significant in certain ranges of input value (resting state pulse interval and difference between seated and standing diastolic blood pressure for usual gait speed).

The linear model and the machine learning model selected similar features from multiple domains (sex, age, education, height/BMI, MOCA, grip strength, chair stands time, and fear of falling) in the prediction of gait speed reserve. The linear modelling was faster and simpler, but the explainable machine learning method was capable of non-parametrically modelling non-linear relationships, inherently capturing feature interactions, and various forms of visualisation of modelled relationships. These can help to provide better insights into the role of features in a model and can allow analysts and clinicians to interact and interpret the models in both more granular and holistic ways. Overall, findings of this investigation support a network physiology approach to the study of physiological reserve and could help policy makers and clinicians design strategies to promote resilience and functional independence in community-dwelling older adults.

9 REFERENCES

1. Wu, T. and Y. Zhao, *Associations between functional fitness and walking speed in older adults.* Geriatr Nurs, 2021. **42**(2): p. 540-543.

2. Aldridge, C., et al., *Walking at work: Maximum gait speed is related to work ability in hospital nursing staff.* J Occup Health, 2020. **62**(1): p. e12171.

3. Bohannon, R.W., *Walking after stroke: comfortable versus maximum safe speed.* Int J Rehabil Res, 1992. **15**(3): p. 246-8.

4. Kawajiri, H., et al., *Maximum Walking Speed at Discharge Could Be a Prognostic Factor for Vascular Events in Patients With Mild Stroke: A Cohort Study.* Arch Phys Med Rehabil, 2019. **100**(2): p. 230-238.

5. Kollen, B., G. Kwakkel, and E. Lindeman, *Hemiplegic gait after stroke: is measurement of maximum speed required?* Arch Phys Med Rehabil, 2006. **87**(3): p. 358-63.

6. Nogueron Garcia, A., et al., *Gait plasticity impairment as an early frailty biomarker.* Exp Gerontol, 2020. **142**: p. 111137.

7. Samson, M.M., et al., *Differences in gait parameters at a preferred walking speed in healthy subjects due to age, height and body weight.* Aging (Milano), 2001. **13**(1): p. 16-21.

8. James, E.G., et al., *Walking Speed Affects Gait Coordination and Variability Among Older Adults With and Without Mobility Limitations.* Arch Phys Med Rehabil, 2020. **101**(8): p. 1377-1382.

9. O'Donoghue, P.J., et al., *Association between gait speed and the SHARE Frailty Instrument in a Falls and Syncope Clinic.* Eur Geriatr Med, 2021. **12**(5): p. 1101-1105.

10. Umegaki, H., et al., *Maximum gait speed is associated with a wide range of cognitive functions in Japanese older adults with a Clinical Dementia Rating of 0.5.* Geriatr Gerontol Int, 2018. **18**(9): p. 1323-1329.

11. Noguerón García, A., et al., *Gait plasticity impairment as an early frailty biomarker.* Experimental Gerontology, 2020. **142**: p. 111137.

12. do Carmo Correia de Lima, M., et al., *Maximum Walking Speed Can Improve the Diagnostic Value of Frailty among Community-Dwelling Older Adults a Cross-Sectional Study.* J Frailty Aging, 2019. **8**(1): p. 39-41.

13. Belle, V. and I. Papantonis, *Principles and Practice of Explainable Machine Learning.* Frontiers in Big Data, 2021. **4**(39).

14. Friedman, J.H., *Greedy function approximation: A gradient boosting machine.* The Annals of Statistics, 2001. **29**(5): p. 1189-1232.

15. Goldstein, A., et al., *Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation.* arXiv pre-print server, 2014.

16. Ribeiro, M.T., S. Singh, and C. Guestrin, *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, Association for Computing Machinery: San Francisco, California, USA. p. 1135–1144.

17. Lundberg, S.M. and S.-I. Lee, *A unified approach to interpreting model predictions*, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, Curran Associates Inc.: Long Beach, California, USA. p. 4768–4777.

18. Lundberg, S.M., et al., *From local explanations to global understanding with explainable AI for trees.* Nature Machine Intelligence, 2020. **2**(1): p. 56-67.

19. Shapley, L.S., *Contributions to the Theory of Games (AM-28), Volume II*, in *17. A Value for n-Person Games*, K. Harold William and T. Albert William, Editors. 1953, Princeton University Press. p. 307-318.

20. Davis, J., et al. *A linear regression-based machine learning pipeline for the discovery of clinically relevant correlates of gait speed reserve from multiple physiological systems*. in *2021 29th European Signal Processing Conference (EUSIPCO)*. 2021. Dublin, Ireland: IEEE.

21. Davis, J.R.C., et al., *Comparison of Gait Speed Reserve, Usual Gait Speed, and Maximum Gait Speed of Adults Aged 50+ in Ireland Using Explainable Machine Learning.* Frontiers in Network Physiology, 2021. **1**.

22. Kearney, P.M., et al., *Cohort Profile: The Irish Longitudinal Study on Ageing.* International Journal of Epidemiology, 2011. **40**(4): p. 877-884.

23. Ewing, J.A., *Detecting alcoholism. The CAGE questionnaire.* JAMA: The Journal of the American Medical Association, 1984. **252**(14): p. 1905-1907.

24. Knight, S.P., et al., *Associations between Neurocardiovascular Signal Entropy and Physical Frailty.* Entropy (Basel), 2020. **23**(1).

25. Donoghue, O.A., et al., *Is orthostatic hypotension and co-existing supine and seated hypertension associated with future falls in community-dwelling older adults? Results from The Irish Longitudinal Study on Ageing (TILDA).* PLoS One, 2021. **16**(5): p. e0252212.

26. Finucane, C., et al., *Age-related normative changes in phasic orthostatic blood pressure in a large population study: findings from The Irish Longitudinal Study on Ageing (TILDA).* Circulation, 2014. **130**(20): p. 1780-9.

27. Frewen, J., et al., *Cognitive function is associated with impaired heart rate variability in ageing adults: the Irish longitudinal study on ageing wave one results.* Clin Auton Res, 2013. **23**(6): p. 313-23.

28. Pardey, J. and S. Jouravleva. *The next-generation holter revolution: from analyse-edit-print to analyse-print*. 2004. IEEE.

29. Nolan, H., L. Newman, and O. Donoghue, *Chapter 5: Objective Indicators of Health and Function (TILDA Wave 3 Report). Available on: https://tilda.tcd.ie/publications/reports/pdf/w3-key-findings-report/Chapter%205.pdf (accessed 26 June 2021).* 2016.

30. Donoghue, O.A., et al., *Cohort Profile Update: The Irish Longitudinal Study on Ageing (TILDA).* Int J Epidemiol, 2018. **47**(5): p. 1398-1398l.

31. Nasreddine, Z.S., et al., *The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment.* J Am Geriatr Soc, 2005. **53**(4): p. 695-9.

32. Arevalo-Rodriguez, I., et al., *Mini-Mental State Examination (MMSE) for the detection of Alzheimer's disease and other dementias in people with mild cognitive impairment (MCI).* Cochrane Database Syst Rev, 2015(3): p. CD010783.

33. Chintapalli, R. and R. Romero-Ortuno, *Choice reaction time and subsequent mobility decline: Prospective observational findings from The Irish Longitudinal Study on Ageing (TILDA).* EClinicalMedicine, 2021. **31**: p. 100676.

34. O'Halloran, A.M., et al., *Sustained attention and frailty in the older adult population.* J Gerontol B Psychol Sci Soc Sci, 2014. **69**(2): p. 147-56.

35. Radloff, L.S., *The CES-D Scale.* Applied Psychological Measurement, 1977. **1**(3): p. 385-401.

36. Duggan, E., et al., *Time to refocus assessment of vision in older adults? Contrast sensitivity but not visual acuity is associated with gait in older adults.* Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences, 2017. **72**(12): p. 1663-1668.

37. Shams, L., Y. Kamitani, and S. Shimojo, *Visual illusion induced by sound.* Cognitive Brain Research, 2002. **14**(1): p. 147-152.

38. Hirst, R.J., et al., *Grey matter volume in the right Angular Gyrus is associated with differential patterns of multisensory integration with ageing.* Neurobiology of Aging, 2020.

39. Hirst, R.J., et al., *The effect of eye disease, cataract surgery and hearing aid use on multisensory integration in ageing.* Cortex, 2020. **133**: p. 161-176.

40. Pedregosa, F., et al., *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 2012. **12**.

41. Ke, G., et al., *LightGBM: a highly efficient gradient boosting decision tree*, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, Curran Associates Inc.: Long Beach, California, USA. p. 3149–3157.

42. Moncada-Torres, A., et al., *Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival.* Scientific Reports, 2021. **11**(1).

43. Ke, G., et al., *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*, in *Advances in Neural Information Processing Systems 30 (NIP 2017)*. 2017.

44. Slack, D., et al. *Fooling LIME and SHAP*. ACM.

45. Tibaek, S., et al., *Reference values of maximum walking speed among independent community-dwelling Danish adults aged 60 to 79 years: a cross-sectional study.* Physiotherapy, 2015. **101**(2): p. 135-40.

46. Izawa, K.P., et al., *Gender-related differences in maximum gait speed and daily physical activity in elderly hospitalized cardiac inpatients: a preliminary study.* Medicine (Baltimore), 2015. **94**(11): p. e623.

47. Kyronlahti, S.M., et al., *Educational Differences in Decline in Maximum Gait Speed in Older Adults over an 11-year Follow-up.* J Gerontol A Biol Sci Med Sci, 2020.

48. Bohannon, R.W., *Comfortable and maximum walking speed of adults aged 20-79 years: reference values and determinants.* Age Ageing, 1997. **26**(1): p. 15-9.

49. Clark, D.J., et al., *Neuromuscular determinants of maximum walking speed in well-functioning older adults.* Exp Gerontol, 2013. **48**(3): p. 358-63.

50. Iwata, A., et al., *Maximum movement velocity of the upper limbs reflects maximum gait speed in community-dwelling adults aged older than 60 years.* Geriatr Gerontol Int, 2014. **14**(4): p. 886-91.

51. Rantanen, T., et al., *Association of muscle strength with maximum walking speed in disabled older women.* Am J Phys Med Rehabil, 1998. **77**(4): p. 299-305.

52. Feng, C., et al., *A Systematic Review and Meta-Analysis of Exercise Interventions and Use of Exercise Principles to Reduce Fear of Falling in Community-Dwelling Older Adults.* Physical Therapy, 2022. **102**(1).

53. Ueno, K., et al., *Usefulness of measuring maximal gait speed in conjunction with usual gait speed for risk stratification in patients with cardiovascular disease.* Experimental Gerontology, 2022. **164**: p. 111810.

54. Ferguson, C.J., *An Effect Size Primer: A Guide for Clinicians and Researchers.* Professional Psychology: Research and Practice, 2009. **40**: p. 532–538.

55. Bongers, K.T., et al., *The predictive value of gait speed and maximum step length for falling in community-dwelling older persons.* Age Ageing, 2015. **44**(2): p. 294-9.

56. Kearney, P.M., et al., *Comparison of centre and home-based health assessments: early experience from the Irish Longitudinal Study on Ageing (TILDA).* Age and Ageing, 2010. **40**(1): p. 85-90.

57. Rizzo, R., et al., *Network Physiology of Cortico-Muscular Interactions.* Front Physiol, 2020. **11**: p. 558070.

58. Middleton, A., et al., *Self-Selected and Maximal Walking Speeds Provide Greater Insight Into Fall Status Than Walking Speed Reserve Among Community-Dwelling Older Adults.* American Journal of Physical Medicine &amp; Rehabilitation, 2016. **95**(7): p. 475-482.

59. Romero-Ortuno, R., et al., *Do older pedestrians have enough time to cross roads in Dublin? A critique of the Traffic Management Guidelines based on clinical research findings.* Age Ageing, 2010. **39**(1): p. 80-6.

60. Schimpl, M., et al., *Association between walking speed and age in healthy, free-living individuals using mobile accelerometry--a cross-sectional study.* PLoS One, 2011. **6**(8): p. e23299.

61. Kasovic, M., L. Stefan, and A. Stefan, *Normative Data for Gait Speed and Height Norm Speed in >/= 60-Year-Old Men and Women.* Clin Interv Aging, 2021. **16**: p. 225-230.

62. Kenny, R.A., et al., *Normative values of cognitive and physical function in older adults: findings from the Irish Longitudinal Study on Ageing.* J Am Geriatr Soc, 2013. **61 Suppl 2**: p. S279-90.

63. Cruz-Jentoft, A.J., et al., *Sarcopenia: revised European consensus on definition and diagnosis.* Age Ageing, 2019. **48**(1): p. 16-31.

64. Moreira, V.G., M. Perez, and R.A. Lourenco, *Prevalence of sarcopenia and its associated factors: the impact of muscle mass, gait speed, and handgrip strength reference values on reported frequencies.* Clinics (Sao Paulo), 2019. **74**: p. e477.

65. Perez-Sousa, M.A., et al., *Gait speed as a mediator of the effect of sarcopenia on dependency in activities of daily living.* J Cachexia Sarcopenia Muscle, 2019. **10**(5): p. 1009-1015.

66. Nishimura, T., et al., *Usefulness of chair stand time as a surrogate of gait speed in diagnosing sarcopenia.* Geriatr Gerontol Int, 2017. **17**(4): p. 659-661.

67. Cai, Y., et al., *Participation in cognitive activities is associated with foot reaction time and gait speed in older adults.* Aging Clin Exp Res, 2020.

68. Fernandez Menendez, A., et al., *The Determinants of the Preferred Walking Speed in Individuals with Obesity.* Obes Facts, 2019. **12**(5): p. 543-553.

69. Naples, J.G., et al., *Impact of Drug-Drug and Drug-Disease Interactions on Gait Speed in Community-Dwelling Older Adults.* Drugs Aging, 2016. **33**(6): p. 411-8.

70. Borda, M.G., et al., *Educational level and its Association with the domains of the Montreal Cognitive Assessment Test.* Aging Ment Health, 2019. **23**(10): p. 1300-1306.

71. Wong, G.K., et al., *Comparison of montreal cognitive assessment and mini-mental state examination in evaluating cognitive domain deficit following aneurysmal subarachnoid haemorrhage.* PLoS One, 2013. **8**(4): p. e59946.

72. Briggs, R., et al., *What Is the Relationship Between Orthostatic Blood Pressure and Spatiotemporal Gait in Later Life?* J Am Geriatr Soc, 2020. **68**(6): p. 1286-1292.

73. Jeon, E.J., et al., *Clinical significance of orthostatic dizziness in the diagnosis of benign paroxysmal positional vertigo and orthostatic intolerance.* Am J Otolaryngol, 2013. **34**(5): p. 471-6.

74. Schniepp, R., et al., *Gait characteristics of patients with phobic postural vertigo: effects of fear of falling, attention, and visual input.* J Neurol, 2014. **261**(4): p. 738-46.

75. Bueno, G.A.S., et al., *Fear of Falling Contributing to Cautious Gait Pattern in Women Exposed to a Fictional Disturbing Factor: A Non-randomized Clinical Trial.* Front Neurol, 2019. **10**: p. 283.

76. Brandler, T.C., et al., *Depressive symptoms and gait dysfunction in the elderly.* Am J Geriatr Psychiatry, 2012. **20**(5): p. 425-32.

77. Abad, C.C., et al., *Cardiac autonomic control in high level Brazilian power and endurance track-and-field athletes.* Int J Sports Med, 2014. **35**(9): p. 772-8.

78. Killane, I., et al., *Relative association of processing speed, short-term memory and sustained attention with task on gait speed: a study of*

community-dwelling people 50 years and older. J Gerontol A Biol Sci Med Sci, 2014. **69**(11): p. 1407-14.

79. Niedermaier, O.N., et al., *Influence of cigarette smoking on human autonomic function.* Circulation, 1993. **88**(2): p. 562-71.

80. Mahoney, J.R., K. Cotton, and J. Verghese, *Multisensory Integration Predicts Balance and Falls in Older Adults.* J Gerontol A Biol Sci Med Sci, 2019. **74**(9): p. 1429-1435.

81. Stapleton, J., et al., *A standing posture is associated with increased susceptibility to the sound-induced flash illusion in fall-prone older adults.* Exp Brain Res, 2014. **232**(2): p. 423-34.

82. Kearney, P.M., et al., *Comparison of centre and home-based health assessments: early experience from the Irish Longitudinal Study on Ageing (TILDA).* Age Ageing, 2011. **40**(1): p. 85-90.

83. Romero-Ortuno, R. and D. O'Shea, *Fitness and frailty: opposite ends of a challenging continuum! Will the end of age discrimination make frailty assessments an imperative?* Age Ageing, 2013. **42**(3): p. 279-80.

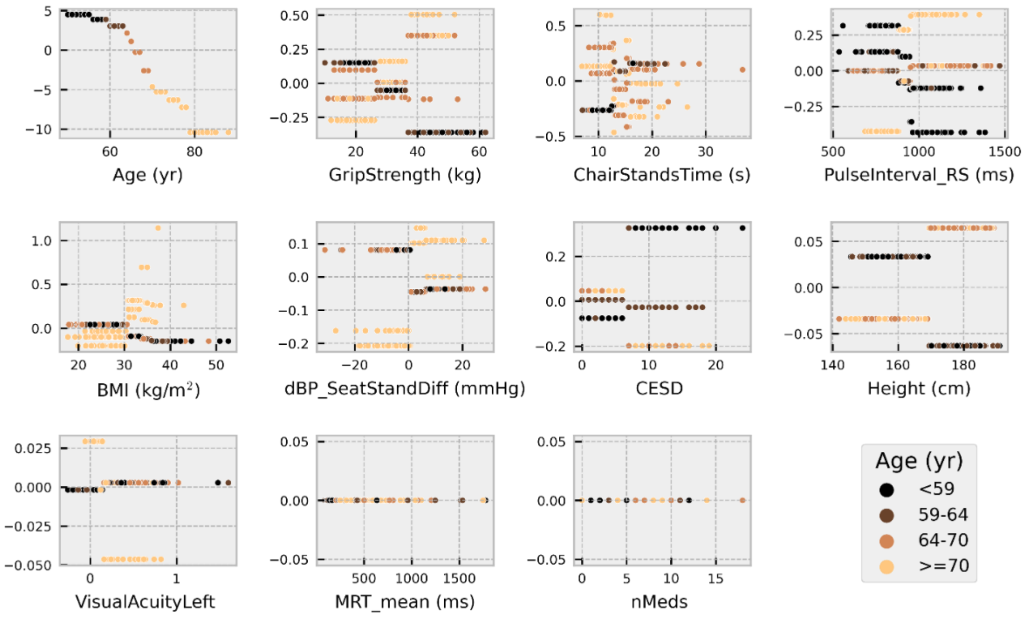
84. Bartsch, R.P., et al., *Network Physiology: How Organ Systems Dynamically Interact.* PLoS One, 2015. **10**(11): p. e0142143.

10 APPENDICES

10.1 APPENDIX A

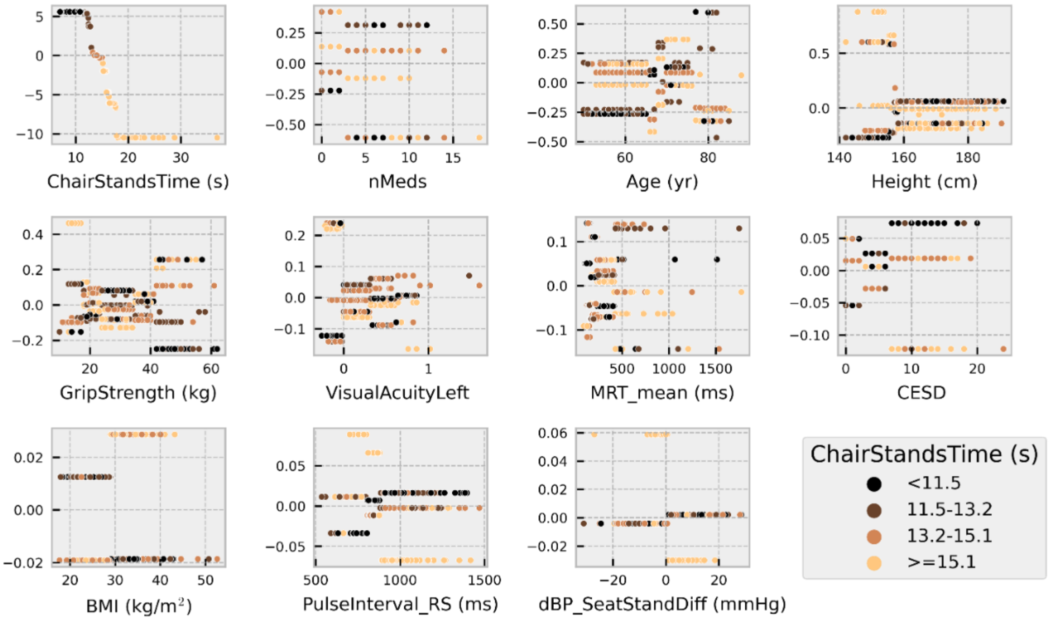
Scatter plots of the four top interaction effects on the **usual gait speed model** (i.e., age, chair stands time, body mass index, and grip strength). In the scatter plots, the points are coloured according to the value of the main interaction feature. The interactions are computed for the features in whatever numerical form they exist in but for ease of visualisation, continuous features are coloured according to what quartile a particular sample's value falls in; black indicates the value is in the lowest quartile and light brown the highest quartile. In each figure, the subplots are ordered from top-left to bottom-right by decreasing mean absolute SHAP interaction value.

10.1.1 Age Interactions



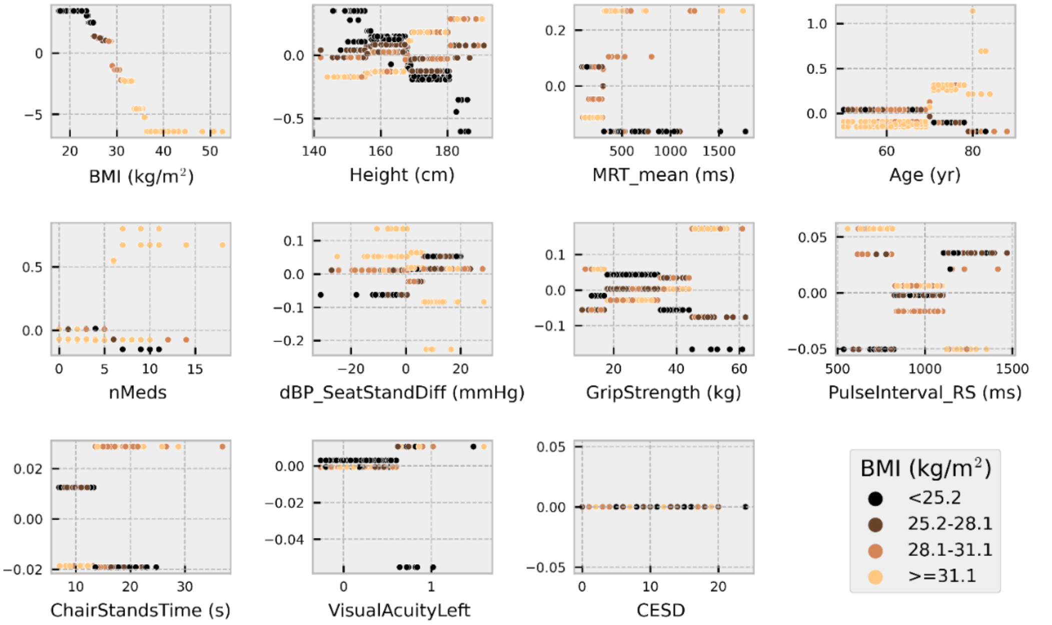
Appendix A Figure 1. SHAP values for the interaction between age and all features in the usual speed model. The x-axes present the values of a feature in the units of that feature. The colour of a sspoint indicates what quartile for age that sample falls in with black being lowest and light brown being highest. The y-axis displays the SHAP value of the interaction.

10.1.2 Chair Stands Time Interactions



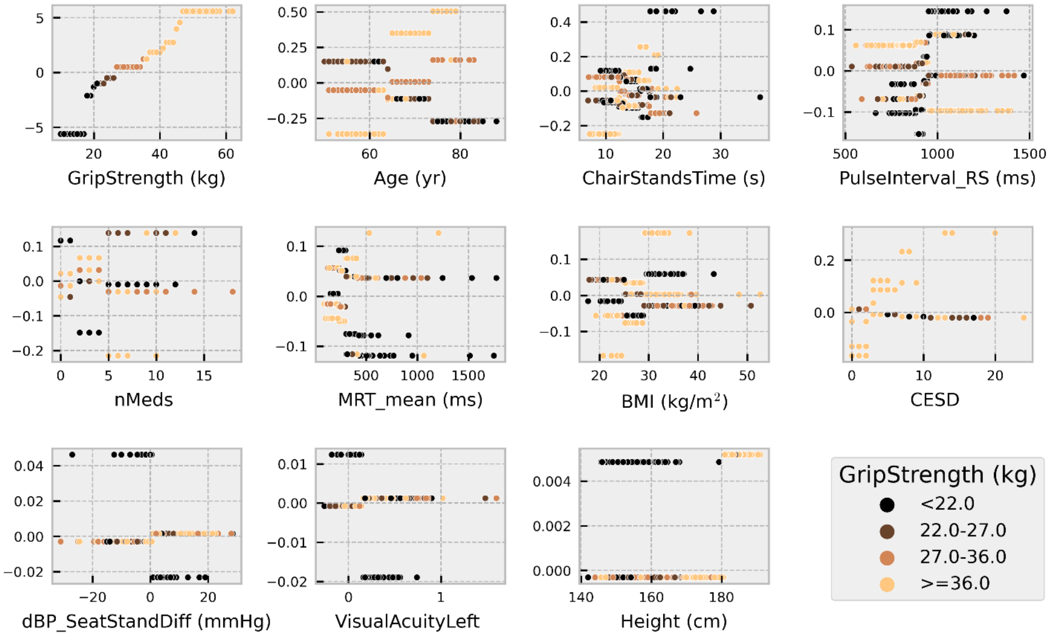
Appendix A Figure 2. SHAP values for the interaction between chair stands time and all features in the usual speed model. The x-axes present the values of a feature in the units of that feature. The colour of a point indicates what quartile for chair stands time that sample falls in with black being lowest and light brown being highest. The y-axis displays the SHAP value of the interaction.

10.1.3 BMI Interactions



Appendix A Figure 3. SHAP values for the interaction between BMI and all features in the usual speed model. The x-axes present the values of a feature in the units of that feature. The colour of a point indicates what quartile for BMI that sample falls in with black being lowest and light brown being highest. The y-axis displays the SHAP value of the interaction.

10.1.4 Grip Strength Interactions

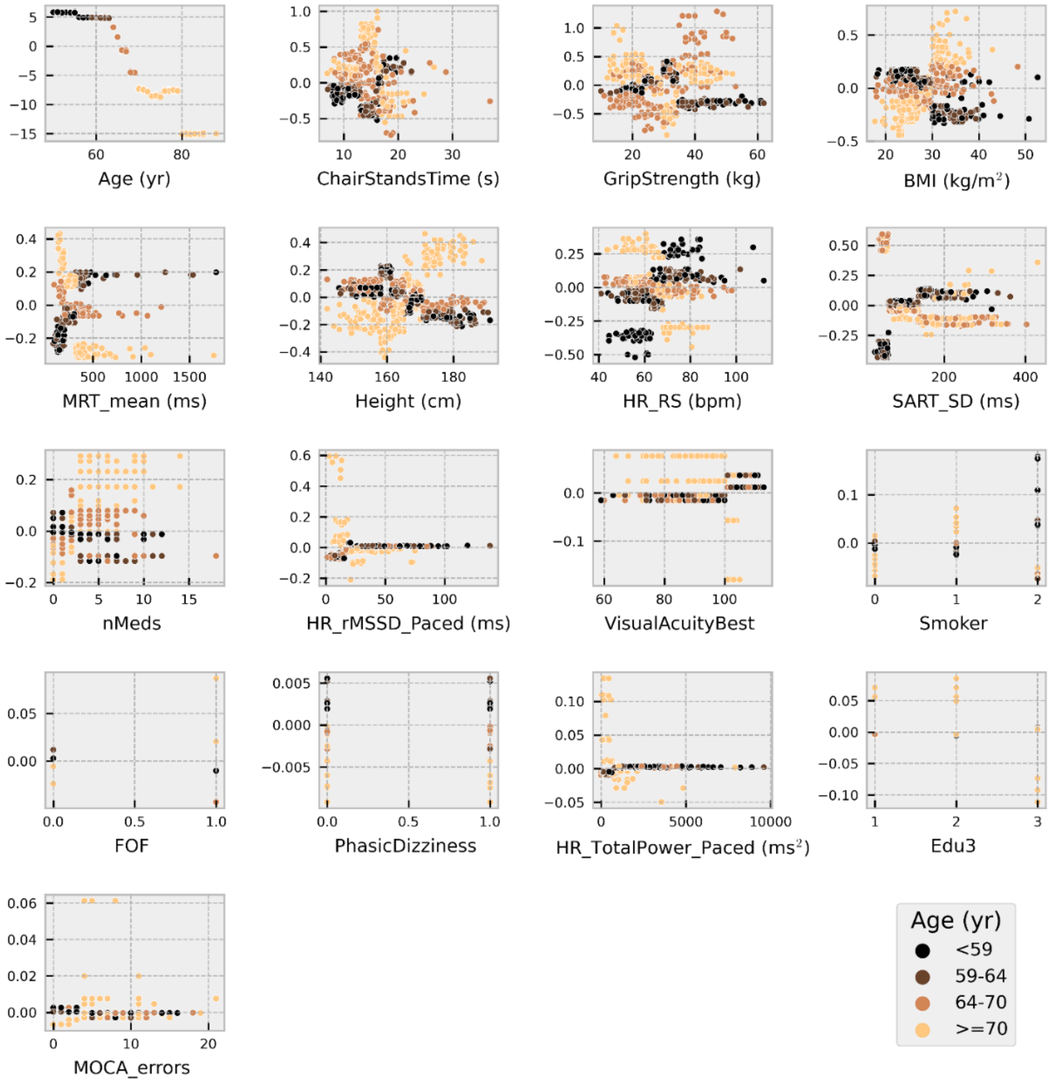


Appendix A Figure 4. SHAP values for the interaction between grip strength and all features in the usual speed model. The x-axes present the values of a feature in the units of that feature. The colour of a point indicates what quartile for grip strength that sample falls in with black being lowest and light brown being highest. The y-axis displays the SHAP value of the interaction.

10.2 APPENDIX B

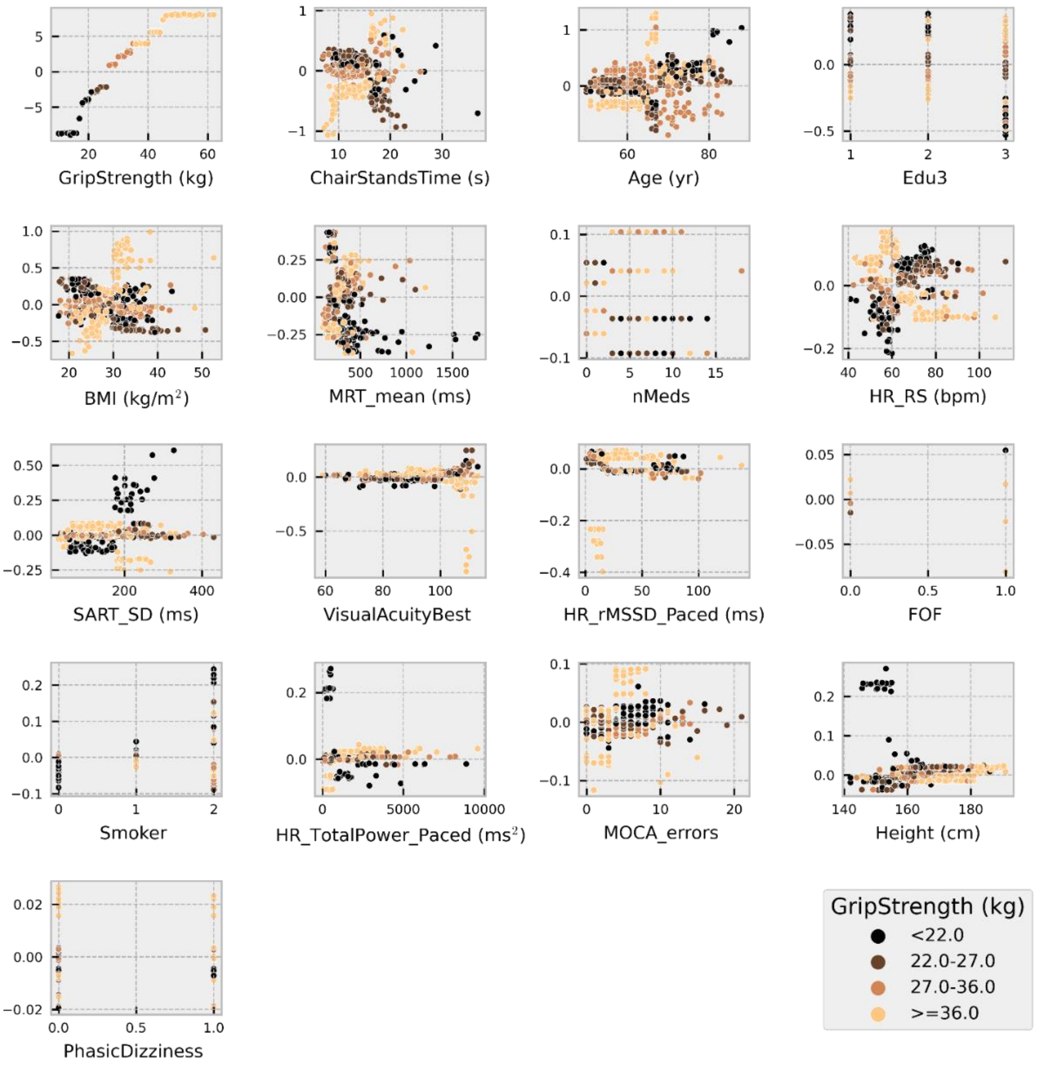
Scatter plots of the four top interaction effects on the **maximum gait speed model** (i.e., age, grip strength, chair stands time, and body mass index). In the scatter plots, the points are coloured according to the value of the main interaction feature. The interactions are computed for the features in whatever numerical form they exist in but for ease of visualisation, continuous features are coloured according to what quartile a particular sample's value falls in; black indicates the value is in the lowest quartile and light brown the highest quartile. In each figure, the subplots are ordered from top-left to bottom-right by decreasing mean absolute SHAP interaction value.

10.2.1 Age Interactions



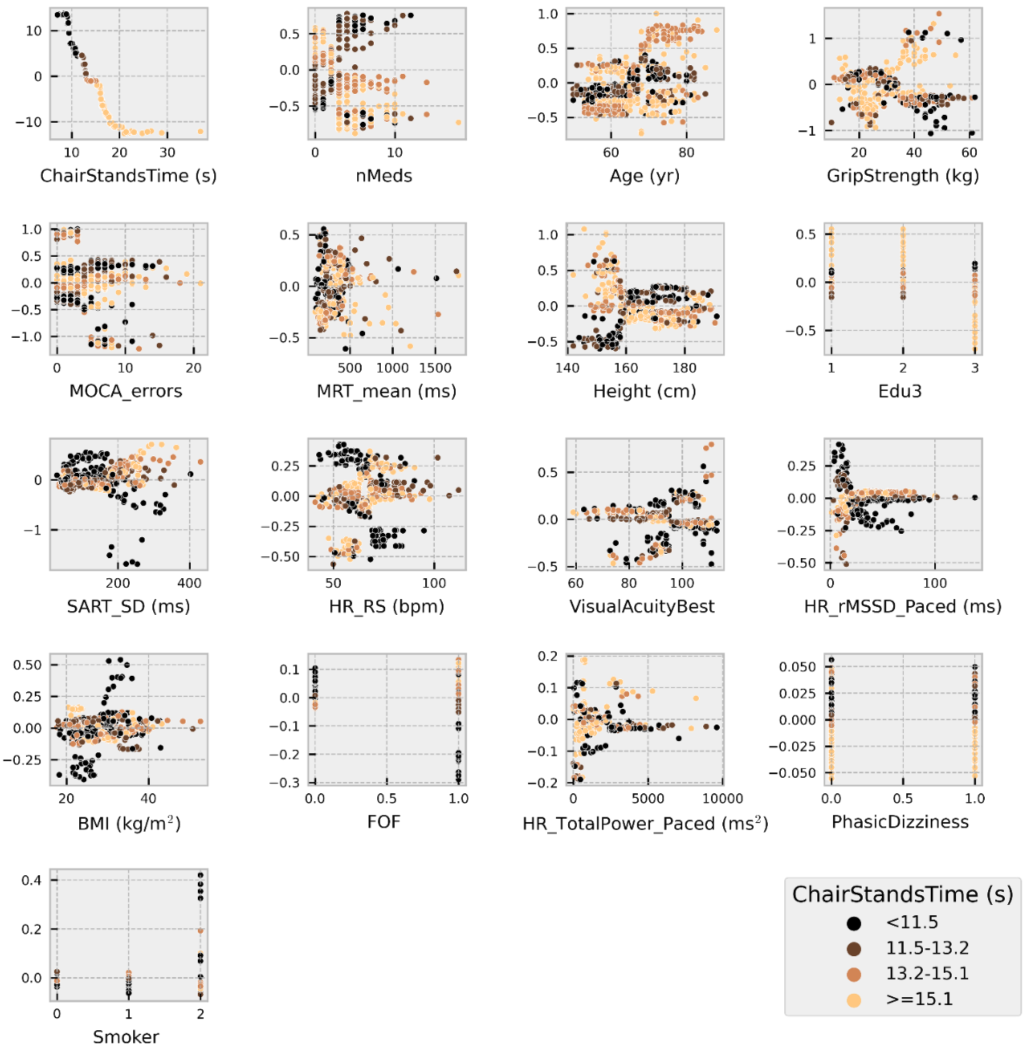
Appendix B Figure 5. SHAP values for the interaction between age and all features in the maximum speed model. The x-axes present the values of a feature in the units of that feature. The colour of a point indicates what quartile for age that sample falls in with black being lowest and light brown being highest. The y-axis displays the SHAP value of the interaction.

10.2.2 Grip Strength Interactions



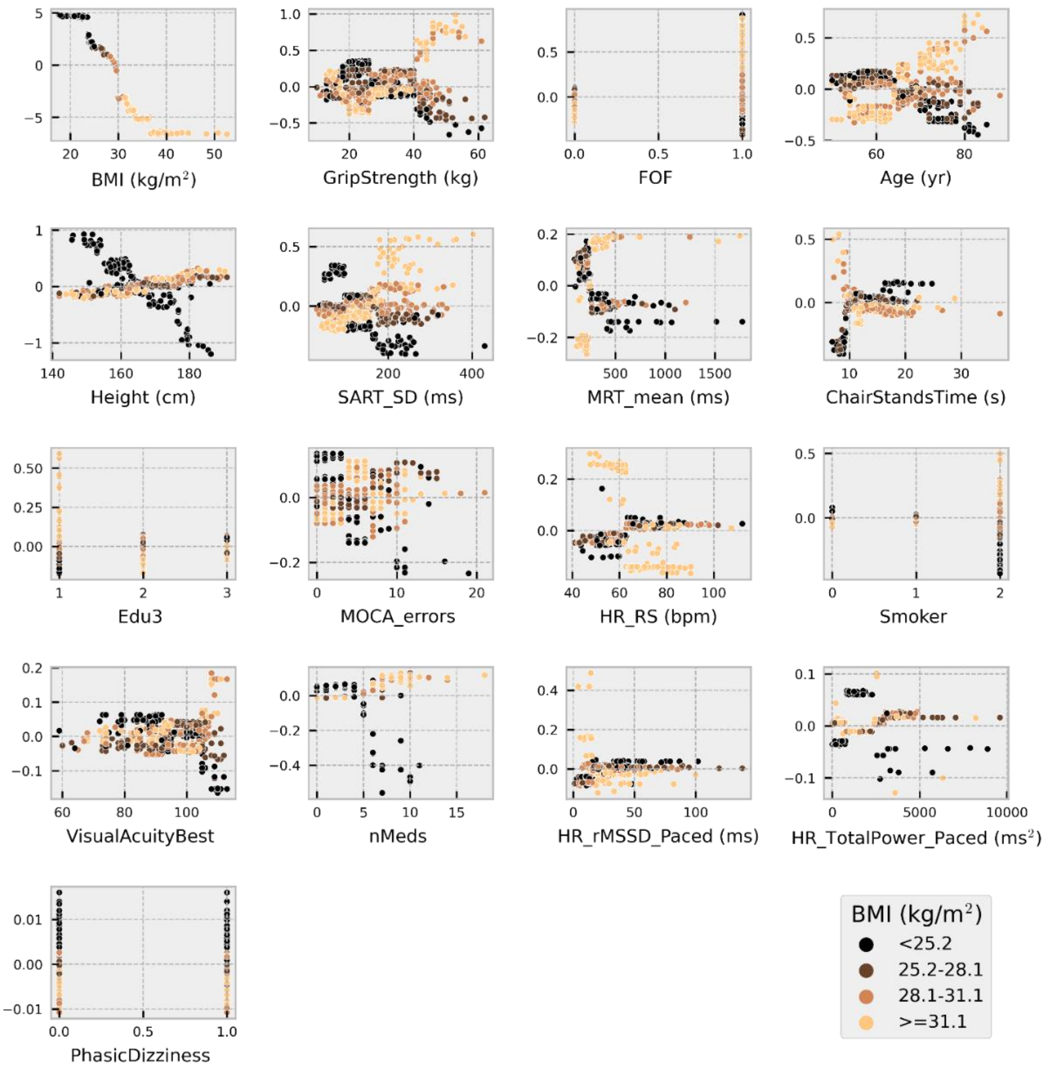
Appendix B Figure 6. SHAP values for the interaction between grip strength and all features in the maximum speed model. The x-axes present the values of a feature in the units of that feature. The colour of a point indicates what quartile for grip strength that sample falls in with black being lowest and light brown being highest. The y-axis displays the SHAP value of the interaction.

10.2.3 Chair Stands Time Interactions



Appendix B Figure 7. SHAP values for the interaction between chair stands time and all features in the maximum speed model. The x-axes present the values of a feature in the units of that feature. The colour of a point indicates what quartile for chair stands time that sample falls in with black being lowest and light brown being highest. The y-axis displays the SHAP value of the interaction.

10.2.4 BMI Interactions



Appendix B Figure 8. SHAP values for the interaction between BMI and all features in the maximum speed model. The x-axes present the values of a feature in the units of that feature. The colour of a point indicates what quartile for BMI that sample falls in with black being lowest and light brown being highest. The y-axis displays the SHAP value of the interaction.