FRAILTY IN OLDER ADULTS LIVING WITH HIV: DATA ANALYTICS AND MACHINE LEARNING TOWARD STANDARDISING A HEALTHY AGEING ASSESSMENT METHODOLOGY

Master of Science Dissertation

Mark R. Tyrrell, MSc Data Science 180590920

Supervisor: Dr. Paolo Missier

School of Computing
Newcastle University
United Kingdom
June 2019

Abstract

Activity levels in mid-adulthood are strong predictors of functional capacity later in life. Older Adults Living with HIV (OALWH) exhibit accelerated geriatric functional capacity profiles and are therefore a particularly vulnerable demographic to the effects of frailty onset. Early and frequent monitoring of patient functional capacity through clinical assessments of frailty can be used for interventions to improve outcomes. This analysis aimed to develop a method of replacing clinical assessments of frailty with patient-generated data analytics sourced from wearable activity trackers. The overall aim was underpinned by two primary research objectives: the use of unsupervised learning methods to identify homogeneous subgroups within a cohort of OALWH, as well as the deployment of interpretable machine learning to build a model for predicting frailty. The analysis found no strong subgroupings in the data, nor was the patient-generated data found capable of producing accurate predictions of frailty. Novelty in this research included the deployment of interpretable machine learning models to provide a predictive analysis of frailty.

1 Introduction

Along with advances in preventative medicine and falling fertility rates, life expectancy in countries throughout the world has advanced drastically since 1950. As a result, the elderly comprise an increasing proportion of the world's demographics [1]. The resulting services required to manage geriatric care present a major challenge to public health systems [2]. New paradigms and methodologies are required to meet this challenge.

The notion of **Healthy Ageing** as defined by the World Health Organization (WHO) is: *"the process of developing and maintaining the functional ability that enables wellbeing in older age"*[3]. Maintenance of functional capacity through the ageing process has been shown to be largely influenced by sustained overall activity levels in mid-adulthood [4]. In collaboration with HIV specialist Dr. Giovanni Guaraldi and the University of Modena and Reggio Emilia, the WHO is searching for healthy ageing tools to support health-care professionals and self-management in the routine care of the elderly. Such tools would utilise early warning metrics, assessing frailty and allowing preventative interventions earlier in the ageing trajectory; thereby supporting improved healthy ageing outcomes.

Traditional methods of monitoring functional capacity centre around complex indices comprised of multiple clinically assessed indicators [5]. This presents a challenge to preventive care due to the high cost and time involved with out-patient interactions. Towards this end, digital innovations including smartphone applications and wearable activity trackers offer the potential of data generation characterising individual functional capacity. Insights from these data could potentially be harvested using analytics and machine learning and exploited to replicate clinically assessed indices.

The ongoing *MySmart Age with HIV* (MySAwH) study [6] is designed to promote healthy ageing in older adults (age 50+) living with HIV (OALWH). Due to the effects of HIV, complications with anti-retroviral (ARV) treatment, and social factors common amongst people living with HIV, OALWH often exhibit diminished capacity at an earlier

chronological age than people without HIV [7]. As such, OALWH represent a particularly vulnerable population that could benefit from comprehensive healthcare involving regular quantification of frailty as a key evaluative component.

The MySAwH study involved 283 subjects based out of three clinics in Italy, Australia and Hong Kong. Each subject was equipped with a wearable activity tracker (Garmin VivoFit2) which provided daily monitoring of step count, calories consumed and sleep duration. In addition, the subjects were required to complete a monthly questionnaire administered via a smartphone app (MysAwHApp). For comparative purposes, a traditional clinical frailty index (FI) assessment was performed at baseline and every 9 months thereafter. The FI assessment utilised biomarkers and HIV-specific indicators collected by clinicians.

The data described above form the basis of the analysis outlined by this report. To avoid confusion between the analysis of these data and the longitudinal study which produced the data, the clinical study will be exclusively referred to as **MySAwH** throughout this document.

1.1 Overall Aim and Objectives

The overarching aim of the analysis was to use the MySAwH dataset to investigate and articulate a method of evaluating frailty using patient-generated data. Current methods of formulating frailty index (FI) rely on clinical visits to extract an array of biomarkers, which is a lengthy and costly procedure. If a model could be demonstrated to provide consistent and accurate predictions of FI given cheap and accessible patient-generated data, it would greatly advance the significance of FI as the basis for diagnostic and prognostic assessments in the context of managing healthy ageing in OALWH. Furthermore, such a model would have natural extensions toward healthy ageing management in geriatric care.

In seeking to achieve the aim of the analysis, a methodology was followed which explored key contributory research objectives. The aims and objectives are articulated below.

Aim:
Develop a methods of replacing clinical assessments of frailty with patient-generated data analytics

Objectives:

1. Analyse the MySAwH activity data using unsupervised learning methods to identify homogeneous subgroups within the cohort.
2. Model the MySAwH activity data employing interpretable supervised learning methods in the prediction of FI, assuming FI is a robust ground truth.

2 Background

2.1 Literature Review

The analysis drew upon existing literature spanning three key areas: 1) the domain - i.e. the clinical study of healthy ageing; 2) the methodology of time series modelling and clustering; and 3) the prediction of FI/ICI using various interpretable machine learning schemes. A review of the respective areas is summarised in the following sections.

2.1.1 The Healthy Ageing Paradigm and its application to OALWH

The healthy ageing approach to geriatric health care centres around prolonging functional capacity through the ageing process by focusing on early interventions [3]. Peeters 2013 showed that physical functioning levels in mid-adulthood are strong predictors of functional capacity later in life [4]. Figure 1 demonstrates this dynamic via the growth in the spread of physical functioning score quantiles with increases in age [4]. The disability threshold indicates a diminished functional capacity level requiring assisted living. Individuals achieving levels close to the maximum physical functioning score could expect approximately 10 additional years of functional independence compared with their peers at the median level.

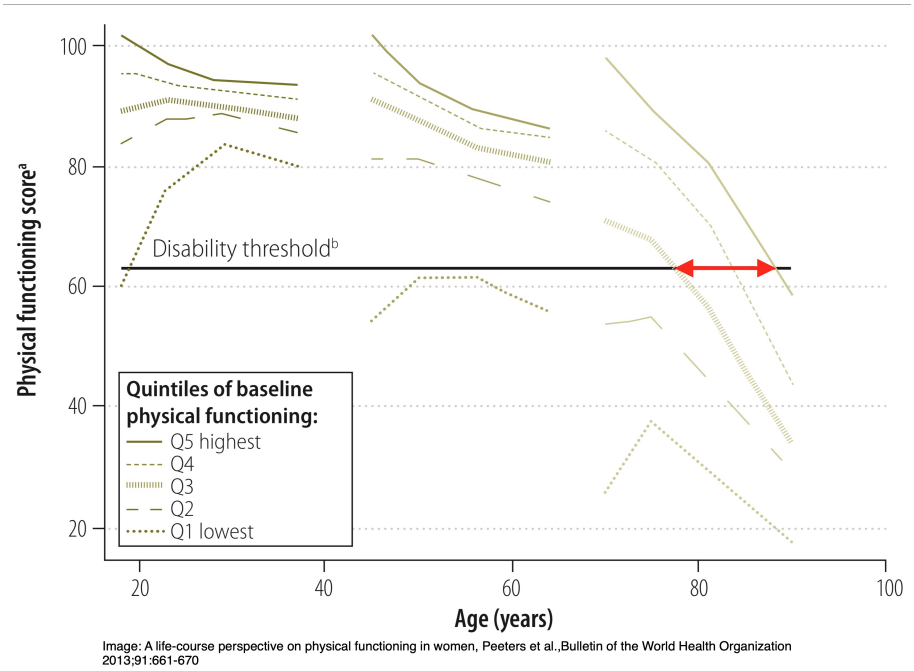


Figure 1: Functional Capacity over the Life Course [4]

Research on the study of healthy ageing in the older adults living with HIV (OALWH) demographic focused on outputs linked to Guaraldi and his collaborators [8, 9, 10, 7]. These studies centre primarily on indices characterising functional capacity. In particular, these studies closely examine Frailty Index (FI); an index which, as it quantifies a general degree of physical impairment, articulates the opposite of functional capacity. FI

is computed using various input variable schemes. Guaraldi et al. 2015, utilises 37 non-HIV clinical variables to construct an FI which was effective at predicting survival and incident multi-morbidity, demonstrating the importance of FI in the clinical management of OALWH [10]. Guaraldi et al. 2018 employs a wider range of variables to compute FI, with 72 indicators including biometric, psychiatric, blood tests, daily life activities, geriatric syndromes and nutrition data [8]. This study also looked at an additional 20 indicators to construct an HIV Index, (HIVI), and Protective Index (PI); both of which supplemented FI. Additionally, Brothers and Rockwood 2018 review the application of FI to OALWH concluding that a standardised method of evaluating frailty amongst people living with HIV has not yet been established [7].

The work of Dr. Matteo Cesari on intrinsic capacity was also reviewed [3, 11, 12]. Dr. Cesari is a Medical Doctor and specialist in Geriatrics and Gerontology at the University of Milan (UNIMI), as well as the MySAwH study's link to the WHO. His work seeks to establish intrinsic capacity as part of the healthy ageing approach to geriatric healthcare. Intrinsic capacity could be established as part of a comprehensive healthcare approach to healthy ageing. Unification of previous attempts to assess frailty in OALWH and the broader population would align well with intrinsic capacity as it assesses similar indicators.

2.1.2 Identification of Group Similarity using Time Series Clustering

The use of activity trackers in the MySWH study provided an opportunity to search for similar patterns of behaviour amongst the subjects. This was effected through the clustering of the time series data. Literature reviewed under this category focused on time series

Time series clustering is a complex operation. As with clustering of static variables, time series clustering involves identifying groups, then minimising intra-group similarity (cohesion), while maximising the inter-group similarity (separation), However, by nature time series exhibit high-dimensionality. Therefore the resulting computational cost of clustering time series is often higher than for static variables. Compared to static variables, time series complexity increases with each additional point in the series. This results in a distance evaluation operation defined by $O(mn)$ computation time, where m is the length of series A and n is the length of series B [13].

Furthermore, methods for determining distance based on a series shape are not straightforward. Humans are capable of intuitively categorising shape similarity, but this operation is algorithmically complex. For instance, Euclidian distance measures simply compute the distance between 2 series at time t; ignoring phase shifted correlative effects. This property of Euclidian distance naturally works well for lock-step (i.e. synchronised) time series, but not for time variate pattern matching.

Dynamic Time Warping (DTW) is significantly better at identifying patterns between similar series [14][15]. In DTW, the Euclidian distance is computed for every point in series A to every point in series B, resulting in a 2 dimensional distance matrix. The DTW distance between the series is then computed from the shortest path through the distance

matrix in terms of aggregate distance. The computational expense of this operation is kept manageable by a user-selected window of size n^2, where n is the number of time series points offset from the matrix diagonal [16].

Aside from computational cost, DTW can sometimes produce poor clustering behaviour. Variation in amplitude can cause the algorithm to respond by warping the time axis unnecessarily, with a single point mapped onto extended sections of the second time series [17]. This results in unintuitive clustering alignments. Additionally, prominent features in one time series can cause the algorithm to miss natural alignments. This can be caused by clustering excessively misaligned intervals. Therefore, careful pre-selection of intervals prior to clustering would appear to be a logical method of improving the algorithmic performance.

Shaped-Based Distance (SBD) offers a computationally faster alternative distance measure to DTW with similar performance [18]. The algorithm relies on the cross-correlation and coefficient normalisation (NCCc) sequence computed between 2 time series. This sequence is computed by convolving the time series to compare different alignments without using warping.

Objective judgements of cluster assignment quality can be effected using cluster validation indices (CVIs). Cluster quality can generally be assessed by a measure of the intra-cluster (i.e. cohesion) and inter-cluster (i.e separation) distances. CVIs exploit the cohesion and separation measures using various methods to produce an index value indicating the overall quality of the cluster operation. Arbelaitz et al. 2013 reviewed an extensive list of CVIs concluding that Silhouette validation was one of the best performing out of 30 indices experimentally compared [19]. As time series clustering algorithms output an aggregate distance value between each series, CVIs designed for static variables are compatible.

Visualisation by t-distributed stochastic neighbour embedding (t-SNE) provides a convenient method of presenting the results of multi-dimensional clustering in 2 dimensions. t-SNE works by preserving pairwise distances in low dimensional space [20]. The algorithm can provide a useful method of visualising multi-dimensional clustering effects beyond validation indices.

2.1.3 Prediction of FI/ICI using Interpretable Machine Learning

The benefits of highly accurate machine learning algorithms such as deep neural networks are often tempered by the inability to explain why the algorithm makes its classification decisions. This lack of intelligibility is particularly problematic in the medical sector, where incorrect classification can be a matter of life and death [21]. Generalised additive models (GAMs) provide high intelligibility in that each term in the model is summative with respect to the response variable, and therefore the effect of each term on the response variable can be analysed separately. However, GAMs often exhibit diminished predictive accuracy compared to more advanced models. Adding pairwise interactions of variables to a GAM, known as a GA2M model, increases accuracy while maintaining strong intel-

ligibility. The GAM portion of the model can be assessed separately from the interaction components, each of which are 2 dimensional and can therefore be easily visualised with respect to the response variable using a heatmap [21].

Other interesting work in the field of intepretable machine learning involves the use of Shapley values to provide attribution of predictors in the predicted model output. This approach utilises game theory fundamentals by measuring the contribution of a given predictor to model output. The algorithm accomplishes this by taking the difference between the model output with the predictor and without it. This task is then repeated for a stochastically determined subset of the predictors to capture interactive components. The resulting differences are then averaged to produce a Shapley value for the predictor which approximates its contribution to model output [22]. The Shapley Additive Explanations (SHAP) package for Python provides the implementation of the Shapley values approach and works well with many common black box machine learning models [23].

2.2 Data Understanding

Data used for this analysis were sourced from the MySAwH study. The data contain a maximum of 30 months of observations on 283 subjects based out of the three clinics in Modena, Sydney and Hong Kong. The primary focus of this analysis involves data extracted from Garmin Vivofit2 activity trackers quantifying subjects' daily step count, calories consumed and sleep activity. Additionally, the predictive modelling analysis also makes use of ecological momentary assessment (EMA) data characterising emotional state, habits, functional abilities and quality of life. These data were collected periodically through a questionnaire administered via the MySAwHApp app. For comparative purposes, the predictive modelling analysis also used the subjects biomarker data. These data form the basis of the frailty index (FI) computation and are comprised of various clinical indicators.

The study commenced in 2016, but did not exhibit uniform rate and volume of take-up at each location. Figure 2 displays the aggregate quantity of activity data observations recorded for each subject's relative study month, by clinic. Take-up was led by Modena, followed by Sydney, then Hong Kong considerably later. As subjects joined the study at different times, the volume of observations grew over time. With only 156 subjects represented by 18 or more complete months of observations, the gradual take-up limited the data available for modelling and longitudinal insights.

The distribution of observations per subject is demonstrated by the box plot in Figure 3. The maturity of the Modena cohort is clearly apparent, with a higher median (586) and tighter interquartile range than Sydney and Hong Kong.

The activity data were assumed to be of generally high quality in this analysis. The activity trackers were assumed to produce reliable approximations of step count, calorie consumption and sleep time. Furthermore, the subjects were assumed to be using the trackers as per instructions; i.e. maintaining an adequate battery charge and wearing the trackers diligently throughout the study period. However, as might be expected the data do reflect limited violations of these assumptions. The sleep tracking capability of the

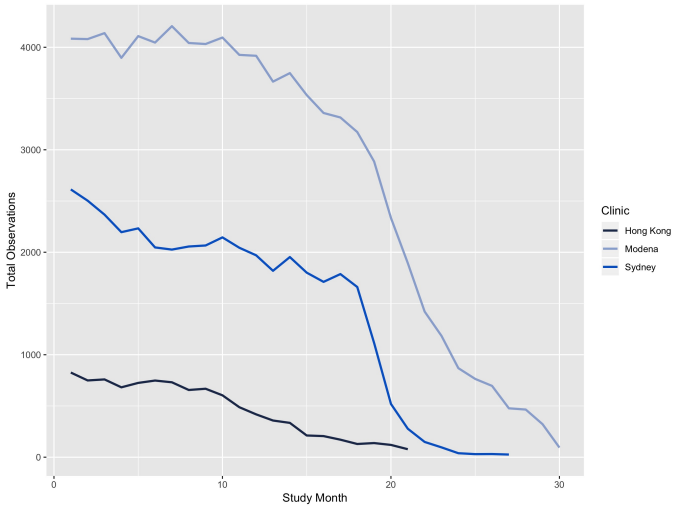


Figure 2: Activity Total Observations by Study Month

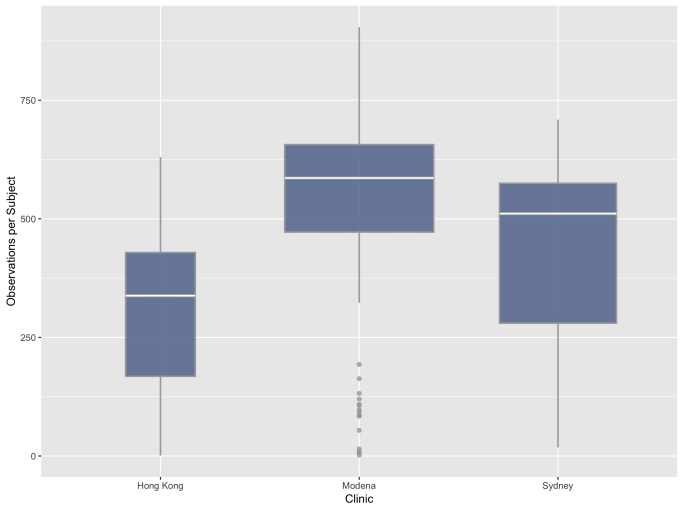


Figure 3: Activity Distribution of Observations per Subject by Clinic (n = 33, 151, 93)

devices appears suspect in that there were multiple instances of 0 values recorded for daily sleep duration. In such cases, sleep duration was imputed by local mean where possible, and global mean where a local mean was not possible to compute. Additionally, daily activity records were occasionally missing for certain subjects.

The EMA data were collected through the MySAwHApp smartphone app by querying subjects on their eating, sleeping, smoking and social habits, as well current mood, stress levels and neurological function. The EMA data also included the international physical activity questionnaire (IPAQ) to gauge self-perception of activity levels. The data were ostensibly collected monthly. In practice, the collection was not always complete. Figure 4 displays the quantity of observations by questionnaire category and clinic. Again, Modena leads in terms of quantity of observations. There is notable variation in

the total observations between the various categories. The variation is likely due to the breadth of the questionnaire. At 50 questions, it is quite possible that subjects became disengaged during response and either failed to complete the survey or gave quick answers.

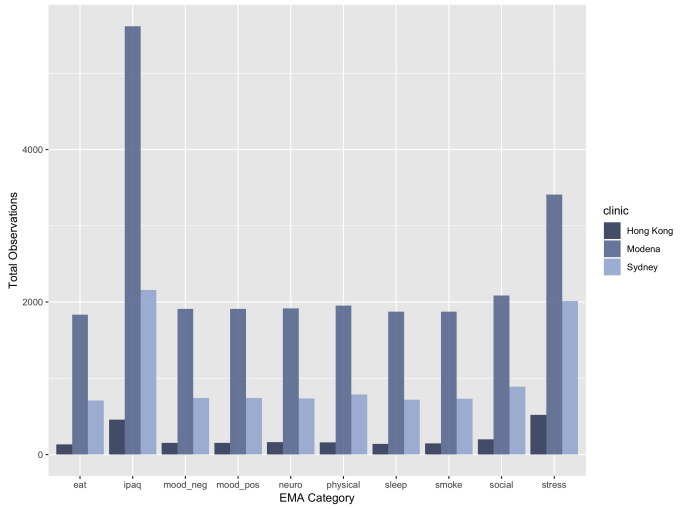


Figure 4: EMA Total Observations by Category and Clinic

The biomarker data were collected during clinical assessments at approximately 9-month intervals throughout the study. The data were primarily comprised of clinical indicators across six categories as displayed in Table 1. Body composition measured subjects' physical attributes with a focus on body mass index. Blood samples looked at common clinical indicators such as HDL and LDL cholestoral levels. Co-morbidity indicators characterised the prevalence of chronic disease in the subjects. HIV-related variables included additional bloodwork as well as reflecting subjects' anti-retroviral (ARV) treatment history. Polypharmacy quantified the number of prescription medicines taken by the subjects, excluding ARV treatements. Additionally, patients were asked to give their employment status.

The purpose of the biomarker data was for clinical computation of FI. Thus, each observation in the biomarkers dataset was accompanied by a corresponding assessed FI value based on an established formula that takes threshold biomarker data as input. These FI values constitute the ground truth for the predictive modelling of FI in this analysis.

2.3 Data Preparation

Prior to analysis, the data were checked for missing values and anomalies. In some cases, records were omitted from analyses due to insufficient or corrupted data. As detailed in Section 2.2, where possible and practical, data were imputed to replace missing values.

A major component of preparation involved synchronising activity records by study month and day of the week. As the subjects commenced the programme on different dates, time series cluster analyses was dependant on the relative study month, rather than

Table 1: Clinically Assessed Biomarker Data

Category	**Indicator**
Body composition	Lipoatrophy Multicenter AIDS Cohort Study (MACS) criteria > 1
Body composition	High or low BMI <18 or >25 kg/m2
Body composition	High waist circumference If female: >88 cm, If male: >102 cm
Bloodwork	High total cholesterol >200 mg/dl
Bloodwork	High low-density lipoprotein >100 mg/dl
Bloodwork	Low high-density lipoprotein <40 mg/dl
Bloodwork	High triglycerides >150 mg/dl
Bloodwork	Abnormal white blood cell counts <4000 cells/microlitri
Bloodwork	Anemia (female: <10gr/dl; male < 12 gr/dl)
Bloodwork	Hepatitis C coinfection Positive
Bloodwork	Hepatitis B coinfection Hepatitis B antigen positive
Bloodwork	Vitamin D insufficiency <30 ng/ml
Bloodwork	Abnormal parathyroid hormone >60 pg/ml
Bloodwork	Elevated D-dimer >Sample mean (358)
Bloodwork	Elevated C-reactive protein >0.7 mg/l
Bloodwork	Hyponatremia <125 mmol/l
Bloodwork	Proteinuria or albuminuria >5 mg/dL
Bloodwork	Elevated aspartate transaminase >31 U/l
Bloodwork	Elevated alanine transaminase >31 U/l
Bloodwork	Abnormal alkaline phosphatase <38 or >126 U/l
Bloodwork	Elevated g-glutamyl transphosphatase >55 U/l
Bloodwork	Low platelets <150 billion/l
Bloodwork	Abnormal potassium <3.5 or >5.3 mEq/l
Bloodwork	Abnormal phosphorus <2.5 or >5.1 mg/dl
Bloodwork	Abnormal thyroid-stimulating hormone <0.27 or >4.2 mIU/l
Bloodwork	Elevated total bilirubin >1.10 mg/dl
Co-morbidities	Cardiovascular disease (clinical diagnosis)
Co-morbidities	Hypertension (blood pressure or treatment)
Co-morbidities	Diabetes type II (fasting glucose > 125md/dl or treatment)
Co-morbidities	CKD (two estimated egfr < 60ml /min/)
Co-morbidities	Cirrohosis (FIB 4 score > 3.25)
Co-morbidities	COPD (FEV1/FVC < 0.7)
Co-morbidities	Osteoporosis (DXA z score < -2.5)
Co-morbidities	Any Cancer (clinical diagnosis with biopsy confirmation)
HIV related variables	Current CD4 cell count (<500 cells/mm^3)
HIV related variables	Nadir CD4 cell count (<200 cells/mm^3)
HIV related variables	HIV VL (>40 copies/ml)
HIV related variables	CD4/CD8 cell ratio (<1.0)
HIV related variables	Duration HIV infection (>10 years)
HIV related variables	Pre HAART start (antiretroviral therapy started before 1/1/1997)
HIV related variables	ART failure (history VL > 1000 copies/ml while on antiretr therapy)
Polypharmacy	Polypharmacy (> 5 drug classes excluding antiret therapy)
Patient related outcomes	Unemployment Self-report

the real date. Furthermore, subjects did not all start recording activity data on the same day of the week. As weekly seasonality was a key point of exploration, the data were synchronised to commence on the first Monday of study month 1.

The EMA questionnaire data largely consisted of Likert scale responses (e.g. *on a scale of 1 to 10 how would you rate your current stress level?*) as well as some binary questions. As Likert scale questions are ordinal categorical variables with an assumed degree of equality between the scale intervals, these data were simply converted to numerical values during the data preparation phase. While there is debate about the validity of assuming equal intervals in Likert scales when using the resulting data for parametric statistics, possible errors introduced are generally very small [24].

In less clear cases with the EMA data, guidance was taken from the MySAwH study which had previously analysed these data for different indices. For instance in the Physical Activity section, when the subjects were asked which activity they had been doing immediately prior to responding, the possible responses were: Physical Activity/Exercise, Reading, Computer, Watching TV, Eating/Drinking, Cooking/household chores, Childcare or taking care of another person, Other. The study had treated this question as a binary response with "Physical Activity/Exercise" qualifying as a positive response, and all others negative. As there was no clear further meaning to be derived from the other responses, this analysis treated the question identically, assigning a binary 1 to "Physical Activity/Exercise" responses.

2.3.1 Time Series Missing Data

The synchronisation operation on the time series involved assigning a study 'day' value to each entry in the time series starting from the first Monday of study month 1. In conducting this operation, a large amount of 'hidden' missing data was discovered. Whereas the time series had already been checked for NAs and suspected erroneous 0 values, it was not obvious that there would also be date records completely omitted from the series. However, this was the case with large consecutive sections of data missing from the majority of the time series.

Figure 5 provides a visualisation of the extent of the missing time series data using the *missingno* package for Python. The plot is best viewed as a wide dataframe representation with the time series for each subject as a column and the first row of the series at the top; black markers indicating data, and white markers indicating the missing data. The line on the right side of the plot is a histogram representing the distribution of missing data.

The extent of the missing data demonstrated by Figure 5 was considerable and presented a serious challenge to time series cluster analyses. Various time series imputation methods were explored in order to salvage as many observations as possible. Kalman smoothing has been shown to compare favourably with other methods of interpolation of univariate time series, and indeed produced the best results this analysis [25]. However, the effect was limited to shorter spans of sequential missing data. Figure 6 demonstrates this dynamic, with the red lines indicating imputed data. The longer imputed span in 6b is an intuitively very poor approximation of the missing sections of the time series, based on previous patterns apparent in the data. Conversely, the short 6-point imputed span displayed in Figure 6a provides a decent estimate of the missing data, with a seasonality element and mean close to what could be intuited from the surrounding series.

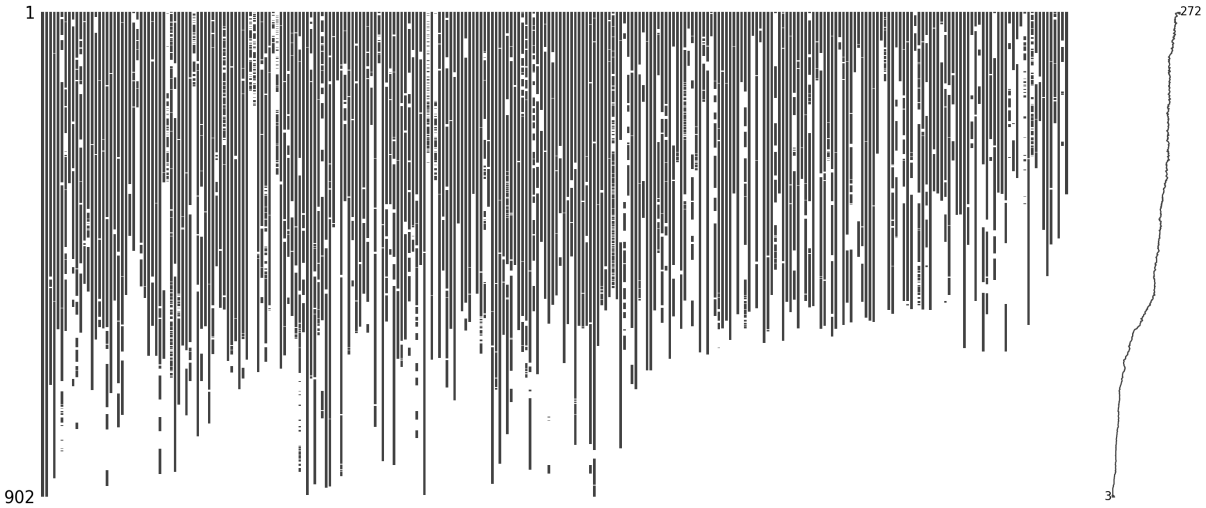
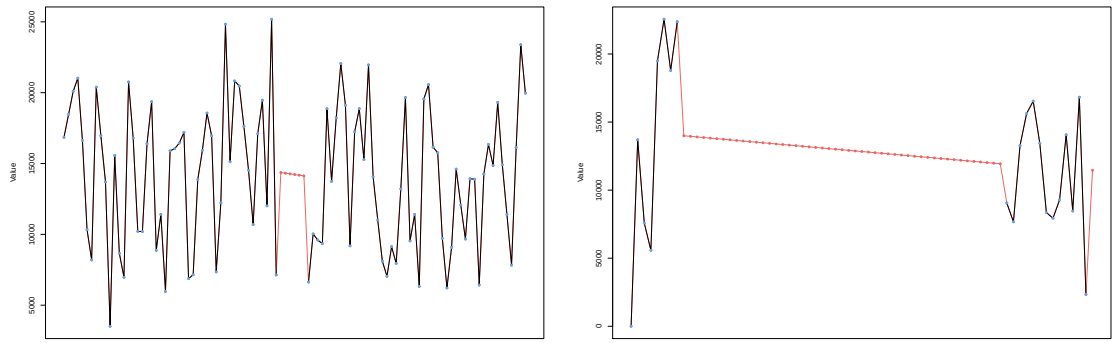


Figure 5: Activity Time Series Missing Data



(a) 6-Point Imputation

(b) 49-Point Imputation

Figure 6: Time Series Imputation: Subject ID 120 (Kalman Smoothing)

The poor imputation performance over long spans would have been detrimental to the clustering of raw time series as the algorithms would tend to erroneously see the imputed regions as key features of similarity. This therefore necessitated discarding any time series with greater than 6 sequential instances of missing entries. As Figure 5 indicates, this action resulted in a large reduction in available data with 73,120 of 130,632 obs representing 122 of the 283 subjects. However, as detailed in Section 4.1.2, raw time series was only one of numerous approaches used to cluster time series. The other methods all used some degree of aggregation and were therefore less affected by low-fidelity approximations of the missing data.

2.3.2 EMA and Biomarkers Missing Data

The EMA and biomarkers datasets also exhibited large quantities of missing data, though they were considerably more intact than the activity data. Figure 7 represents the EMA dataset, with the white markers indicating missing data. As demonstrated, the missing data were sparse and appeared to be structured loosely by questionnaire section. This would indicate that some subjects skipped sections entirely. The EMA missing data were imputed for each variable by local mean based on subject id where possible, and by the global mean otherwise.

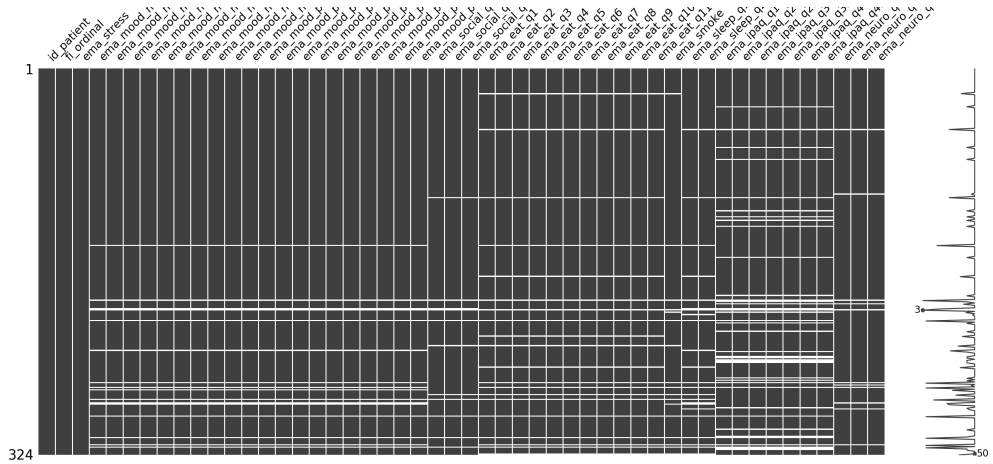


Figure 7: EMA Missing Data

Missing data in the biomarkers dataset is represented by Figure 8. This dataset exhibited consistent missing entries on seven variables: Vitamin D, PTH, Na, UProtein, AST, K and P. As with the EMA dataset, the biomarkers missing data were imputed for each variable by local mean based on subject id where possible, and by the global mean otherwise.

Cleaning the biomarker data presented a major challenge in that the source file (*FileAllData.xlsx*) contained numerous instances of missing FI assessments and dates. All observations containing missing entries for the ground truth, FI were removed as there was to alternative. However, the resulting dataset of 571 observations still contained 186 observations with missing dates. The dates of each FI assessment provided a qualifier window for segmenting the activity and EMA to the corresponding ground truth (FI), therefore the absence of these dates would have resulted in severely reduced data for modelling. As the data were already limited in size, this was particularly problematic.

The solution involved a search for additional data sources. Clinical assessment dates corresponding to the *FileAllData.xlsx* dates were found in the *score.data.xlsx* file. These data did not include an FI score variable, however the dates were usable when matched with the original working file. After joining the data, 41 observations remained with

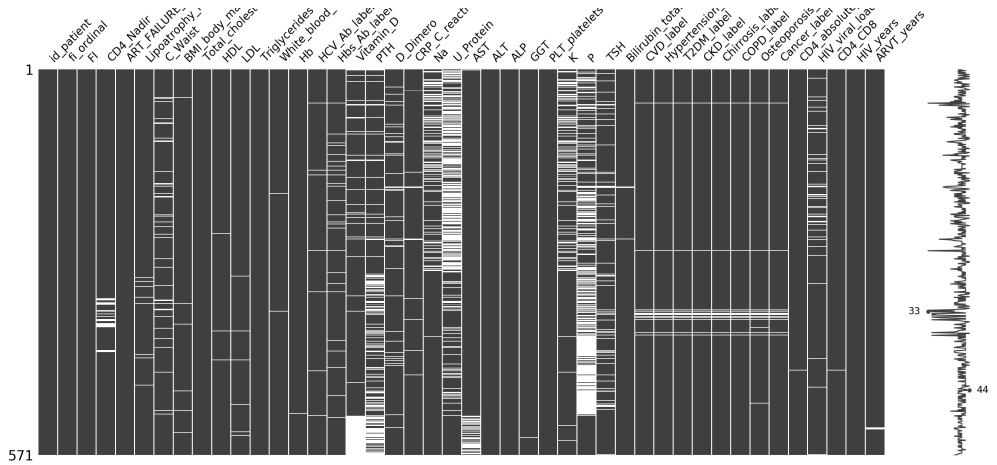


Figure 8: Biomarkers Missing Data

missing dates. These missing dates were then imputed by searching for non-missing dates for the same subject id under a different FI assessment interval. If another date was available, the 9 or 18 months difference were used to impute a date for the missing entry by referring to the corresponding interval (eg. Baseline, 9-Months, 18-Months). For instance, a subject with a Month 9 FI assessment on 01/01/2017 and a missing date for their Month 18 FI assessment would be assigned a Month 18 FI assessment date of 01/09/2017. This processing reduced the total number of missing FI assessment dates to 11, resulting in 560 usable observations.

2.4 Reproducibility

The Analysis is composed of multiple scripts written in R, and Python. The R scripts comprise approximately the first half of the work. These scripts are focused on exploring, cleaning, wrangling the data, in addition to time series clustering using the *dtwclust* package. The Python scripts comprise the second half, and are primarily concentrated on pre-modelling feature set processing, including correlation analysis and missing data imputation. The actual modelling was conducted in Jupyter Notebooks due to limitations in the Explainable Boosting Machines (EBM) *interpret* and the XGBoost interpretability plugin Shapley Additive Explanations *SHAP* package.

There are minor exceptions to this schema. Exploration of missing data in the interim dataframe output by the `garmin_win (TS_Cluster_Preparations.R)` script was examined in Jupyter Notebook to allow firstly for the use of the excellent *missingna* Python package, and secondly for the better presentation and viewing capability of Jupyter.

Due to the complexity of the analysis and the need to work between R and Python, no streamlined reproducibility functionality was built-in. In order to reproduce the analysis, scripts must be run manually and in some cases sequentially. The output dataframes from computationally expensive processing have been cached as *.Rdata* files in the /cache folder

to allow for quick access. Additionally, many dataframes were exported as *.csv* files to the wrangled data folder to allow access by downstream Python scripts.

All code and data used to produce the analysis are provided in an archived (zip) folder using the standard R *Project Template* format. Full documentation and reproducibility instructions are included in the *README* files in the relevant folders.

3 Analytical Structure

As stated in Section 1.1, the analysis sought to develop methods of evaluating frailty using the patient-generated MySAwH data set. This aim was underpinned by two primary research objectives:

1. Analyse the MySAwH activity data using unsupervised learning methods to identify homogeneous subgroups within the cohort.

2. Model the MySAwH activity data employing interpretable supervised learning methods in the prediction of FI, assuming FI is a robust ground truth.

The first of these objectives focused on varied cluster analyses of the activity time series data, while the second focused on predictive modelling of that same data using various interpretable machine learning algorithms. These tasks were linearly complementary, with results of the first objective potentially informing inputs to the second. However, the extensive breadth of documenting the respective methodologies and findings necessitated a logical division in this paper. Each research objective was therefore assigned an individual section in the report detailing the respective methods and results.

4 Objective 1: Identification of Sub-groups

The activity data was explored using unsupervised learning in order to identify similar groupings amongst the subjects that could provide insight, as well as contributing to a feature set for predictive modelling of the frailty index. This effort focused on time series cluster analyses conducted using the various methods detailed in the following section, in turn followed by the results of the analyses.

4.1 Methods

Various methods were used to explore the activity data in order to identify homogeneous groupings amongst the subjects. This effort focused primarily on time series cluster analyses, as well as $k$means analysis of time series summary statistics and derivations of the raw step count, calories and sleep duration data.

4.1.1 Challenges

A major component of the time series clustering task involved deciding the width of the time domain used as an input to the clustering algorithm; i.e. whole series, subsections, summaries, seasonality or trends. While evaluating the use of long sections of the series, for instance trends or whole series, comparison was complicated because of the following factors:

1. Subjects live in vastly different climates with very different weather and out-of-sync seasons. For instance, midsummer in Sydney is midwinter in Modena, whereas Hong Kong experiences a subtropical climate year round. These climate variances naturally affect daily activity levels.

2. Subjects live in different types of cities. Hong Kong is vastly different than both Sydney and Modena in terms of urban landscape, public transport and culturally-informed habits. Densely populated cities with good public transit are more likely to encourage higher activity levels than cities with high levels of car ownership.

3. The initial study month (0) was not synced. Subjects started the study at multiple different times of year. Therefore even if point number 1 (different climates) was somehow accounted for, each individual time series would have also needed to be indexed for the time of year the subject started the study.

4. The data were limited for many subjects. Of the 283 subjects in the study, 25 percent of the time series contained less than 365 days of activity observations.

With these complexities in mind, several different approaches were employed towards cluster analysis of the activity data, with varying results. These included clustering by raw time series, seasonality, trend, weekly and monthly summarisations, and summary statistics.

4.1.2 Clustering Approaches

Clustering of the activity time series observations was initially explored by clustering the raw data without any type of summarisation processing. As summarisation reduces time series dimensionality and noise, analysis of the raw daily time series data was expected to reveal only weak similarities between subjects.

Clustering of the data was next explored by decomposition; extracting the seasonality, trend and residual error from the time series. Decomposition was accomplished using the *decompose* function in the R *stats* package. The decomposition function first computes the overall trend for a given series using a moving average, which is then negated from the time series. Seasonality is then computed by averaging each point within a specified period, for each occurrence of the period in the time series. The error is then obtained by negating the trend and seasonality components from the original time series. The decomposition algorithm works optimally when the time series covers an integer number of complete periods. Therefore, prior to clustering, the activity data were processed to ensure that each series was a multiple of a 7-day weekly base frequency.

The averaging function of seasonality over a recurring time interval was expected to show similarities in the subjects' activity patterns. The data were first analysed by monthly seasonality over a 28-day period. The time series were further examined by clustering on weekly seasonality. Ideally the activity data would have been comprised of a highly granular series (eg. hourly readings) from which to extract a meaningful 7-day seasonality. This would have provided an intuitive summarisation of individual activity characteristics over a salient period. However, the data contain only seven readings per week, minimising the possibility of deriving strong findings from this clustering approach.

Decomposition by trend allowed more flexibility than seasonality as long or short intervals could be extracted for separate cluster analysis. The data were clustered by trend over quarterly and yearly intervals, as well as for each of study month 1 through 12 in isolation. A meta analysis was then conducted on the resulting monthly cluster assignments to ascertain the strength of activity pattern similarities amongst individual subjects over the first year of the study. It was posited that if subjects followed similar identifiable activity patterns over the course of the selected interval, these patterns should manifest in a secondary cluster analysis.

The methodology for secondary clustering utilised cluster assignments over each month as input data. As the cluster assignments were nominal categorical variables, the standard partitional kmeans algorithm was avoided as cluster means for one-hot encoded input vectors are not necessarily reflective of true cluster characteristics [26]. Alternatively, the frequency-based kmodes clustering algorithm was employed, with the Cao method of cluster initialisation [27].

Further clustering analysis was conducted by reducing the dimensionality of the time series by weekly and monthly summarisations. Summary statistics were computed for each activity type and the resulting reduced times series were clustered. Prior to clustering, the

summarised time series were filtered to include only full samples for series of a meaningful length. The resulting summarised time series consisted of 52 intervals of 1-week and 12 intervals of 1-month.

4.1.3 Kmeans Clustering Time Series Summary Statistics

Kmeans clustering of the time series Activity data involved removing the temporal dimensions by aggregating by summary statistics, and adding derived variables for each time series over selected sub-intervals.

The interval selected for summarisation was informed by the requirements of the predictive modelling task. Intervals were assigned with respect to the period between successive FI assessments. Each subject had FI assessments performed approximately every 9 months over the course of the MySAwH study. The dates of each FI assessment were used as an interval for segmenting the activity and EMA data to the corresponding ground truth (FI). For instance, for a subject with a Month 9 FI assessment on 01/01/2017 and a Month 18 FI assessment on 01/09/2017, all Activity and EMA observations in the interval 01/01/2017 through 01/09/2017 were summarised and labelled with the Month 18 FI value. Therefore, feature construction was accomplished by summarising all activity data observations in the interval preceding a given FI assessment by mean, variance, minimum and maximum.

Prior to computing summary statistics, the data were further stratified into weekday and weekend periods for a given FI interval. It was assumed that stratification of this type might reveal differences in patterns between the working week and more leisurely weekend.

In addition to the summary statistics, an ordinal categorical variable was derived for each activity type over the same interval. The variables qualified activity levels according to commonly defined thresholds as presented in Table 2 [28]. Walking activity in minutes per day was computed from step mean based on an average adult step rate of 100 steps per minute [29]. The step mean for each time series interval was divided by 100, and categorised according to the daily walking minutes thresholds.

Table 2: Activity Thresholds

Walking (min/day)	**Total physical activity** (MET.minutes/week)	**Sleep Duration** (hours/night)
< 20 (Low)	< 919 (Low)	< 7 (Poor)
21 to 30	919 to 1902	$7 - 8$ (Good)
31 to 60	1903 to 3706	> 8 (Poor)
> 60 (High)	> 3706 (High)	

Total physical activity, in metabolic equivalent of task (MET) minutes per week, was computed from the mean of the daily total calories data. This involved extracting the subject's weight from multiple sources in the clinical data, and imputing missing values based on mean weights by sex for the OALWH sample. The daily $MET \cdot minutes$ were computed using the conversion detailed in Equation 1 [30][31], then multiplied by 7 to

compute the projected $MET \cdot minutes/week$.

$$MET \cdot minutes = \frac{200 \cdot Kcal}{3.5 \cdot Weight} \tag{1}$$

Sleep duration in hours was computed by dividing the total recorded daily sleep seconds by 3600. Prior to this computation, the sleep duration data were imputed to correct corrupted entries. In total, 14,608 observations in the original data were comprised of zero values, indicating prima facie that the subject had not slept during the interval. However, these were assumed to be erroneous entries resulting from improper use of the activity tracker by the subjects. The erroneous sleep data were therefore imputed using the mean sleep duration for each subject id. Where the mean for each subject id was insufficient, the global mean sleep duration was imputed instead.

4.1.4 Clustering Algorithms and Distance Measures

The analysis made use of two primary clustering algorithms. Agglomerative hierarchical clustering using complete linkage, and partitional clustering. Distance measures included dynamic time warping (DTW) and shape-based distance (SBD). This resulted in 4 algorithmic configurations applied to each clustering approach: hierarchical clustering with DTW, hierarchical clustering with SBD, partitional clustering with DTW, and partitional clustering with SBD.

As detailed in Section 2.3, prior to clustering the data were processed in order to synchronise all time series to begin on the first recorded Monday of study month 1. All time series clustering utilised the *tsclust* function provided by the R *dtwclust* package [16]. The comparison included partitional clustering and agglomerative hierarchical clustering algorithms, with various combinations of distance measures including dynamic time warping (DTW) and shape-based distance (SBD).

4.1.5 Cluster Validation and Visualisation

Objective judgements of cluster assignment quality were validated using the Silhouette score cluster validation index (CVI). This index has been shown to provide competitive performance when measured against other widely used CVIs [19]. In order to assess the optimal k value for each time series clustering approach, the clustering algorithm was applied for a range of $k = 2 : 20$ and the CVI computed for each value of k. This process was repeated for each of the 4 configurations of clustering algorithm and distance measure. The results were then represented in an elbow plot.

For the Kmeans clustering of time series summary statistics, t-SNE visualisations were used in addition to CVI in order to visually represent any apparent clustering in the multidimensional data in 2 dimensional space. This additionally allowed for stratifying the data by descriptor variables to identify any potential overlap with cluster assignments strongly apparent in the visualisations. The descriptors included subjects' clinic, sex, age, frailty index (FI) and clinical biomarkers. Such overlap would have provided strong insight into patient activity correlations.

4.2 Results

Clustering of the activity time series data using the various methods produced poor results. No pervasive, strong evidence of similar subgroups in the cohort was observed.

4.2.1 Clustering Raw Time Series

The elbow plot in Figure 9 displays the cluster validation using silhouette score over $k = 2 : 20$ for step count across multiple different intervals. Figure 9a and 9b document clustering validation over the course of a single month, while figure 9c and 9d apply to intervals of the first 3 and 12 months of the study respectively.

As demonstrated, the CVI for all intervals were uniformly poor. The results appear to vary with the length of the interval. The worst results were exhibited by the longest period of 12 months. The same effects were observed on clustering of calories and sleep duration activity data. Based on these results, it would appear that clustering by raw time series data is not a viable avenue for extracting meaningful inference on activity-based sub-groupings within the cohort.

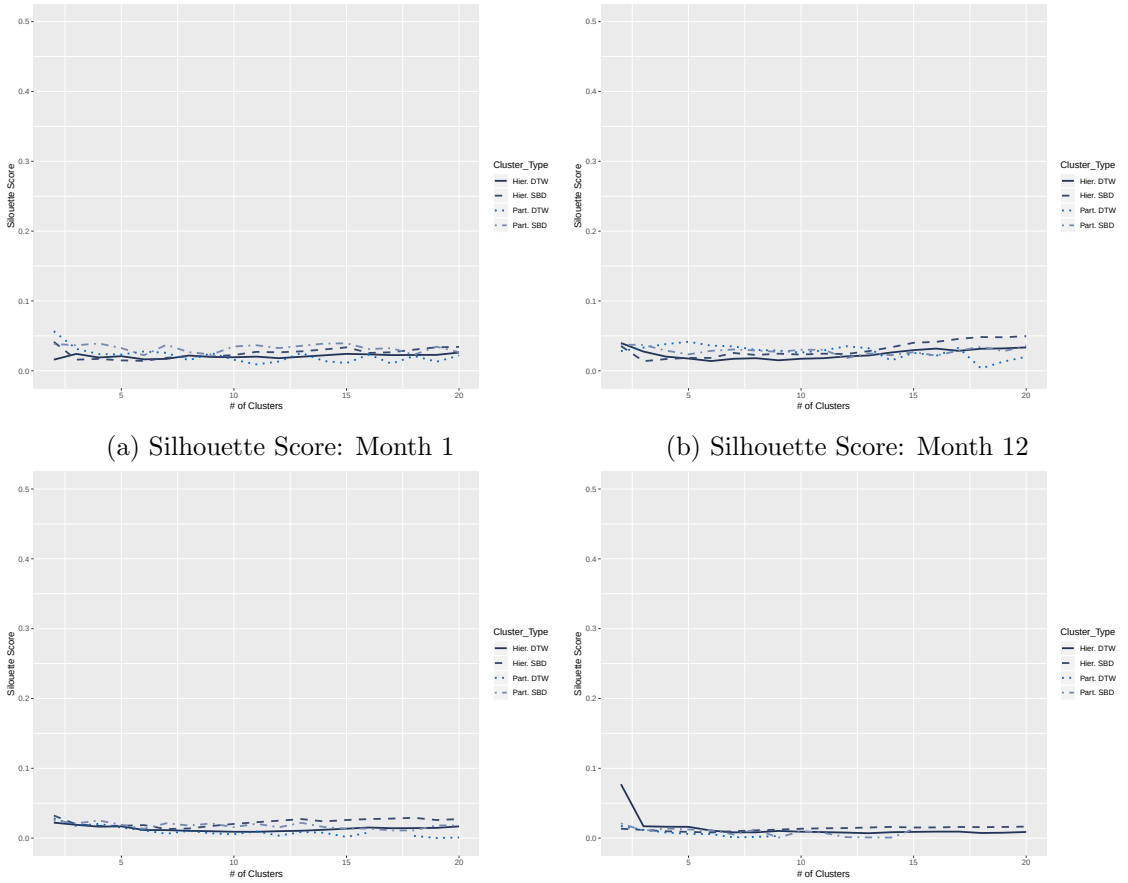
4.3 Clustering by Seasonality

The results of clustering on a 28 day monthly step count seasonality are demonstrated by Figure 10a. The plot contains the results of $k = 20$ hierarchical clustering using SBD distance on the activity step data. While there are certain clusters which appear to be more cohesive than others, there are also many clusters with only a few assignments. Moreover, some of the clusters appear very similar. It is therefore difficult to discern strong natural groupings in the time series, despite the high k value.

The monthly seasonality was further evaluated using CVI. Figure 10b displays the Silhouette score for 4 different clustering configurations using various clustering algorithms and distance measures. For all clustering configurations, the silhouette score reaches a maximum of 0.087 out of a possible 1 indicating very low intra-cluster cohesion and inter-cluster separation, and therefore a lack of distinct clusters apparent in the data.

The results of DTW clustering on weekly step count seasonality for $k = 4$ are displayed in Figure 11a. The plot appears to show more clearly identifiable groupings than the plot for monthly seasonality. Additionally, the validity of these clusters is notably higher than for monthly seasonality when evaluated using CVI. However, the Silhouette score for the best performing clustering configuration was still only 0.283 out of a possible 1 (ref. Figure 11b). Though considerably better than monthly seasonality clustering, this score indicates a lack of strong sub-groupings within the data.

The seasonality-based clustering analyses were repeated for the calories and sleep duration time series. Similar results were observed, with consistently higher Silhouette scores on the weekly seasonality clusters compared with monthly. The Silhouette scores for each of these analyses are provided in Table-3. None of the cluster configurations on either of



(a) Silhouette Score: Month 1

(b) Silhouette Score: Month 12

(c) Silhouette Score: Months 1:3

(d) Silhouette Score: Months 1:12

Figure 9: CVI for Raw Step Count Clustering over Various Intervals

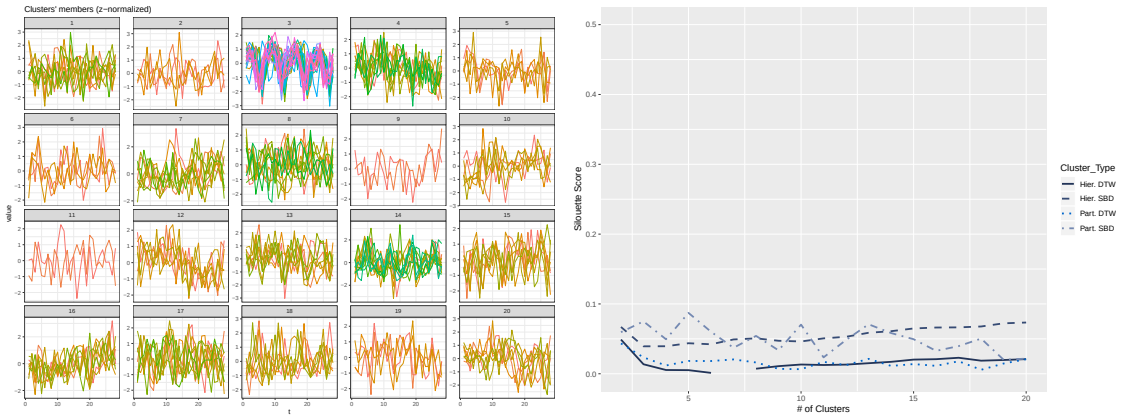
the two additional activity types produced a Silhouette score indicative of distinct clustering in the data.

Table 3: Maximum Silhouette Score by Seasonality for each Activity

Activity	**Interval**	**Max Score**	**k**	**Cluster Type**
Step	Week	0.283	2	Partitional SBD
Step	Month	0.087	5	Partitional SBD
Calories	Week	0.304	2	Partitional SBD
Calories	Month	0.080	2	Partitional SBD
Sleep	Week	0.232	17	Partitional SBD
Sleep	Month	0.043	2	Partitional DTW

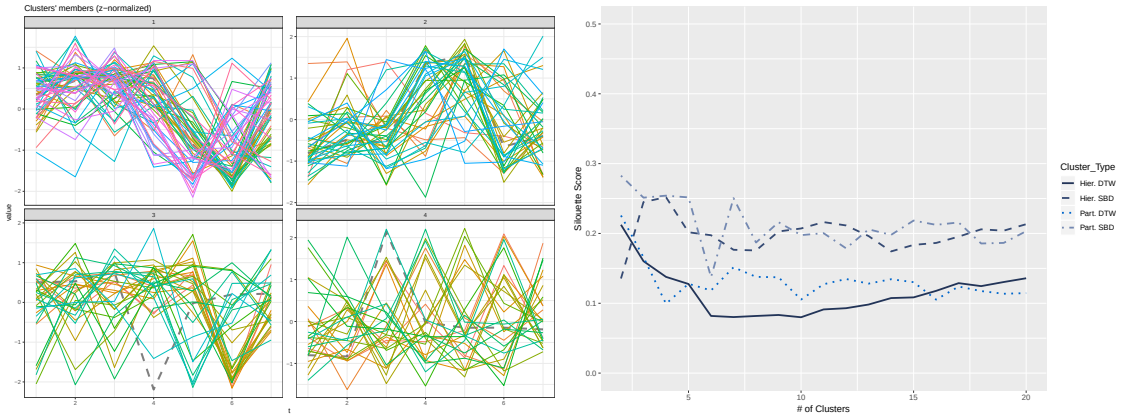
4.4 Clustering by Trend

Initial exploration of decomposition by trend over various long term periods returned generally low CVI scores. Figure 12 displays the results of CVI evaluation over $k = 2 : 20$



(a) Clustering Hier. SBD k = 20 (b) Silhouette Score for k = 2:20

Figure 10: Clustering Step Count by 28-Day Seasonality

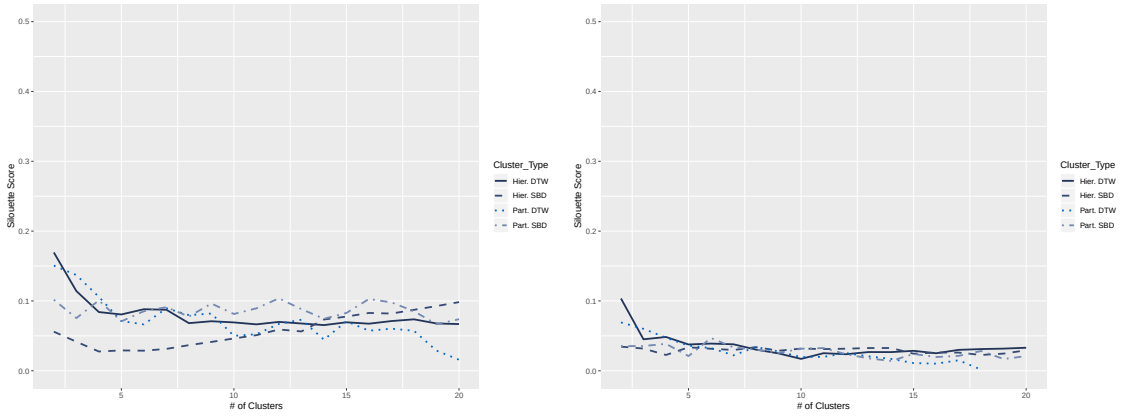


(a) Clustering Part. SBD k = 4 (b) Silhouette Score for k = 2:20

Figure 11: Clustering Step Count by 7-Day Seasonality

clusters for the first 3 months and 12 months (12 and 48 weeks respectively) of the study for 201 subjects. CVI was naturally higher for the shorter interval of 3 months with a maximum Silhouette score of 0.239 for Partitional SBD at $k = 3$ compared with 0.124 for the 12 months interval at $k = 2$. The actual 'months' used in this case refer to 28-day base periods in order to provide a repetitive interval of uniform length for the decomposition algorithm.

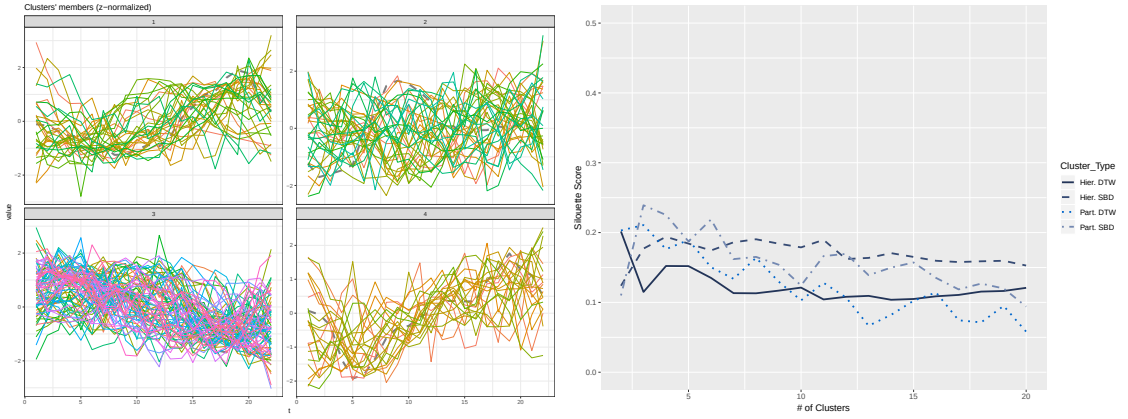
As observed in Section 4.2.1, clustering the time series over shorter intervals consistently yielded better cluster results. With this dynamic in mind, clustering by trend was also examined on a single monthly (28 day) basis. Figure 13a displays the results of clustering the trend over the first study month for $k = 4$. The k value in this case was determined by the relatively good Silhouette score for partitional SBD at $k = 4$ as displayed in Figure 13b. The time series plots show four moderately clear step count activity trends over the course of the month. Plot 1 appears to show low initial activity, followed



(a) 3 Month Trend (k = 2:20) (b) 12 Month Trend (k = 2:20)

Figure 12: Silhouette Score: Step Count by Trend

by a gradual upward trend. Plot 2 is very noisy, but could be described as a steady but mild upwards trend. Plot 3 shows a downwards trend, while plot 4 shows an initial drop in activity, followed by an upward trend. Table 4 provides a breakdown of the cluster assignment by quantity and proportion. Forty-seven percent of the subjects exhibited the uniform activity levels associated with cluster 3.



(a) Clustering Part. SBD: k = 4 (b) Silhouette Score k = 2:20

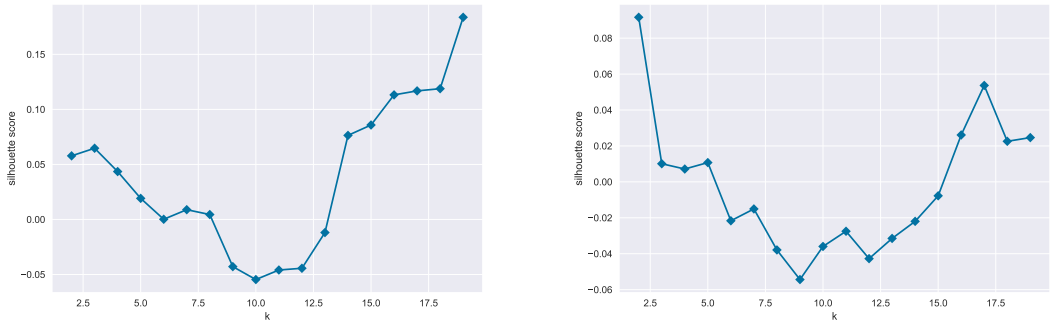
Figure 13: Clustering Step Count by Trend over Study Month 1

The trend over month-1 clustering was then repeated for all months 1 through 12. The cluster assignments for each of these months were then analysed to reveal consistency in assignment through each month using *k*modes clustering. Figure 14 displays the results of the *k*modes cluster analysis for step count trend over study months 1-3 and 10-12. As demonstrated by the silhouette plots, there is no apparent consistent clustering pattern in the cluster assignments over extended periods. The same effects were observed for months 4-6 and 7-9, as well as for longer periods. The longer periods exhibited even weaker clustering. As with seasonality-based clustering, clustering by trend over long periods does

Table 4: Part. SBD Cluster Assignments: Trend over Study Month 1

Cluster	Quantity	Proportion
1	30	0.19
2	35	0.22
3	75	0.47
4	21	0.13

not appear to reveal strong sub-groupings within the data.



(a) Silhouette Score: Months 1:3

(b) Silhouette Score: Months 10:12

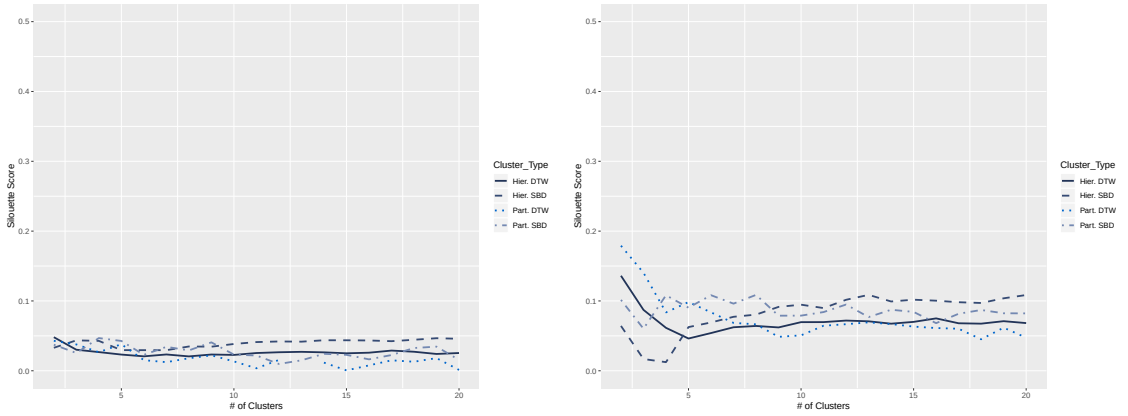
Figure 14: Kmodes Clustering: Step Count Trend Cluster Assignments

4.5 Clustering Summarised Time Series

As with procedures in the preceding sections, clustering of the summarised time series returned generally low cluster validation scores. Figure 15 displays the resulting CVI for weekly and monthly summarisations evaluated over $k = 2 : 20$. The weekly summarisations produced a uniformly low silhouette score, based on a 52-point summarised time series taken from a 1-year period. The monthly summarised times series produced slightly higher scores, based on a 12-point time series also taken from a 1-year period. This is in keeping with observed results in the preceding sections, i.e. a negative correlation of CVI score with time series interval length.

4.6 Kmeans Clustering of Time Series by Summary Statistics

Following stratification and variable derivation, the activity data were analysed by $k$means clustering and validated using CVI. Figure 16a displays the results of the silhouette score for $k = 3 : 20$. While there appeared to be some clustering apparent at $k = 5$, the cohesion and separation between clusters were not overly strong. The data were also visualised using t-distributed stochastic neighbour embedding (t-SNE) which preserves multi-dimensional pairwise distances in 2 dimensions. The t-SNE visualisation displayed in Figure 16b demonstrates the challenges with clustering these data. While the maximum CVI score appears at $k = 5$, the clusters apparent in Figure 16b are not clearly delineated,



(a) Silhouette Score: Step Mean Weekly (b) Silhouette Score: Step Mean Monthly

Figure 15: Clustering Summarised Time Series

with multiple possible densities apparent.

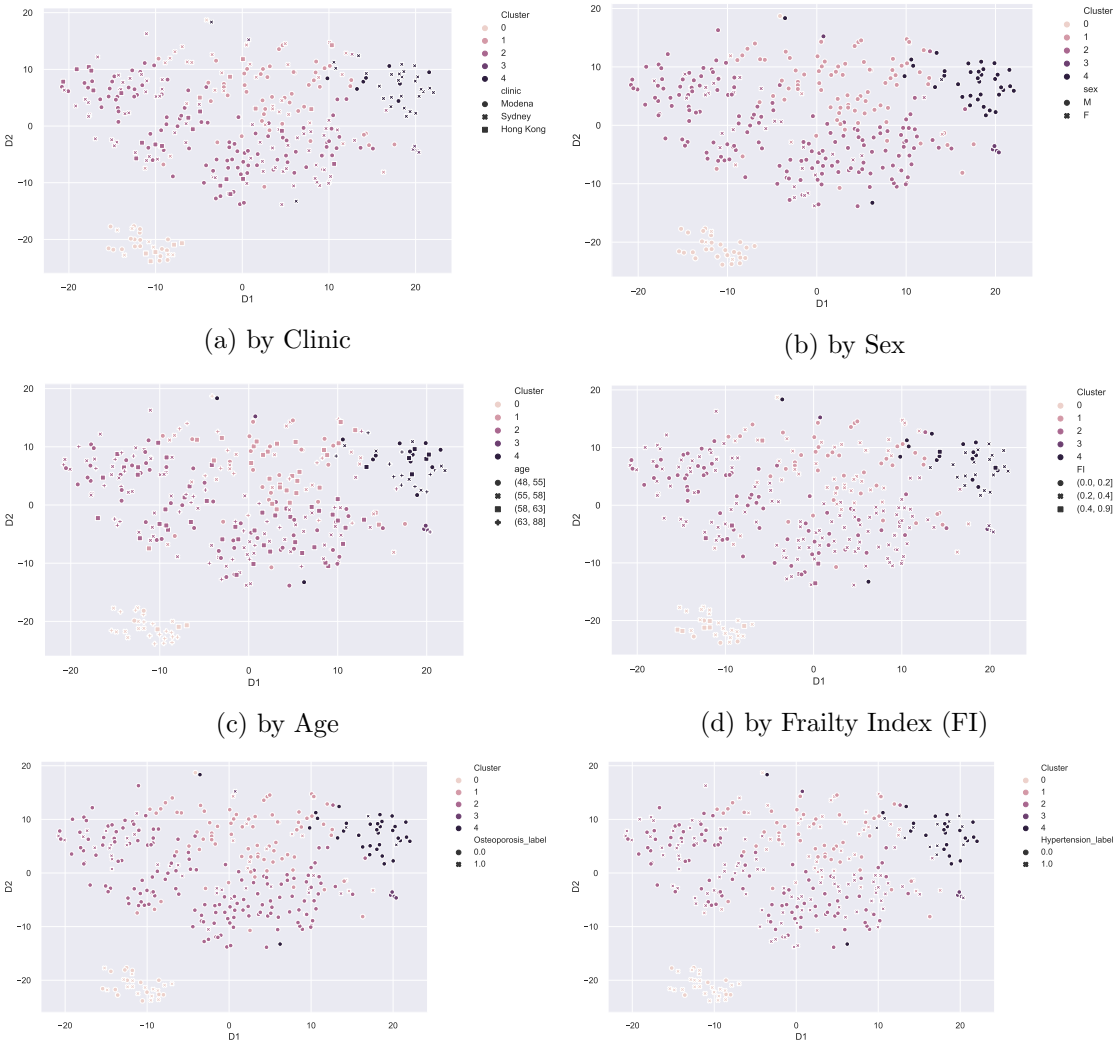
(a) Silhouette Score ($k = 3 : 20$)

(b) t-SNE Visualisation

Figure 16: Kmeans Clustering of Activity Summary Features

Despite failing to provide an indication of overall cluster validity, Figure 16b does provide an interesting line of inquiry with regard to the peripheral densities which align with the $k$means cluster assignment. In the visualisation, the colour channel corresponds to the cluster assignment for $k = 5$. As demonstrated, 2 of the large peripheral densities align with algorithmic cluster assignments (0 and 4) as well as the smallest density (cluster 4).

Further analysis was conducted using t-SNE visualisations and cluster meta data as demonstrated in Figure 17. None of the cluster-aligned peripheral densities demonstrated any notable overlap with the age, sex, clinic and FI stratifications. These stratifications appeared to be distributed randomly across all marks. However, the t-SNE visualisation did reveal an apparent over-representation of osteoperosis and hypertension cases in cluster 0, the most prominently distinct density apparent in the visualisations (ref. Figures 17e 17f).



(a) by Clinic

(b) by Sex

(c) by Age

(d) by Frailty Index (FI)

(e) by Osteoperosis

(f) by Hypertension

Figure 17: t-SNE Visualisation of Activity Summary Features Stratified

5 Objective 2: Predictive Modelling of Frailty Index

Modelling of frailty index (FI) was conducted using various models and multiple data from the MySAwH study. The methods involved with preparing the feature sets and designing the model comparison are described in the following section. The performance of the resulting models is discussed immediately afterwards.

5.1 Methods

The following sections detail the feature set construction, modelling plan, models used and dimensionality reduction methods used to predict frailty index (FI) using the patient-generated data.

5.1.1 Feature Set Construction and Modelling Plan

The modelling phase required construction of separate feature sets built on the Activity and EMA data. These feature sets were based on FI sourced from the Biomarkers dataset as ground truth. Both the Activity and EMA feature sets were constructed by segmenting the observations for a given subject by FI assessment interval, i.e. the interval between successive FI assessments. Each subject had FI assessments performed approximately every 9 months over the course of the MySAwH study. Referring to Figure 18, the dates of each FI assessment were used as an interval for segmenting the activity and EMA data to the corresponding ground truth (FI). For instance, for a subject with a Month 9 FI assessment on 01/01/2017 and a Month 18 FI assessment on 01/09/2017, all Activity and EMA observations in the period 01/01/2017 through 01/09/2017 were summarised and labelled with the Month 18 FI value.

Figure 18: Labelled Feature Vector by FI Assessment Interval

Summarisations used on the Activity data in the interval preceding a given FI assessment included mean, variance, minimum and maximum for each activity type, as well as derived threshold variables. All of these summarisations were further stratified by weekday and weekends. As this feature set was first used as an input to the time series clustering

models covered previously, details of the methodology used for the summarisations are provided in Section 4.1.3.

The cleaned Biomarkers dataset contained 560 FI assessments. However, a large portion of these were baseline FI assessments, with no preceding Activity/EMA data. Therefore, the resulting usable observations were considerably reduced. The finalised Activity feature set contained 352 labelled feature vectors, and the EMA feature set, 324.

Figure 19 provides an outline of the model comparison schema. As illustrated by Figure 19, the Activity and EMA feature sets were used to train various predictive models of FI as ground truth. Additionally, a model was trained using the Biomarkers in addition to the Activity and EMA data. The Biomarker data were used to compute the ground truth (FI) using a clinical formula. Therefore this 'control' model was included to provide a comparative benchmark for the other models, as well as to provide an indication of the validity of FI as a ground truth. The performance of each model was then compared to assess the relative strength of the patient-generated data in prediction of FI.

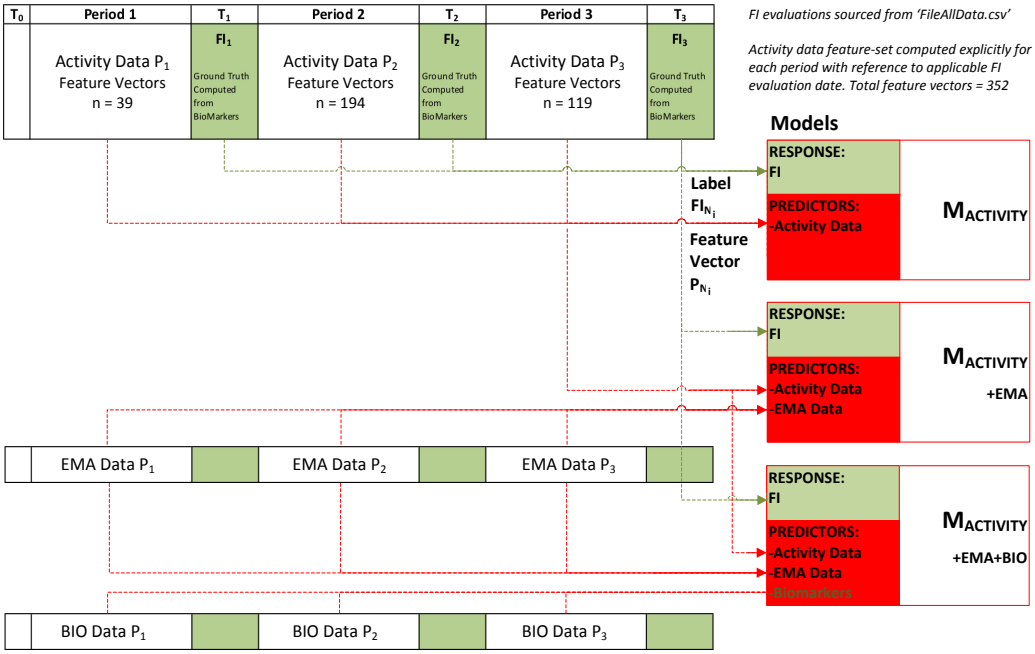


Figure 19: Comparative Modelling of Frailty Index (FI) using various predictors

In order to ensure a common scale for modelling, all feature sets were standardised prior to model input. Additionally, 10 percent of the data were split from the main training set and allocated to a test set for out-of-sample performance validation. Model performance was evaluated using R^2 to measure fit to the training data, while the root mean squared error (RMSE) was used to measure performance accuracy on the out-of-sample test set.

5.1.2 Algorithms

In keeping with the domain requirements, only interpretable machine learning models were used in the prediction of FI. As the purpose of the analysis was to replicate the readily understandable computation of FI using a patient-generated data-driven approach, it was crucial that the model provide interpretable predictions and not be a black box. This constraint informed the models used for the analysis. Linear Regression is an inherently interpretable model where the coefficients for each term signify the effect on the predicted value. Regression Trees are also inherently interpretable, except in cases of very large trees with extensive size and depth. Generalised Additive Models (GAM) provide interpretability via the analysis of each term's link functions. Ensemble boosting methods such as XGBoost can be interpreted with Shapley values.

All machine learning was deployed in Python using relevant packages. A Linear Regression model utilised the *sklearn* package to provide a base modelling of FI using each feature set. A Regression Tree model was then deployed using Microsoft's *interpret* package.

The GAM model used the Explainable Boosting Machines (EBM) algorithm, which comprises the primary functionality of the *interpret* package. The API provides an interface to a Generalised Linear Model with Interactions (GA^2M). As GA^2Ms are built on linear models, the link functions comprising each term in the model can be assessed independently for contribution to the predicted value. The package provides accessible breakdowns and visualisations of local interpretability, i.e. contribution of each term to a given prediction. The single terms can be viewed as 2 dimensional plots showing the dynamic between the response vs. predictor variable. The 3 dimensional pairwise interactions can be viewed as a heatmap where the colour gradient marks the response variable, FI. The model also provides global interpretability, i.e. rankings of overall feature importance.

All EBM models were trained use the top 5 most contributing interactions between terms. This functionality allows the user to simply specify an integer number n of interactions. The algorithm then tests the interactions of all terms and selects n interactions with regard to the effect on the response variable. The model also allows manual specification of individual interactions, however this functionality was not used in this analysis as it would have required detailed domain knowledge.

In addition to GA^2M, the gradient boosting algorithm XGBoost was also used to predict FI. Though not inherently interpretable, XGBoost was deployed alongside the *SHAP* package for Python. This package provides interpretability via Shapley values. Though less intuitively decipherable than linear model interpretability, Shapley values provide a robust understanding of how different terms contribute to predicted values.

SHAP (SHapley Additive exPlanations) provides interpretability for each predictor variable by computing models for a stochastically determined subset of all possible combinations of predictors, then taking the difference between the average value contributed by a single predictor and the predicted value [23]. The algorithm is able to work with

any machine learning model and is thus highly flexible. Figure 20 provides a high level overview of how SHAP is integrated into a model to explain predictions.

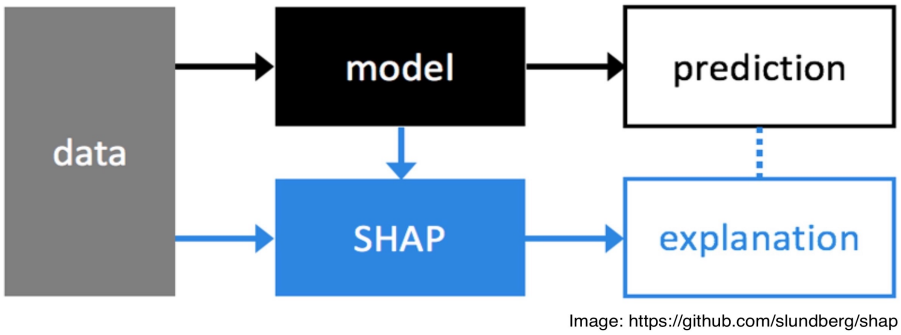


Figure 20: Overview SHAP Interpretability

5.1.3 Variable Collinearity

Variable multi-collinearity can affect the interpretability of linear models by introducing variability in the weights or link functions for each term [32]. This variability can lead to ambiguity of meaning, as the weight values cannot then be relied upon to provide an understanding of the contribution of each term to model predictions. Highly correlated variables must then be managed before modelling, either through a regularisation function or other method of dimensionality reduction [33].

Prior to modelling, the feature sets were assessed for highly correlated variables. Figure 21 presents a heatmap demonstrating the relative correlation of the variables in the activity dataset. Extreme red or blue shades indicate high positive or negative correlation, while yellow font indicates a correlation index value greater than 0.7. As shown, the dataset contained numerous very highly correlated variables. This was to be expected, as the variables were all derived from only 3 original raw variables.

Following the correlation analysis, the dimensionality of the activity data was reduced by removing correlated variables with a correlation index value above a threshold of 0.7 [33]. The reduced dataset is displayed in Figure 22. As demonstrated, over the half of the original 30 variables were cut, with only 13 features remaining in the reduced dataset.

Figure 23 presents the correlation analysis for the variables in the EMA dataset. As with the activity data, the EMA data contained numerous very highly correlated variables. However, in this case the correlations were more structured. Closer inspection of Figure 23 reveals the largest contiguous batch to be a result of high sectional correlation stemming from the EMA positive and negative mood sections. This effect could indicate possible problems with EMA survey frequency or questionnaire design. Excessively frequent survey requests from the MySAwHApp smartphone application could have led to entry fatigue. Moreover, the mood section of the questionnaire involves Likert scale questions where the subject is prompted to assess their level of 20 different emotional states on a scale of 1 to 5. These factors could have contributed to disengagement of the subjects and resulting

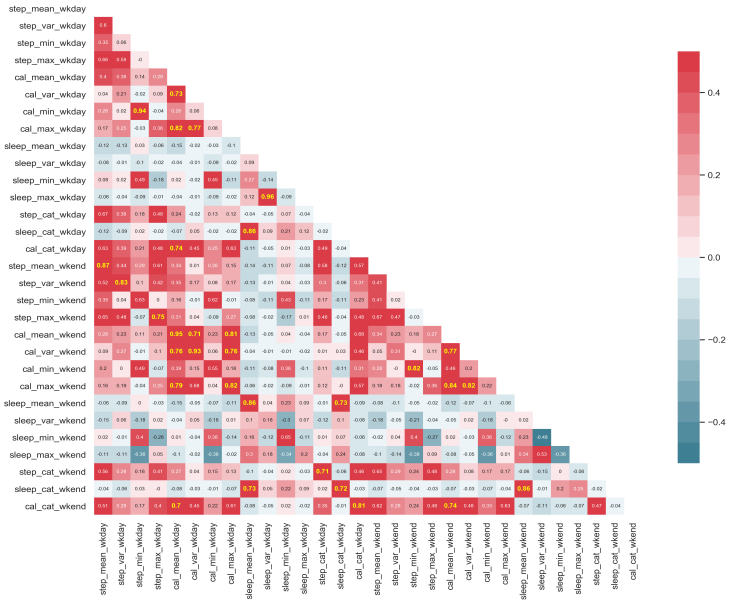


Figure 21: Correlation Analysis: Activity Data

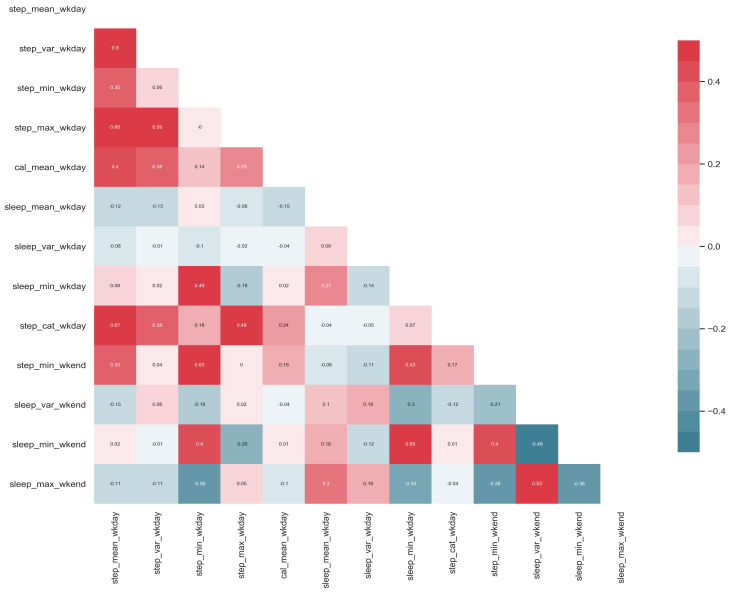


Figure 22: Correlation Analysis: Activity Data (Reduced)

inaccuracy in the responses.

The dimensionality of the activity data was reduced by removing correlated variables

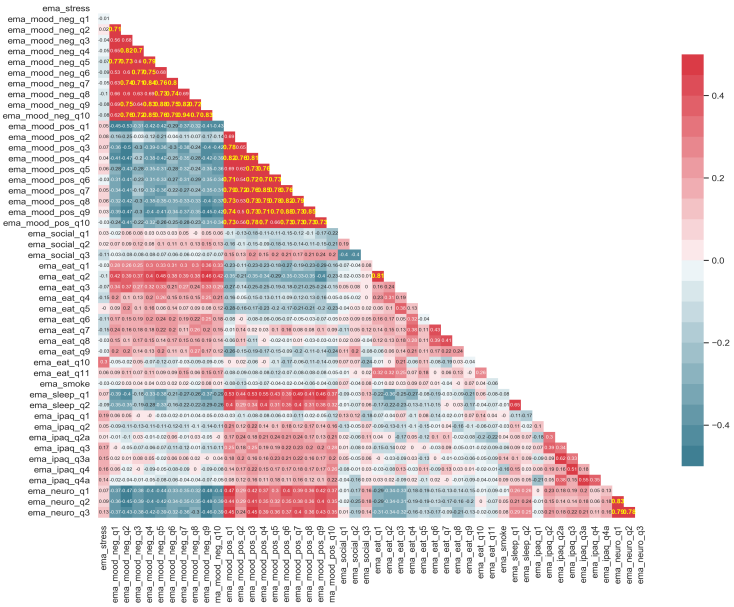


Figure 23: Correlation Analysis: EMA Data

with a correlation index value above a threshold of 0.7. The reduced dataset is displayed in Figure 24. As demonstrated, over the half of the original 30 variables were cut, with only 13 features remaining in the reduced dataset. The bulk of the removed variables were associated with the highly correlated mood questions.

Figure 23 presents the correlation analysis for the variables in the biomarkers dataset. Unlike the activity and EMA datasets, the biomarkers data were not overly correlated. As these data were only used for comparative purposes in the predictive modelling analysis, the dataset was not subjected to variable reduction.

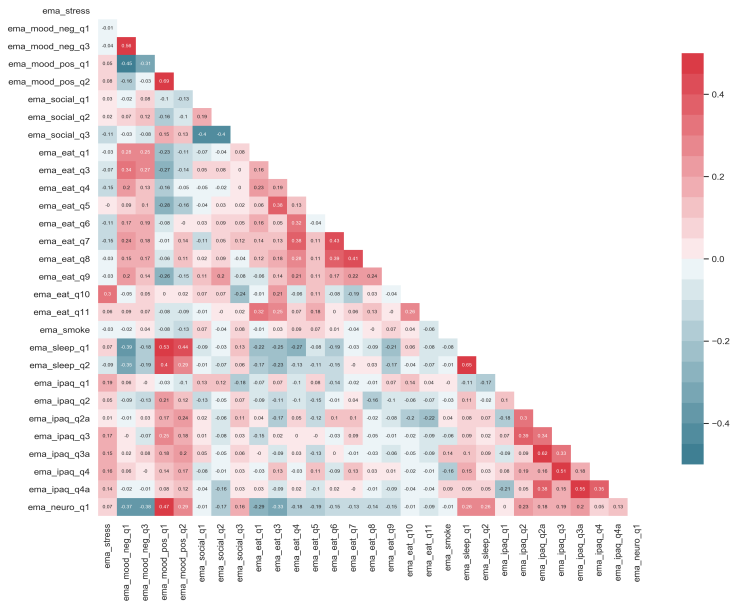


Figure 24: Correlation Analysis: EMA Data (Reduced)

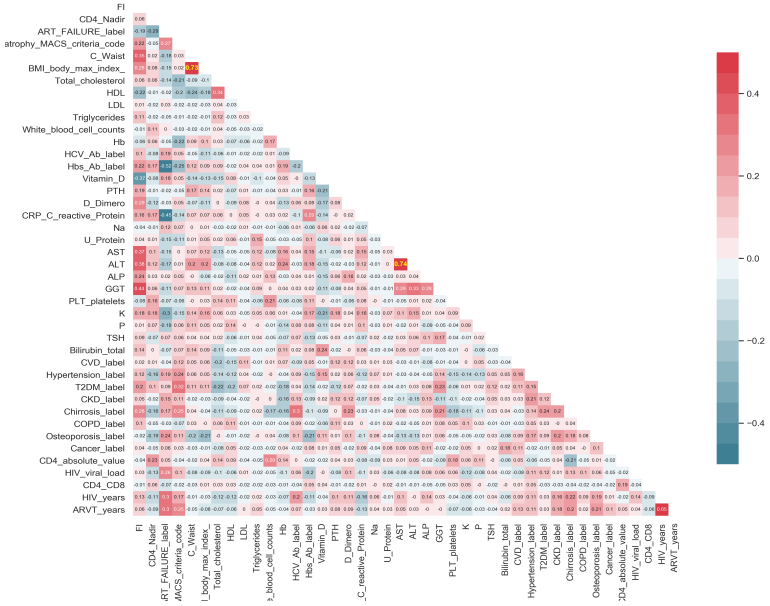


Figure 25: Correlation Analysis: Biomarkers

5.2 Results

Despite intensive feature engineering, variable reduction and powerful models, no model was found capable of providing strong predictive performance of frailty index (FI). Furthermore, the control model performance appears to invalidate the use of FI as a robust ground truth for machine learning models. The following sections document the results of the modelling efforts.

Figure 26 provides a histogram of the response variable, frailty index (FI). The distribution is characterised by a mean of 0.229 and a standard deviation of 0.092. The distribution of FI is important, because it provides context for understanding the significance of the model validation metrics.

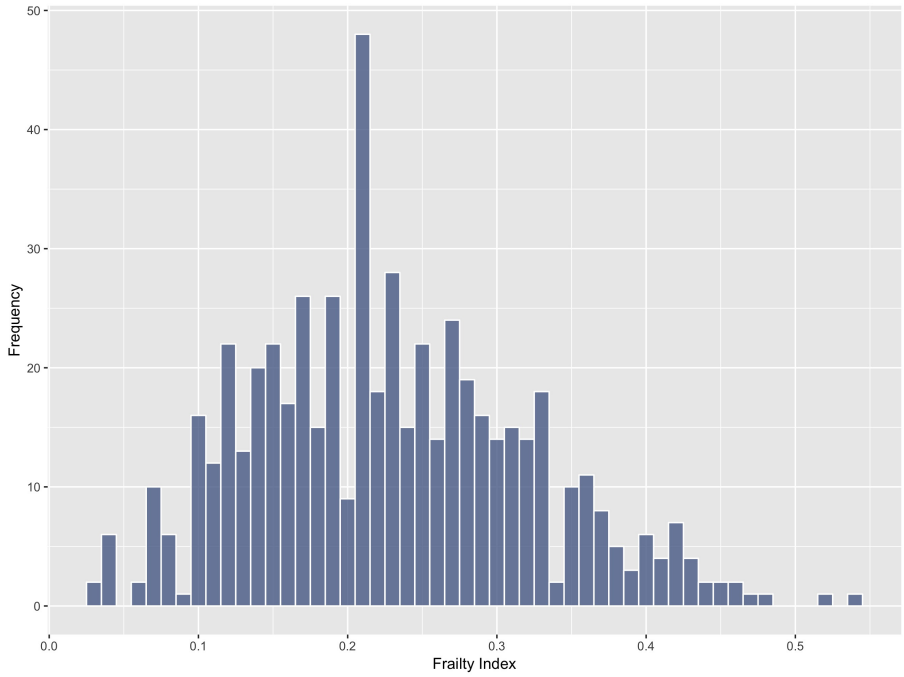


Figure 26: Distribution of Frailty Index Scores (n = 560)

5.2.1 Model Performance Comparison

Training the Activity data using a linear regression model produced poor results. The model exhibited low R^2, indicating a poor explanation of total variance, and relatively high root mean squared error (RMSE) stemming from predictions made on out of sample data. Referring to Table 5, the highest R^2 exhibited for the patient-generated data was 0.126 for the Activity model with an RMSE of 0.1. As the standard deviation of the FI distribution is 0.092, the RMSE is very high. As expected, the control model using Activity, EMA and Biomarkers data did better, with an R^2 of 0.217. However, the RMSE for this model was still high at 0.093.

Table 5: Model Comparison by Performance

Model	**Feature Set**	**RMSE**	**R^2**
Linear Regression	Activity	0.099	0.126
Linear Regression	EMA	0.111	-0.152
Linear Regression	Activity + EMA	0.103	0.048
Linear Regression	Activity + EMA + BIO	0.093	0.217
Regression Tree	Activity	0.100	0.140
Regression Tree	EMA	0.110	-0.140
Regression Tree	Activity + EMA	0.100	-0.010
Regression Tree	Activity + EMA + BIO	0.080	0.390
EBM (GA^2M)	Activity	0.100	0.090
EBM (GA^2M)	EMA	0.110	-0.060
EBM (GA^2M)	Activity + EMA	0.100	0.040
EBM (GA^2M)	Activity + EMA + BIO	0.060	0.580
XGBoost	Activity	0.078	NA
XGBoost	EMA	0.079	NA
XGBoost	Activity + EMA	0.079	NA
XGBoost	Activity + EMA + BIO	0.057	NA

Interestingly the R^2 decreased for the Activity + EMA model vs. the simpler model with only Activity data. This would seem to indicate an inherently weak signal for the EMA data in linear regression modelling of FI. Therefore a model was constructed using EMA exclusively, with the results again displayed in Table 5. As suggested by the joined model performance, the R^2 for the EMA model was considerably worse at -0.152 than the joined model and the Activity only model. Therefore the EMA data appear to be a very poor predictor of FI.

Table 5 also presents the result of Regression Tree modelling of FI using the reduced feature sets. As with Linear Regression, the performance of the two patient-generated models was poor. Again, the model including the Biomarkers exhibited much better performance. This model also improved on performance compared to Linear Regression, with an R^2 of 0.39 and an RMSE lower than one standard deviation of FI at 0.08.

The EMA data again pulled the model performance down when combined with Activity features. The same poor performance was again assessed in a separate EMA-only model, which produced an R^2 of -0.14, underlining the relative unsuitability of the EMA data for modelling FI when compared to the Activity feature set.

Performance with the more advanced Explainable Boosting Machines (EBM) model was equivalent to the previously tested models on the patient-generated data. EBM exhibited R^2 scores of 0.09 for the model built on Activity data, and 0.04 for the combined Activity + EMA model. RSME was 0.1 for both models. With EBM, the control model exhibited very good R^2 at 0.58 and RSME at 0.6.

The ensemble boosting model XGBoost provided the best overall performance on the out-of-sample validation. The Activity model scored an RMSE of 0.078, with the Activity + EMA scoring only slightly worse at 0.079, and the control model scoring 0.057.

5.2.2 Model Interpretability

Despite the poor overall predictive performance measures of each model, the process did provide some intuitively logical insight into the relationships of select predictor variables with the response variable, FI. The EBM implementation with *interpret* package and XGBoost with *SHAP* package provide an excellent API for understanding both global effects of predictors on model output, as well as explaining individual predictions.

Figure 27 demonstrates the global interpretability of the Activity feature set modelled with EBM. The contributory importance of the top 15 predictor variables on the response variable are detailed in Figure 27a. The bar chart reveals the weekday step count mean to be the primary contributor to FI, with a mean contribution in excess of approximately 0.004. Other weekday step and sleep variables comprise the remainder of the top 5 contributors.

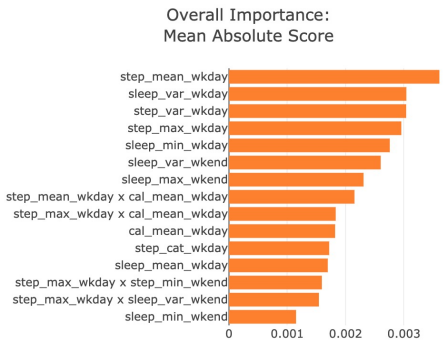
Figure 27b breaks down the 2 dimensional dynamic between weekday step count mean and FI, showing how FI responds to changes in step count. The gray regions around the plotted line indicate confidence intervals. The histogram at the bottom represents the distribution of the predictor variable. The plot shows a clear negative correlation between step count and FI, with high relative FI values at low daily step counts, and FI decreasing with increased step counts.

Unfortunately many of the other global contribution plots for the top predictors produced somewhat ambiguous results. Referring to Figure 27c, it is difficult to ascertain the relationship between the predictor variable, weekday sleep variance, and the response, FI. The response appears to increase with along with sleep variance before dipping.

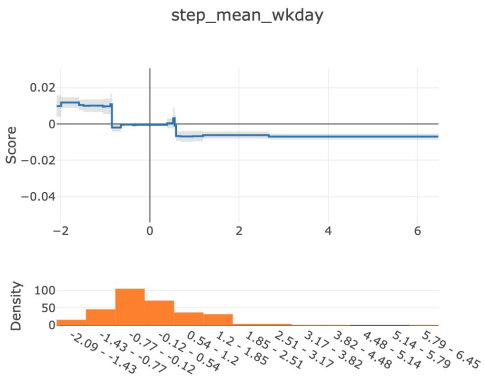
The pairwise interaction plots also did not provide clear results. Figure 27d displays the interaction between weekday step count mean and calories mean. The predictor variables comprise the x and y axes, while the colour marker signifies FI. The resulting heatmap is inconclusive, with no clear and consistent relationship apparent between the interacting terms and FI.

The modelling of FI using XGBoost provides further insight into the relationship between the predictor variables and the model output. As with the EBM API, the SHAP provides an easily accessible summary of relative predictor importance. Figure 28a displays the top 15 predictor variables by importance. Of the top 5, weekday step mean, maximum and variation match predictors found in the top 5 for the EBM model highlighting the predictive strength of these variables across models.

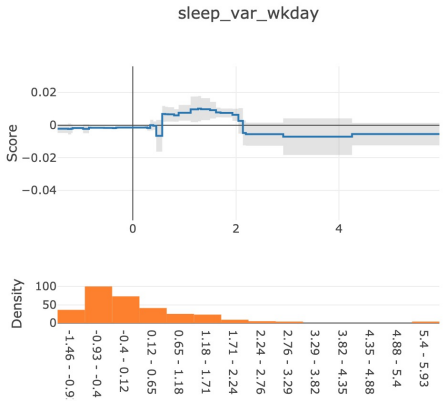
The package also returned scatter plots approximating the relationship of each predictor with the response, FI for each observation. Figure 28b demonstrates how the Shapley value for weekday maximum count varies with changes in the weekday step maximum, thus providing an approximation of the model output (FI). The plot appears to show



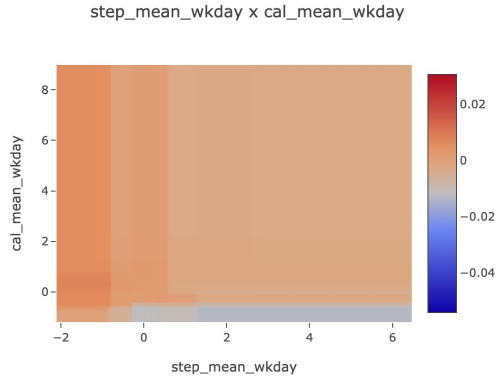
(a) EBM (Activity) Predictor Contribution



(b) Weekday Step Mean



(c) Weekday Sleep Variance



(d) Weekday Step Mean * Calories Mean

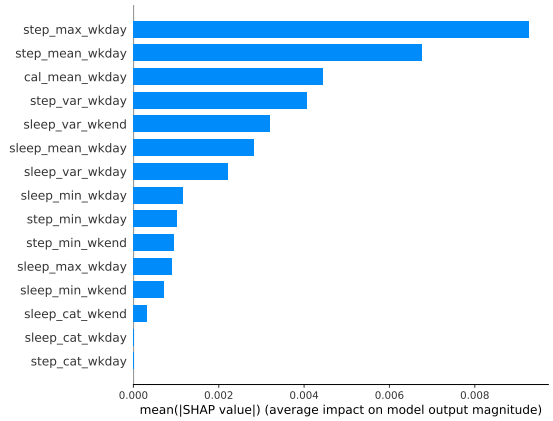
Figure 27: EBM (Activity) Global Contribution

a negative exponential correlation, with the Shapley value initially decreasing sharply with increases in maximum step count, and then levelling out. The plot also provides insight into the top interaction variable with weekday maximum step count, i.e. weekday mean step count. The colour marker indicates the corresponding level of weekday mean step count for each observation. The interactive relationship between the two variables is logical, in that high step count maximum values translate to high means - a dynamic typical of the mean as a measure of central tendency, as it is susceptible to extreme values.

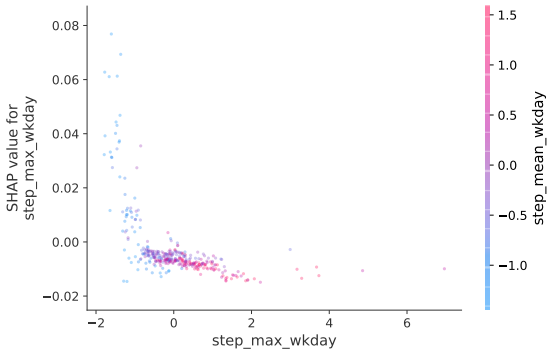
Figure 28c provides another example of approximated global predictor affect on model output, this time for the weekday step count mean with weekend sleep variance as an interaction term. The plot characterises a gradual negative exponential correlation with model output. This observation supports the findings in 27b for the same variable. However the interaction with weekend sleep variance appears to be randomly distributed across all observations, without any clear pattern.

As with the EBM interpretation of predictors on model output, not all the variables exhibited logical relationships with the approximated model output. Figure 28d represents

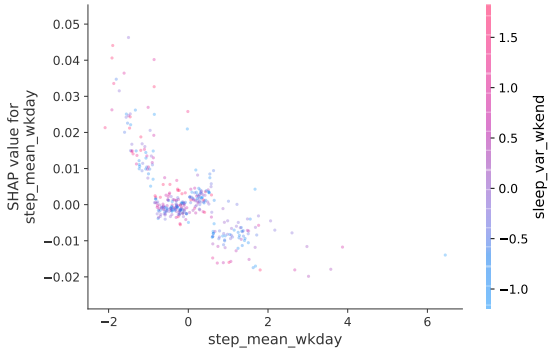
the effects of weekday calorie count. It is difficult to decipher any clear relationship between the predictor and FI in this case. The interaction variable, weekday step mean does however show the expected relationship with calories, i.e. a positive correlation.



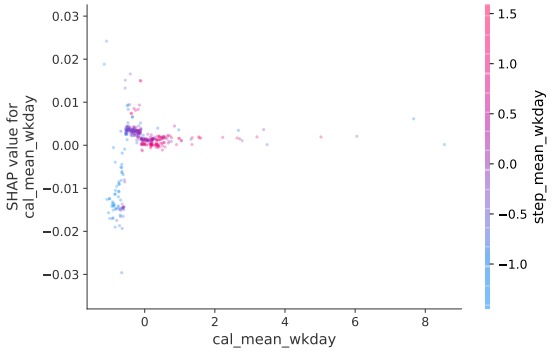
(a) XGB (Activity) Predictor Contribution



(b) Weekday Step Max



(c) Weekday Step Mean



(d) Weekday Calories Mean

Figure 28: XGBoost (Activity) Global Contribution

6 Discussion

Overall the results of the analyses did not reveal substantive findings pertinent to the major research objectives.

6.1 Objective 1

The time series clustering returned poor results with low cluster validity observed in most cases for all of the main clustering approaches. In the case of raw time series clustering, this was likely due to excessive variance in the raw time series - i.e. there was no smoothing or summarisation applied to the data. For all approaches, the worst results were observed over the longest periods for all categories of activity data. This correlation between longer intervals and diminished cluster validity was possibly caused by the higher dimensional space of the longer intervals. The longer intervals could have introduced a greater likelihood of variance, and therefore any two series could have appeared more dissimilar to distance measuring algorithms.

Clustering of the weekly seasonality offered better validity and was a potentially valuable point of inquiry. However, the low granularity of the weekly seasonality decomposition calls into question the significance of any conclusions derived from analyses based on these data. At only seven points, the time series is very short and does not provide a granular representation of subjects' activity patterns over the course of a week.

The decomposed time series single-month trend also showed promising cluster validity. Visual analysis of the resulting cluster plot revealed trends that were intuitively understandable as characterising groupings of typical human activity. This point of analysis returned some interesting insight, in that almost half of the subjects exhibited a downward trend in activity levels throughout the first month of the study. This would indicate that the greater majority of subjects actually became less active in the first month of the study.

While no strong overall groupings were found, the clustering of time series by stratified weekday and weekend summary statistics reveals some minor insights. Visualisation of the clustering by t-SNE visualisation with overlaid stratification revealed an apparent over-representation of osteoperosis and hypertension cases in one of the distinct clusters. These two indicators are indicative of decreases functional capacity and therefore likely represent a distinct subgroup in terms of activity levels.

Challenges with the experiment design likely contributed to the lack of cohesive findings with time series clustering. The subjects were dispersed amongst 3 very different climates and residing very different cities in terms of transport and lifestyle. Additionally the data were minimal to begin with, and contained numerous cases of missing entries.

Possible improvements to future experiments could include a more rigidly controlled deployment of activity trackers. Following up with subjects to ensure correct and regular usage of the activity trackers would likely yield better results and avoid cases of excessive missing data. Furthermore, a more granular data collection taking hourly readings could

provide better insights into similarities in activity patterns when clustered on a weekly seasonality.

6.2 Objective 2

Modelling of frailty index using patient-generated data did not produce strong predictive performance. The more advanced models produced the best results with XGBoost providing the minimum predictive error. This improved performance likely reflects the increased complexity of the model when compared with Linear Regression, Regression Trees and Generalised Additive Models with Interactions (GA^2M).

The comparative modelling schema did provide insight into the validity of the frailty index (FI) as ground truth. The control model using the clinically assessed biomarker data produced much better predictive performance than the models which used patient-generated data exclusively. However, the absolute performance of the clinical-generated data model was not to the level expected. These clinical data directly inform the computation of the ground truth via a clinical formula. Therefore a model trained on these data should theoretically exhibit very strong predictive performance. The lack of very strong observed performance by the model appears to indicate that the frailty index data used in this analysis is not a robust ground truth for machine learning

The interpretability of the models provided some interesting insights for the top predictors by contributory importance. Weekday step count appears to be the primary contributor to both the EBM and XGBoost models, perhaps reflecting subjects commutes to work or the nature of their jobs. This could also possibly be due to the longer interval length compared with weekends providing a larger sample size for certain summary statistics (eg. maximum). The plots of single predictor effects on model output also showed a clear negative correlation between step count and FI, with high relative FI values at low daily step counts, and FI decreasing with increased step counts. This again likely indicates that subjects who walk to work or have more active jobs are more likely to have lower measurable frailty.

Other global contributions for the top predictors were less clear. The EBM model output for weekday sleep variance appears to increase with along with sleep variance, which makes intuitive sense, as regular sleep patterns would tend to relate positively with overall health. However, the response then dips inferring the opposite, i.e. increased variance in sleep patterns correlates with decreased frailty. The histogram for weekday sleep variance may give some indication as to the problem. With the decreasing FI occurring over the long tail of the sleep variance distribution, it is likely the low quantity of samples is not providing an accurate reflection of the true effect. The large confidence intervals over this range support this assessment.

As an outcome of this analysis, a suggested improvement for future studies would be to reexamine the use of frailty index (FI) as a basis for assessing functional capacity in older adults living with HIV (OALWH). Based on the findings of this analysis, there does not appear to be a strong relationship with clinically assessed frailty and general activity

levels. A more relevant index could simply be based on activity level summary statistics or categorical thresholds.

7 Conclusion

In conclusion, this analysis did not succeed in developing a method of replacing clinical assessments of frailty with patient-generated data analytics. Clustering of activity data time series did not reveal strong subgroupings within the data. Additionally, despite numerous modelling configurations and feature engineering, the activity data feature set did not produce viable predictive accuracy when trained on frailty index as ground truth.

Overall the time series cluster analyses did not result in strong validated clustering observed amongst the cohort. In some potentially observed clustered groups, low granularity presented a barrier to drawing inference from the results. In others, the cluster assignments were only valid over short intervals. Visualisations of stratified clusters did reveal potential relationships with certain clinical indicators.

Prediction analysis of the data was affected by an unclear relationship of the standardised method of assessing frailty (frailty index) with general activity levels. Based on the results of this analysis, frailty index does not appear to be a robust ground truth for training a model using activity data.

In closing, further research in this area is required to develop methods of using patient-generated data to provide frequent and automatic assessments of functional capacity. Higher granularity activity trackers could contribute to research outcomes. Additionally, future studies in this line would benefit from discarding frailty index, and experimenting with a different target index in order to quantify functional capacity in older adults living with HIV and other reduced-mobility demographics.

References

[1] Kaare Christensen, Gabriele Doblhammer, Roland Rau, and James W Vaupel. Ageing populations: the challenges ahead. *The Lancet*, 374(9696):1196 – 1208, 2009.

[2] David Reeves, Stephen Pye, Darren M Ashcroft, Andrew Clegg, Evangelos Kontopantelis, Tom Blakeman, and Harm van Marwijk. The challenge of ageing populations and patient frailty: can primary care adapt? *BMJ*, 362, 2018.

[3] WHO. World report on ageing and health. Report P-28, World Health Organization of the United Nations, 2015.

[4] Geeske Peeters, Annette J Dobson, Dorly JH Deeg, and Wendy J Brown. A life-course perspective on physical functioning in women. *Bulletin of the World Health Organization*, 91(9):661–670, 2013.

[5] Jose Lara, Rachel Cooper, Jack Nissan, Annie T. Ginty, Kay Tee Khaw, Ian J. Deary, Janet M. Lord, Diana Kuh, and John C. Mathers. A proposed panel of biomarkers of healthy ageing. *BMC Medicine*, 13(1):1–8, 2015.

[6] MySAwH Consortium. My smart age with hiv homepage. Available at https://www.mysmartage.org/ (2019/06/01).

[7] Thomas D. Brothers and Kenneth Rockwood. Frailty: a new vulnerability indicator in people aging with hiv. *European Geriatric Medicine*, 10(2):219–226, Apr 2019.

[8] I. Franconi, O. Theou, L. Wallace, A. Malagoli, C. Mussini, K. Rockwood, and G. Guaraldi. Construct validation of a Frailty Index, an HIV Index and a Protective Index from a clinical HIV database. *PLoS ONE*, 13:e0201394, October 2018.

[9] G Guaraldi, A Malagoli, O Theou, TD Brothers, LMK Wallace, R Torelli, C Mussini, S Sartini, SA Kirkland, and K Rockwood. Correlates of frailty phenotype and frailty index and their associations with clinical outcomes. *HIV Medicine*, 18(10):764–771, 2017.

[10] Giovanni Guaraldi, Thomas D Brothers, Stefano Zona, Chiara Stentarelli, Federica Carli, Andrea Malagoli, Antonella Santoro, Marianna Menozzi, Chiara Mussi, Cristina Mussini, Susan Kirkland, Julian Falutz, and Kenneth Rockwood. A frailty index predicts survival and incident multimorbidity independent of markers of hiv disease severity. *AIDS (London, England)*, 29(13):1633—1641, August 2015.

[11] M. Cesari. Intersections between Frailty and the Concept of Intrinsic Capacity. *Innovation in Aging*, $1(\text{suppl}_1) : 692 - -692, 062017$.

[12] Matteo Cesari, Islene Araujo de Carvalho, Jotheeswaran Amuthavalli Thiyagarajan, Cyrus Cooper, Finbarr C Martin, Jean-Yves Reginster, Bruno Vellas, and John R Beard. Evidence for the Domains Supporting the Construct of Intrinsic Capacity. *The Journals of Gerontology: Series A*, 73(12):1653–1660, 02 2018.

[13] Philippe Esling and Carlos Agon. Time-series data mining. *ACM Comput. Surv.*, 45(1):12:1–12:34, December 2012.

[14] Tak Chung Fu. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164–181, 2011.

[15] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. Time-series clustering - A decade review. *Information Systems*, 53:16–38, 2015.

[16] Alexis Sarda-Espinosa. Comparing time-series clustering algorithms in r using the dtwclust package. 2017.

[17] Eamonn J. Keogh and Michael J. Pazzani. Derivative Dynamic Time Warping. pages 1–11, 2001.

[18] John Paparrizos and Luis Gravano. K-shape: Efficient and accurate clustering of time series. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2015-May:1855–1870, 2015.

[19] Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza, Jesús M. Pérez, and Iñigo Perona. An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1):243–256, 2013.

[20] Laurens van der Maaten and Hinton Geoffrey E. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 164(2210):10, 2008.

[21] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1721–1730, New York, NY, USA, 2015. ACM.

[22] Scott M. Lundberg and Su-In Lee. Consistent feature attribution for tree ensembles. (Section 5), 2017.

[23] Scott Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. (Section 2):1–10, 2017.

[24] Geoff Norman. Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, 15(5):625–632, 2010.

[25] Steffen Moritz, Alexis Sardá, Thomas Bartz-Beielstein, Martin Zaefferer, and Jörg Stork. Comparison of different Methods for Univariate Time Series Imputation in R. 2015.

[26] Xiaohui Huang, Yunming Ye, Liyan Xiong, Raymond Y.K. Lau, Nan Jiang, and Shaokai Wang. Time series k-means: A new k-means type smooth subspace clustering for time series data. *Information Sciences*, 367-368:1–13, 2016.

[27] Fuyuan Cao, Jiye Liang, and Liang Bai. A new initialization method for categorical data clustering. *Expert Systems with Applications*, 36(7):10223–10228, 2009.

[28] Sophie Cassidy, Josephine Y. Chau, Michael Catt, Adrian Bauman, and Michael I. Trenell. Cross-sectional study of diet, physical activity, television viewing and sleep duration in 233 110 adults from the UK Biobank; The behavioural phenotype of cardiovascular disease and type 2 diabetes. *BMJ Open*, 6(3):1–11, 2016.

[29] Sophie Cassidy, Harley Fuller, Josephine Chau, Michael Catt, Adrian Bauman, and Michael I. Trenell. Accelerometer-derived physical activity in those with cardio-metabolic disease compared to healthy adults: a UK Biobank study of 52,556 participants. *Acta Diabetologica*, 55(9):975–979, 2018.

[30] Herrmann SD Meckes N Bassett Jr DR Tudor-Locke C Greer JL Vezina J Whitt-Glover MC Leon AS Ainsworth BE, Haskell WL. The compendium of physical activities tracking guide. *Healthy Lifestyles Research Center, College of Nursing Health Innovation, Arizona State University*, 2011. Accessed: 2019-08-10.

[31] Barbara E. Ainsworth, William L. Haskell, Stephen D. Herrmann, Nathanael Meckes, David R. Bassett, Catrine Tudor-Locke, Jennifer L. Greer, Jesse Vezina, Melicia C. Whitt-Glover, and Arthur S. Leon. 2011 compendium of physical activities: A second update of codes and MET values. *Medicine and Science in Sports and Exercise*, 43(8):1575–1581, 2011.

[32] David A. Belsley. A Guide to using the collinearity diagnostics. *Computer Science in Economics and Management*, 4(1):33–50, 1991.

[33] Robert M. O'Brien. A caution regarding rules of thumb for variance inflation factors. *Quality and Quantity*, 41(5):673–690, 2007.

8 Apppendix A: Responsible Research Innovation

A standardised, patient-generated method of assessing functional capacity has broad application to ageing populations at large, as well as to vulnerable subsets such as older adults living with HIV (OALWH) in providing higher frequency accuracy evaluations of frailty index (FI). This would in turn provide primary health care workers with the insight necessary to launch early interventions where applicable, resulting in better healthy ageing outcomes for ageing adults and OALWH. However, careful implementation of the predictive model must be observed in order to prevent underestimation of FI which could result in missed opportunities for intervention. The risk of underestimation is mitigated by the use of an interpretable model, allowing diagnosticians to carefully measure the contribution of any particular factor to the FI computation.

Stakeholders for this study include the subjects, researchers, and public healthcare systems.

Outcomes from this study are potentially transformational in that the proposed method would be part of a paradigmatic shift in geriatric medicine. The deployment of personalised, patient-generated FI assessments would be supportive of the move towards a more holistic system of public healthcare.

Alternative applications of this method could potentially involve use in the domain of competitive sports. Recent disputes at international sporting competitions have centred on the physiological and pharmacological definitions of sex with regard to eligibility to compete in a given gender bracket. The current system measures hormone levels such as testosterone to identify a competitor as eligible to compete in mens or womens events. Evaluation based on functional capacity could potentially provide a fair and gender-neutral assignment of the competitor to a particular classification.

Further development of wearable sensors with IoT connectivity could greatly expand the potential of this research. Such devices could provide a broader, more granular and dynamic assessment of an subject's functional capacity. When combined with an automated monitoring platform, physicians and patients could be provided with up to the minute analysis and recommendations based on FI.

As this effort is aimed at deployment within the confines of public healthcare systems, existing regulations on patient confidentiality and data rights already provide adequate controls.

8.1 Apppendix B: Research Data Management Plan

8.1.1 What data will be produced?

No data will be produced by this analysis. The data used are sourced entirely from the MySAwH study and consist of 3 main types: 1) clinically assessed biomarkers; 2) patient-generated activity data from a wearable device; and 3) patient-generated emotional status data from the MySAwH App. These data were provided in open-source comma-separated

values (.csv) format. As these data are highly sensitive, the version provided for this analysis has been anonymised.

8.1.2 How will your data be structured and stored?

Data will be stored on a private Gitlab repository and accessible only to the researchers involved.

8.1.3 How will the data be shared during and after the project?

As the data are highly sensitive, they will be shared only with the researchers involved with the analysis and the MySAwH study. Following completion of the analysis, a stripped-down anonymised version of the data may be released by the MySAwH study. However, this is problematic due to the possibility of reidentification.

8.1.4 Outline the approach to data selection and long-term preservation

Data will be preserved on Gitlab for the duration of the analysis. Following completion of the analysis, data will likely need to be migrated to an NCL-hosted repository. Base file format will remain as comma-separated values (.csv). As csv is the dominant open data exchange format, it provides inherent future-proofing to ensure long term accessibility.

8.1.5 Who has responsibility for implementing the DMP and are resources required?

Mark Tyrrell of NCL is responsible for implementing the DMP vis-a-vis this analysis. The MySAwH study researchers are the original custodians of the data and will maintain control following this analysis.