FUNCTIONAL DATA METHODS FOR UNDERSTANDING HUMAN PHYSIOLOGICAL SYSTEM RESPONSES TO EXERCISE

Nicholas Joseph Coronato

Albany, New York

Bachelor of Science, U.S. Military Academy, 2012

A Thesis submitted to the Graduate Faculty

of the University of Virginia in Candidacy for the Degree of

Master of Science

Department of Engineering Systems and Environment

University of Virginia

May 2022

Mehdi Boukhechba, Chair

Donald E. Brown, Advisor

Arthur Small

Dan Cooper MD (UC Irvine)

Functional Data Methods for Understanding Human Physiological System Responses to Exercise

Nicholas Joseph Coronato

(ABSTRACT)

Physiological response to physical exercise through analysis of cardiopulmonary measurements has been shown to be predictive of a variety of diseases. Nonetheless, the clinical use of exercise testing remains limited because interpretation of test results requires experience and specialized training. Additionally, the type and duration of the exercise testing most effective for prediction of fitness and disease remains controversial. This research examines the use of advanced machine learning methods to understand physiological mechanisms and to predict exercise test completion in a protocol consisting of multiple exercise bouts. Cardiopulmonary signals of 81 healthy children were captured breath-by-breath during these exercise bouts. We explored machine learning strategies to model the relationship between the physiological time series, the participant's demographic variables, and the binary outcome variable indicating whether the participant completed the test. The best performing model, a generalized spectral additive model with functional and scalar covariates, achieved 93.6% classification accuracy and an F1 score of 93.5%. Additionally, functional analysis of variance testing showed that participants in the 'quit' and 'not-quit' groups have significantly different functional means in three signals: heart rate, oxygen uptake rate, and carbon dioxide uptake rate. Overall, these results show the capability of functional data analysis to identify key differences in the exercise-induced responses of participants in multiple bout exercise testing.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The human cardiovascular and associated systems are dynamic and highly interrelated. The measurements resulting from exercise stress testing have been shown to be predictive of a variety of diseases and fitness levels (Schaefer et al. 2020, Guazzi, Raimondo, et al. 2007, Flynn et al. 2009, Myers et al. 2014, Mirizzi et al. 2016, Baril 2018, Hearn et al. 2018, Bazoukis et al. 2021). Standard exercise testing procedures, commonly applied for diagnostic purposes, produce outputs that must be interpreted by trained practitioners with an understanding of the underlying physiology and kinetics of the system, as well as an ability to interpret multiple time series (Neder et al. 2021). We are motivated to investigate alternative types of exercise test in order to assess their utility for predicting health outcomes. The duration and design of gold-standard "ramp" style exercise tests may not be optimal to identify cardiovascular signal patterns for clinical diagnosis. This is especially controversial in the case of pediatric patients, where the traditional cardiopulmonary tests may fail to mimic the nature physical activity patterns of children (Gilliam et al. 1981, Bailey et al. 1995, Bar-Yoseph et al. 2022, Armstrong 2019).

By applying machine learning techniques to multiple bout exercise testing, we seek to lay the foundation for quicker and more consistent interpretation of patterns in physiological time series that may eventually aid researchers in caring for their patient. Our work with functional data analysis shows that it may be a highly useful

method for clustering patients into groups of interest based on their exercise-induced cardiovascular signals. We propose that functional data analysis is particularly useful in the multiple bout exercise scenario, where repeated on-transient and off-transient patterns of the response variable are highly non-linear but especially informative.

After presenting the relevant background for this area of research in Section 1.1 and our research questions in Section 1.2, we describe the sample data set (2.1) and three exploratory data mining techniques (2.2.1-2.2.3) before arriving at Functional Data Analysis in Section 2.2.4.

1.1 Background

1.1.1 Exercise Testing Data

Exercise testing for diagnostic purposes is conducted by measuring physiological responses during graded physical exercise. Typically this is done by measuring gas exchange and cardiac condition in order to score the performance of coordinated human biological subsystems. Cardiopulmonary Exercise Testing (CPET or CPX) is the most common and widely used exercise test. CPET specifically measures responsiveness of the pulmonary, cardiovascular, neuropsychological, skeletal muscular, and hematopoietic systems. Since about the 1920s, $\dot{V}O_{2peak}$ has been considered the most widely used biomarker for aerobic fitness, and it is most commonly measured through CPET protocols as the "gold-standard" (Armstrong 2019). CPET has the advantage of being low-risk and non-invasive (Bartels and Prince 2021). The entire test lasts 40-60 minutes but includes only approximately ten minutes of exercise. Assessments during CPET are generally applicable to patients of all backgrounds and fitness ability

levels. Exercise can be conducted on either a treadmill or stationary cycle ergometer, each with their own advantages ("ATS/ACCP Statement on Cardiopulmonary Exercise Testing" 2003). The choice of exercise modality is typically driven by (1) the principle of specificity (i.e. a runner should be tested on a treadmill and cyclist on a cycle ergometer) or (2) safety considerations (the stationary cycle ergometer presents a lower fall risk). Typically, an initial rest phase is followed by a warm-up period and then a "ramp" protocol, in which work rate is gradually increased until the patient reaches volitional fatigue or test termination criteria are met (Liguori et al. 2021).

Figure 1.1: A child performs CPET on a cycle ergometer. (Dan M. Cooper, R. Bar-Yoseph, et al. 2019)

Through continuous electrocardiogram and other measurements, the relevant output data commonly recorded through CPET are: lung function through flow volume loops; blood pressure; oxygen consumption during exercise ($\dot{V}O_{2max}$); ventilatory

anaerobic threshold (VAT); heart performance; ventilation (VE), carbon dioxide volume (VCO_2), blood gas measurement, and various slopes such as the oxygen uptake efficiency slope (OUES - VE/$\dot{V}O_2$) and VE/$\dot{V}CO_2$. Lactate and end tidal volumes such as $PETCO_2$ and $PETO_2$ are also reported (*Cardiopulmonary Exercise Testing (CPET)* 2017). The data is described in a 9-panel report (Fig. 1.2) which can be leveraged for clinical diagnosis of cardiopulmonary abnormalities (Sietsema et al. 2020).

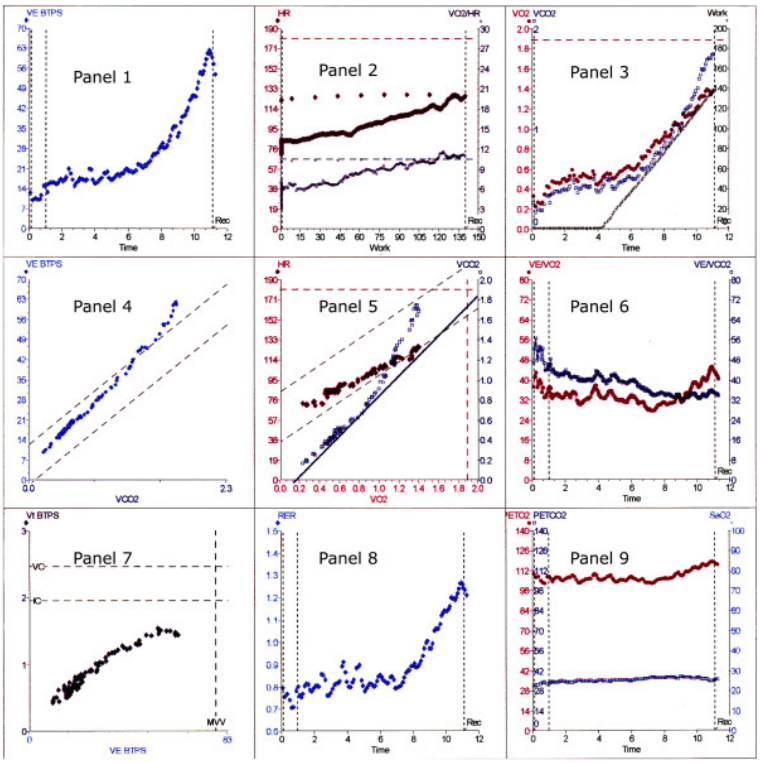


Figure 1.2: Traditional nine-panel plot. This format emanates from Wasserman and colleagues and is by far the most common. The 9-panel format allows 15 variables to be plotted on 9 graphs. (Older 2013)

Armed with the ability to analyze CPET reports, medical professionals or scholars should be able to definitively diagnose – or predict – these abnormalities. The subsequent therapy or rehabilitation program would then be more appropriately targeted at improving the precise symptoms and weaknesses in each individual patient.

Research suggests that prescription of rehabilitation exercise programs is inadequate with just the analysis of cardiopulmonary signals from a resting patient. More high quality predictive data is produced with exercise stress testing as compared to resting physiology; physical exertion induces more useful physiologic signals. Therefore, effective implementation of CPET in cardiac rehabilitation and intervention programs can help optimize patient health outcomes (Guazzi, Bandera, et al. 2017). However, there exist numerous limitations to CPET interpretation and it is reportedly underutilized as a clinical tool (Neder et al. 2021).

1.1.2 Beyond CPET

Though CPET has long been the global standard, some researchers have been investigating more appropriate ways to capture health and fitness information in pediatric patients. (Gilliam et al. 1981; Bailey et al. 1995; Bar-Yoseph et al. 2022) propose that an alternative to CPET could be more suitable for younger patient populations. Among other key differences, it has been noted that gas exchange and ventilatory signals tend to show greater variation in children than adults (Potter et al. 1999). The present study utilizes a protocol termed "Multiple Brief Exercise Bouts" (MBEB) which follows the reasoning that natural patterns of physical activity in children are characterized by relatively short bursts (seconds to minutes) of exercise at various intensities interspersed with rest. By observing the same gas exchange and frequency variables as CPET over a more appropriate fitness test protocol, we hope to glean important physiological insights about square wave exercise cardiovascular dynamics in pediatric populations.

1.2 Research Questions

a. RQ1: Can we use machine learning techniques to accurately predict which individuals will quit exercise based solely on FDA of their cardio-respiratory signals? Can we make this prediction with reasonable accuracy after only four exercise bouts?

b. RQ2: To what extent do the machine learning techniques use gender, maturational status, and body mass to predict the physiological responses of children during multiple brief exercise bouts (MBEB)?

Chapter 2

Methods

2.1 Sample Data Set Description

The data used throughout this study includes demographic descriptors of 81 healthy children who participated in the MBEB protocol on a cycle ergometer at both high- and low-intensity workloads. All participants were screened and determined to be healthy based on interviews to identify any congenital or chronic diseases and conditions that would impair physiological responses to exercise. Extremely physically active participants (e.g., those considered to be elite athletes involved in routine intensive exercise training) were also excluded. The primary demographics under investigation were the participant's gender, maturational status (i.e. puberty level), and body mass. After an initial ramp test to determine individual anaerobic thresholds, the work rate was calculated for each participant as: low-intensity work rate, 80% of the lactate/anaerobic threshold (LAT); and high-intensity work rate, above the LAT and approximately equivalent to 80% of $\dot{V}O_{2peak}$.

Participants were assigned into the following eight groups: early pubertal females at 1) low work intensity MBEB and 2) high work intensity; early pubertal males at 3) low work intensity and 4) high work intensity; late pubertal females at 5) low work intensity and 6) high work intensity; and late pubertal males at 7) low work intensity and 8) high work intensity. The 81 children were described as:

- 17 early puberty females
- 25 late puberty females
- 20 early puberty males
- 19 late puberty males

The participants were asked to try to complete ten consecutive bouts of exercise. In this research, the MBEB protocol consisted of 2-minutes of constant work rate exercise at the individualized intensity (known as a square wave modality). Between exercise bouts, participants were instructed to rest for 1 minute, during which time they were asked to confirm their willingness to continue with the next bout. Measurements were taken breath-by-breath for each of the standard CPET signals. The protocol was approved by the UC Irvine Institutional Review Board. Informed consent was obtained from legally authorized guardians and, where appropriate, assent from the participants themselves.

After time-interpolation to achieve second-by-second data for every participant, the final set included 266,416 discrete observations of all demographic, frequency, and gas exchange variables. Participants were labeled 'quit = 0' if they successfully completed ten bouts at high intensity, and '1' if they quit during the course of MBEB (prior to bout 10 completion). Of the full participant set, 42 were labelled as 'non-quitter' and 39 were labelled as 'quitter'.

As a visual introduction to the data, Fig. 2.1 is a plot of the time series data for heart rate for all participants. See appendix A for the observations of other response variables.

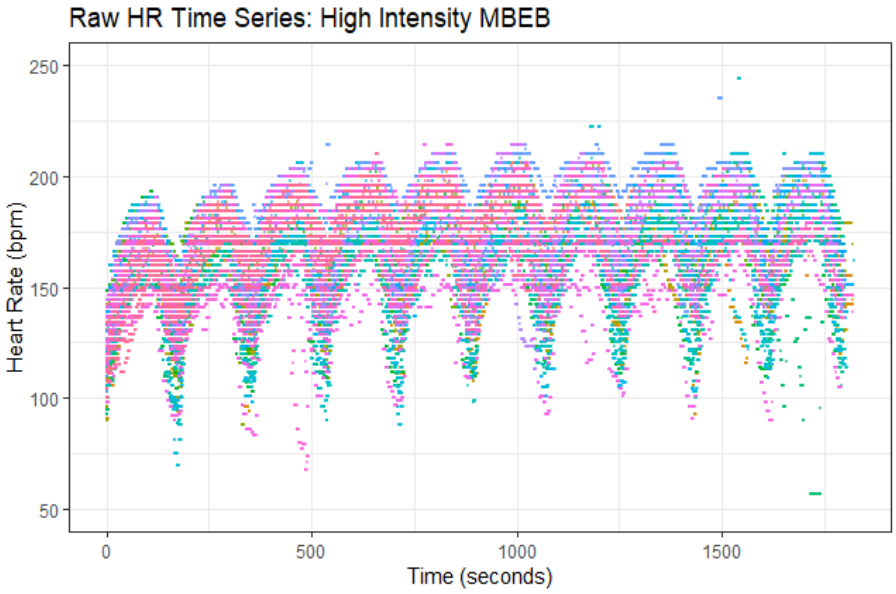


Figure 2.1: Ten bouts of Heart Rate, originally observed breath-by-breath and time interpolated to a second-by-second representation. Each participant's observations are shown with a unique color.

2.2 Data Analysis

The following sections describe four distinct approaches to answering our research questions with the given data. The first three approaches (section 2.2.1-2.2.3) provided key insight into the interacting variables and the dynamics of the cardiovascular system. Ultimately, they failed to address our research goals, but are described here to highlight their role in the research process. We found much greater utility by employing a method called functional data analysis, which is detailed in section 2.2.4.

2.2.1 Mixed Effects Multilevel Modeling (MLM)

As part of exploratory data analysis, we proposed that multilevel regression modeling (MLM) of mixed effects is a highly suitable approach for this class of hierarchical

analytical problems (Bar-Yoseph et al. 2022). According to Hox's textbook on the subject, "multilevel models are designed to analyze variables from different levels simultaneously, using a statistical model that properly includes the various dependencies" (Hox, Moerbeek, and van de Schoot 2017).

The MBEB data set used in this study can certainly be viewed as hierarchical; three explanatory variables (gender, maturational status, and consecutive bout number) exist at all levels of exercise intensity, and one single outcome variable (a physiological response) is measured repeatedly for each subject. Initial analysis revealed high levels of intra-class correlation, with clustering of individuals within puberty and gender groups and variables measured at two separate exercise intensities (treatments). Table 2.1 depicts the multi-level structure when considering the average response in each of the first five exercise bouts as a repeated measure.

Table 2.1: Multilevel Diagram in Table Format for the MBEB Study

Sub-index	**Level**	**Variables**
I (2)	Treatment	Work Intensity (High/Low)
J (4)	Population Group	Gender (Female/Male) Puberty Status (Early/Late)
K (5)	Child	Bout number (repeated measure)

Cross-level investigations were required to address our research hypotheses. The purpose for using a linear mixed effects model in this study was to effectively capture linear trends of the physiological responses from bout-to-bout, while allowing for between-subject variation through random slopes and intercepts.

Response variables were aggregated as the average measurement in each particular bout (Bout 1 through Bout 5). Through post-hoc statistical testing of the estimated

marginal means, we identified some differences between gender and puberty groups by the slope of their average physiological responses across exercise bouts. This linear univariate approach required us to create a separate model structure for each physiological variable and each phase of the exercise bout (HR_{on}, HR_{off}, RR_{on}, RR_{off}, etc.). We started by considering all possible combinations of the multi-level variables and their interactions, and iteratively removed variables that did not show statistical significance. As a result of this step-wise modeling process, each final model was fit with a unique combination of main and interaction effects; significance level of the coefficients were inconsistent across models. Fig. 2.2 is a table showing an example of the mixed effects model coefficients for heart rate. Fig. 2.3 is the corresponding plot of estimated predicted values for the heart rate mixed model.

The mixed effects modeling technique was useful for statistical inference but required too many separate models and effect interactions to draw useful conclusions. The restrictive nature of linear mixed effects models required us to simplify the response variable to an average value per bout; the true wave-like shape of an MBEB time series was lost in this process. Also, this method could not help us address our ultimate research goal: to determine if or when an individual would quit the MBEB event.

	HR (exercise)			HR (rest)		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	139.50	135.58 – 143.42	**<0.001**	137.01	132.55 – 141.47	**<0.001**
Gender [MALE]	0.80	-2.85 – 4.45	0.668	3.09	-1.66 – 7.84	0.203
Tanner [Late]	-3.95	-8.54 – 0.63	0.091	17.24	12.47 – 22.00	**<0.001**
bout	8.91	8.03 – 9.79	**<0.001**	4.24	3.72 – 4.76	**<0.001**
intensity [low]	-18.15	-21.80 – -14.49	**<0.001**	-33.50	-36.39 – -30.60	**<0.001**
Tanner [Late] * bout	2.65	1.46 – 3.84	**<0.001**			
Tanner [Late] * intensity [low]	-0.31	-5.25 – 4.63	0.901			
bout * intensity [low]	-5.36	-6.57 – -4.15	**<0.001**	-1.66	-2.39 – -0.94	**<0.001**
(Tanner [Late] * bout) * intensity [low]	-1.67	-3.30 – -0.03	**0.045**			
intensity [low] * Tanner [Late]				-10.55	-13.03 – -8.06	**<0.001**
intensity [low] * Gender [MALE]				-5.73	-8.21 – -3.25	**<0.001**
Random Effects						
σ^2	120.60			94.40		
τ_{00}	44.51_{ID}			89.59_{ID}		
τ_{11}	$0.28_{ID.bout}$			$0.10_{ID.bout}$		
ρ_{01}	1.00_{ID}			1.00_{ID}		
N	81_{ID}			81_{ID}		
Observations	965			959		
Marginal R^2 / Conditional R^2	0.796 / NA			0.865 / NA		

Figure 2.2: Example table of coefficient estimates for the mixed effects multilevel model that was fit to the bout average heart rate responses. This linear model is fit to each participant's average on-transient (left) and off-transient (right) heart rate response. 'Tanner' is the shorthand for 'Tanner score,' which describes the participant's puberty level. Note that the models for exercise and rest periods include different combinations of main and interaction effects and therefore have separate fit statistics (R^2).

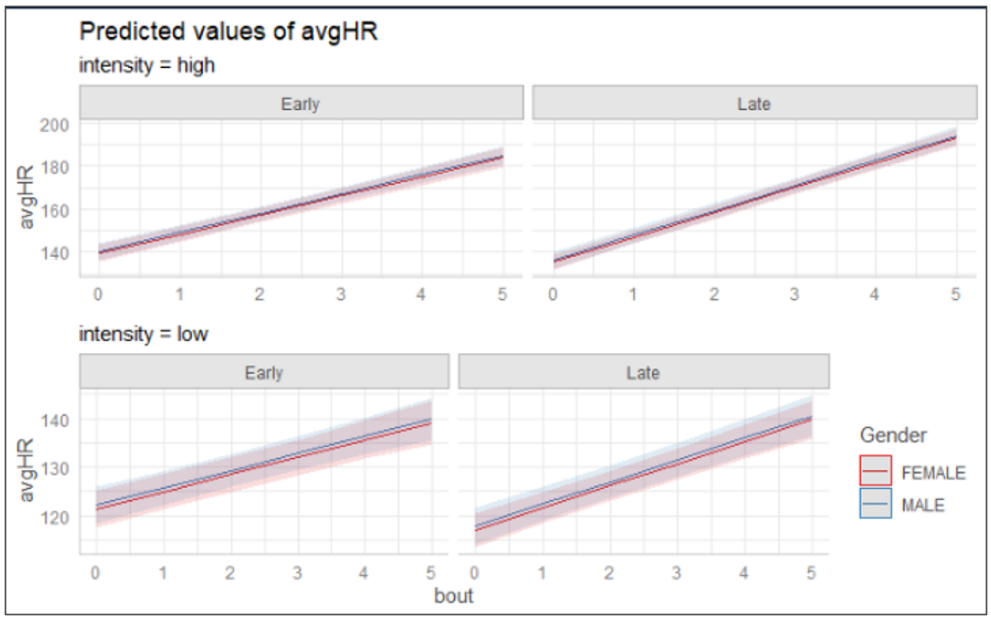


Figure 2.3: The estimated predicted values provide a representation of the linear relationship between time and the response variable (here, average HR). The slopes of these lines can be interpreted by considering the interaction of gender, puberty level, and work intensity. As expected, there is a positive linear trend as the exercise test proceeds through five bouts while the participants do not experience sufficient time to recover their HR. Unsurprisingly, this slope is more pronounced on the high-intensity MBEB test. Note that the shaded areas represent confidence intervals for the predictions, which are consistently overlapping.

2.2.2 Estimating the Time Constant (τ)

A second approach was to estimate the time constant (τ) of the exponential equation that could best fit the data during on-transient (exercise) and off-transient (rest) periods. The simplest equation commonly used to describe heart rate and gas exchange kinetics for on-transient exercise is:

$$Y_{t,on} = Y_{BL} + A(1 - e^{(t-TD)/\tau}) \tag{2.1}$$

where Y_t is the response variable, Y_{BL} is the baseline (starting point) of that variable, A is the maximum value (amplitude), t is the observation time, TD is a time delay, and τ is the time constant that generally describes the 'shape' of the curve. An inverse of this equation is used to describe the off-transient process:

$$Y_{t,off} = Y_{EX} + A(1 - e^{(t-TD)/\tau}) \tag{2.2}$$

where Y_{EX} is the response variable's value at the end of exercise (beginning of off-transient phase).

These functions were fit to each on- and off-transient phase of the available data. The time constant was estimated by iterative nonlinear least squares optimization. This time constant τ was averaged per individual, then became the variable of interest across exercise bouts, and we attempted to reveal differences between gender or puberty subgroups.

This area of research brought us to a better understanding of the various cardiovascular dynamics at play during MBEB but did not give sufficient insight as to

whether we could predict who would quit exercise before completing ten bouts. Additionally, though it worked reasonably well for heart rate and gas exchange metrics, this method was deemed to be an oversimplification of the dynamic processes for several response variables. In the current understanding of oxygen uptake kinetics, a time delay is incorporated for three distinct phases of exercise onset: phase 1, phase 2, and a phenomenon known as the 'slow component' (Stirling, M. Zakynthinaki, and Saltin 2005). This requires a more complex version of equations 2.1 and 2.2. Recent literature has proposed a set of nonlinear ordinary differential equations that enable more accurate estimation of instantaneous physiologic measurements (M. S. Zakynthinaki 2015, Mazzoleni et al. 2016). More research is required to evaluate these approaches in the case of MBEB-type exercise.

2.2.3 Signal Entropy for Classification

The third technique was to explore the variability in each participant's time series as an indicator of their respective fitness levels. Entropy, annotated by $\mathrm{H}(X)$ in the literature, is a measure of uncertainty of a random variable X. In a time series setting, it is the rate of information production (human breathing or heart rate, for example.) Physiologic or biological systems provide abundant opportunity for the study of entropy and time series volatility. (Costa and Ary L Goldberger 2002) posited that entropy increases with the degree of disorder and is at a maximum with completely random systems; furthermore, a lower degree of biological entropy can be associated with "diseased" physiologic signals.

Another concept which surfaces throughout the literature is that physiological systems (such as the human cardiovascular system) have multiple measures of entropy

that describe their ability to adapt to stress. When comparing healthy individuals to diseased or aging patients, studies generally agree that the greatest amount of generalized entropy exists amongst healthy individuals. This is indicative of the theory that, as humans age and become less "healthy," their systems lose robustness and cannot readily adapt (Costa and Ary L Goldberger 2002; Peng, Costa, and Ary L. Goldberger 2009). By analyzing "loss of complexity" as a feature, we could improve our diagnostic approaches for a wide class of diseases.

Acknowledging the complexity and inter-related nature of human biological systems, we understand that MBEB data is highly dimensional when considering the covariates at work during exercise. We investigated the predictive capability of entropy by reducing each MBEB time series to a single value of sample entropy (SampEn) for each signal that describes the trajectory's overall variability. For participants in this study, SampEn was calculated for three physiological response variables (HR, RR, $\dot{V}O_2$). Using the SampEn values from each participant and four demographic variables (gender, puberty level, body mass, and lean body mass), we attempted principal component analysis and other feature reduction techniques with logistic regression. Our goal was again to predict whether an individual would be correctly classified as a quitter, given their demographics and the SampEn values of their MBEB time series.

This procedure was inconclusive beyond the suggestion of which features appeared most important to classification (see Fig. 2.4). Fig. 2.5 shows a projection of the entropy data into two dimensions (the first two principal components). The 'quit' and 'not-quit' participant groups were virtually indistinguishable. We used entropy scores in a logistic regression procedure with elastic net regularization and cross-validation to predict the output class 'quit' or 'non-quit.' This regression performed at a maximum accuracy rate below 80% (see the bottom of Table 3.1 for results). We found this

predictive performance to be well below our best functional data models (described in Chapter 3), which consistently achieved accuracy scores above 90%. More research is required to understand the role of entropy in repeated exercise bout time series.

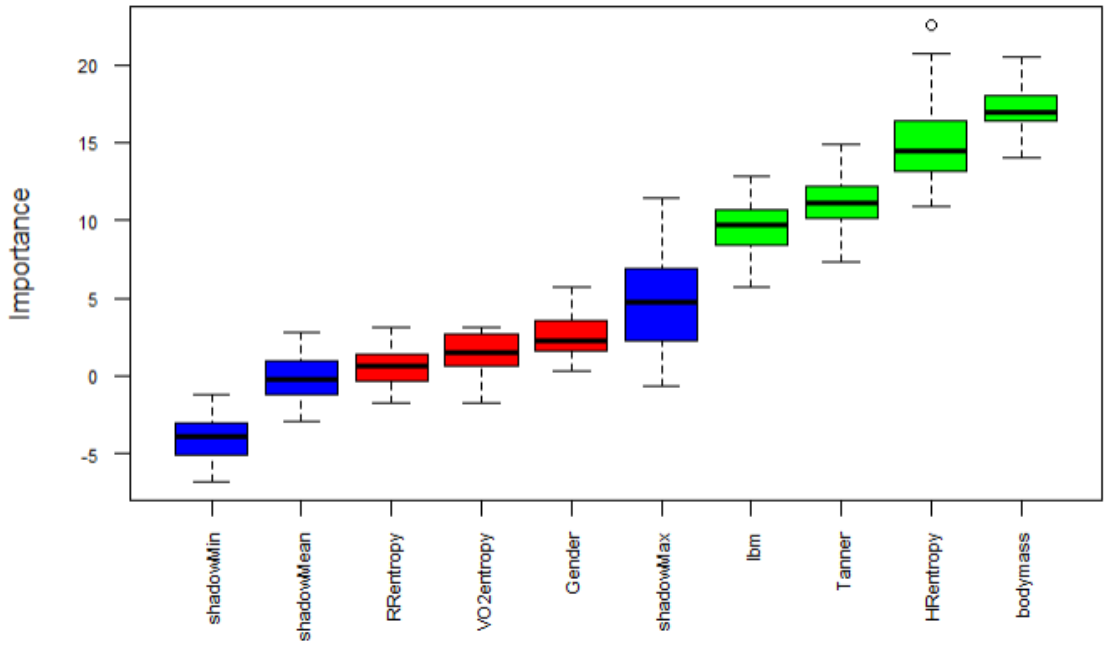


Figure 2.4: Feature importance plot generated by the *BORUTA* package in R, which employs recursive feature elimination. The HR entropy appeared important for classification, but entropy of RR and $\dot{V}O_2$ were far less predictive. One potential explanation is that heart rate measurements are generally not very noisy, thus the variability captured by SampEn is more informative to the underlying physiology. 'Tanner' refers to the Tanner scoring method (puberty status); the abbreviation 'lbm' stands for lean body mass.

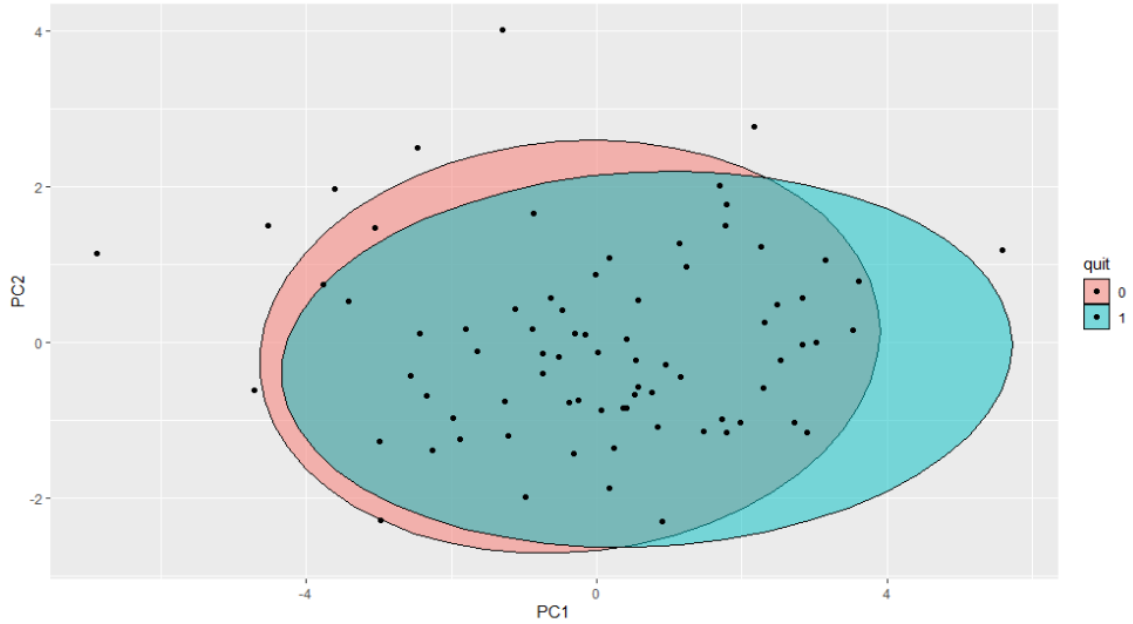


Figure 2.5: The polygons in this plot show the clustering of 'quitters' (blue) and 'non-quitters' (red) when projecting seven sample entropy and demographic features into the first two principal components. The two classes are virtually indistinguishable.

2.2.4 Functional Data Analysis

There is at least one common issue with each of the above data mining techniques for resolving our research questions. None of the proposed procedures provided conclusive evidence as to whether signals derived from MBEB can aid with fitness prediction or discovering differences between puberty and gender groups. Additionally, we seek to produce easily interpretable and physiologically sound results for medical providers. Therefore, we propose Functional Data Analysis (FDA), which is the focus for the remainder of this paper.

FDA is a highly flexible technique which can deal with non-independent and correlated repeated measures. The functional nature of CPET-derived observations encourages us to assume that the data are realizations of stochastic processes in continuous time. FDA's prominence has grown simultaneously with the emergence of electronic devices that accurately capture a continuous stream of physiological data; FDA can help leverage that data towards meaningful empirical conclusions.

A 2013 systematic review (Ullah and Finch 2013) provides a useful introduction to contemporary applications of functional data (FD). This type of analysis has been applied in a variety of time series experiments, particularly those with a biomedical context; FDA has proven powerful in analysis of human growth curves (James O. Ramsay and Silverman 2010), gait analysis (Røislien et al. 2009), fetal heart rate monitoring (Ratcliffe, Heller, and Leader 2002; Ratcliffe, Leader, and Heller 2002), and prediction of maximal oxygen consumption during exercise (Matabuena, Francisco-Fernández, and Cao 2017). Additionally, (Matabuena, Vidal, et al. 2019) proposed FDA to reduce predictive error in the estimation of maximum heart rate by avoiding the problems of high dimensionality and collinearity. A ramp exercise protocol

was used in that research, and the authors called for exploration into the predictive capacity of FDA with square wave exercise modalities.

When implementing FDA, data observations do not need to be equally spaced and missing observations are handled relatively well. Exceptionally noisy signals (such as respiratory rate in our data) benefit from the smoothing procedure, which is the key first step in FDA. The functional data (FD) itself can be more visually informative than the set of finite discrete observations and allows us to draw modeling and prediction information by applying multivariate statistical concepts. (J. O. Ramsay and Dalzell 1991) present several practical reasons for considering FD:

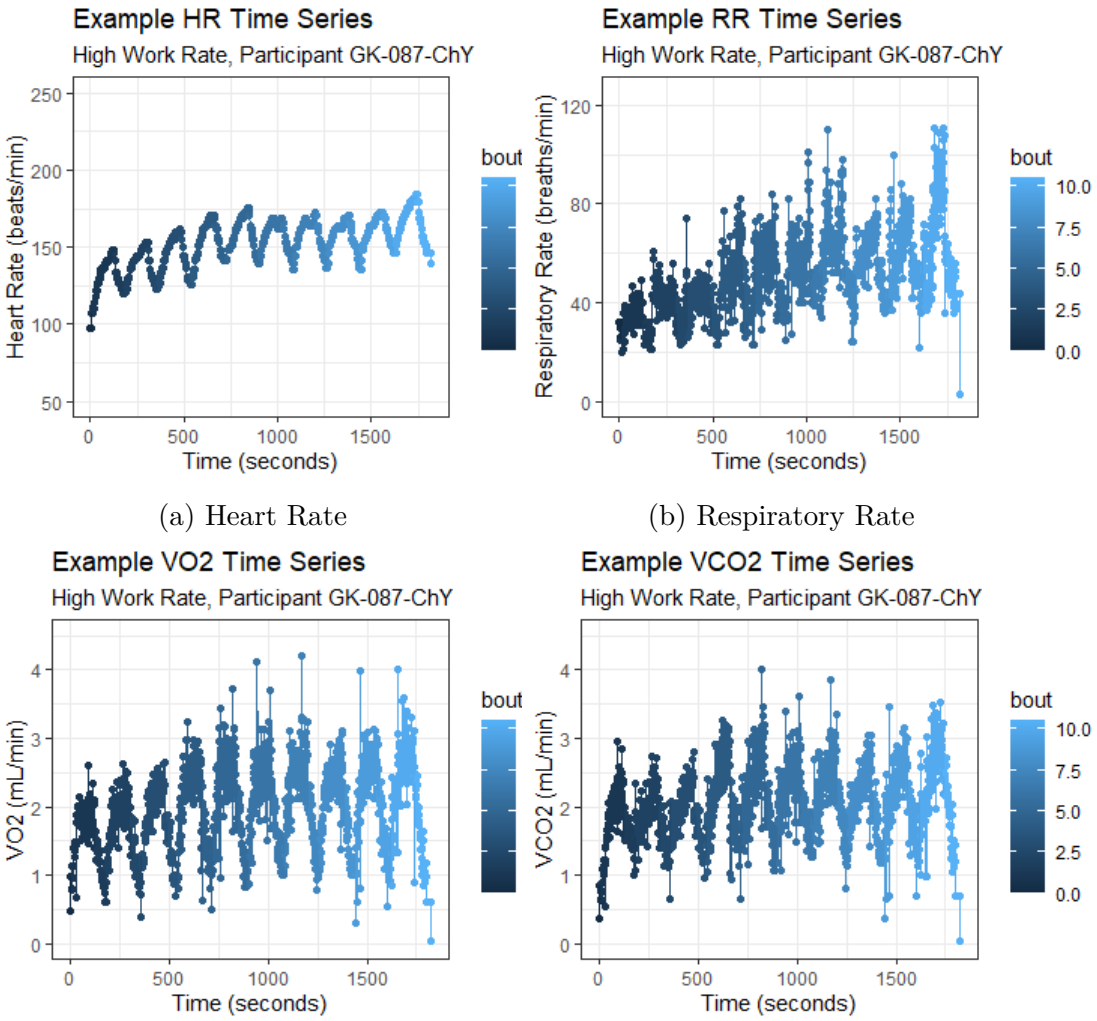
1. smoothing procedures can yield functional representations of a finite set of observations;
2. it is more natural to think through modeling problems in a functional form; and
3. the objectives of an analysis can be functional in nature, as would be the case if finite data were used to estimate an entire function, its derivatives, or the values of other functionals.

The time series measurements of our MBEB experiment are discrete and sometimes noisy observations of a continuous and dynamic process, therefore FDA seems highly appropriate. After transforming the breath-by-breath or second-by-second time series into a collection of smooth FD curves, we can apply supervised or unsupervised machine learning techniques to answer both research questions.

Data Conversion Procedure

The first step of functional data analysis was to convert the raw time series into functional data objects by choosing the appropriate basis transform and smoothing parameters. To address RQ2(b), we included only high-intensity test results for the first 720 seconds of MBEB. The purpose was to analyze only the first four bouts of MBEB, as all 81 participants completed a minimum of four bouts. All data mining and classification modeling was performed in Rstudio (Version 1.4.1103). FDA was conducted in R using the *'fda'* package (version 5.5.0) (J. O. Ramsay, Graves, and Hooker 2021) and the *'fda.usc'* package (version 2.0.2) (M. F. Febrero-Bande et al. 2020; M. Febrero-Bande and de la Fuente 2012).

The four variables of interest for our research question were heart rate (HR) (beats per minute), respiratory rate (RR) (breaths per minute), oxygen uptake rate ($\dot{V}O_2$) (mL/min), and carbon dioxide uptake rate ($\dot{V}CO_2$) (mL/min). The data was organized such that each of these response variables constituted its own independent time series. Through exploratory analysis, we determined that each of these signals has a distinct pattern characterized by variation and noise. Sample representations of each signal type are presented in Fig. 2.6. Plots of the full data set are available in Appendix A.



(a) Heart Rate

(b) Respiratory Rate

(c) Oxygen Uptake Rate

(d) Carbon Dioxide Uptake Rate

Figure 2.6: One participant's second-by-second signals for the full MBEB session. In general, Heart Rate (a) was the signal that contained the least noise in our data set; individual exercise bouts are very easily discerned. Respiratory Rate (b) was the signal that contained the most within- and between-subject variation in our data set; individual exercise bouts are difficult to discern.

A B-spline basis representation was determined to provide an excellent fit to each of the four time series. The splines were generated using 725 total basis elements of 6^{th} order B-splines, with internal knots corresponding to the start- and stop-exercise time points. The smoothing procedure was controlled by a roughness penalty, which resulted in reasonably smooth functions without wild variations in the approximating function. Penalized smoothing was done by applying harmonic acceleration operators to the functional data and searching across values of λ (smoothing parameter) until an acceptable generalized cross-validation (GCV) error level and degrees-of-freedom (DoF) were reached in the smoothed estimate. In other words, each of the response curve sets were deemed appropriately smooth for this particular application. This process is introduced in Chapter 5 of (J. O. Ramsay, Hooker, and Graves 2009). Fig. 2.7 explains this procedure visually.

Three individuals were removed due to irregularities in their time series (likely the result of error during data collection). This left 78 sample curves for analysis. The final ratio of quitters to non-quitters was 39:39.

After transforming the discrete observations to FD objects, we considered the crucial step of curve registration. Registration is important because it allows us to align the curves (by time warping or otherwise) and remove phase variation while maintaining the amplitude variation (Kokoszka and Reimherr 2021). In other words, we transform the 'clock time' of our functional data to a standardized 'system time'. (Marron et al. 2015) presents the definitive foundation for the necessity of registration in misaligned data sets. For our data set, automated continuous registration resulted in minimal phase shifting, as the original time series were nearly perfectly aligned by nature of the testing protocol; exercise bouts began and terminated near the same moment in time for all participants. We measured the proportion of total variation

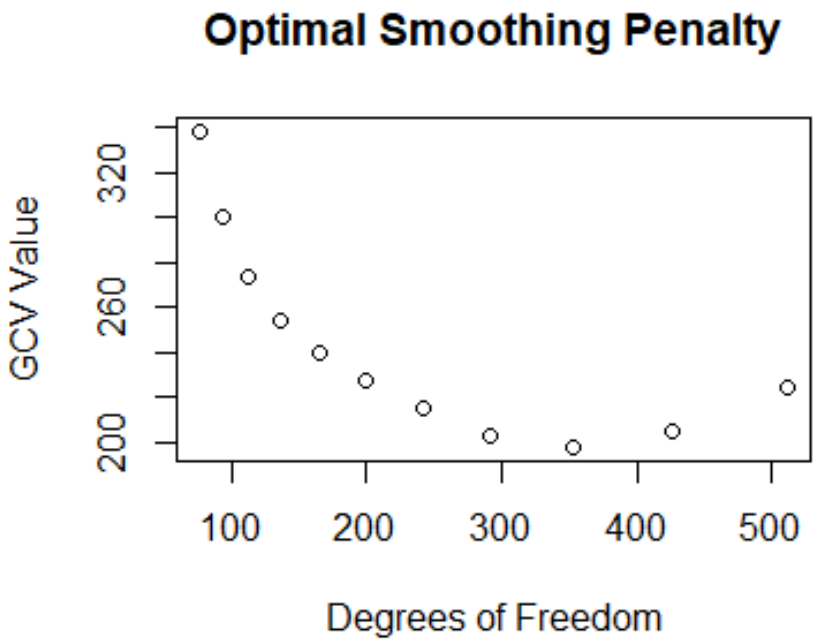


Figure 2.7: Example estimation of the smoothing parameter λ. An appropriate level of smoothing was determined by visual inspection of the relationship between GCV and DoF in the smoothed model, after applying a harmonic acceleration operator that penalizes departures from a shifted sine. This procedure is explained in depth in (J. O. Ramsay, Hooker, and Graves 2009). This figure shows a minimal GCV when the model contains 350 DoF, which corresponds to a λ near 200. Thus, 200 was chosen as the smoothing penalty for the set of HR curves, and the fit was validated after visual inspection of the smoothness (see Fig. 2.8). This process was repeated for RR, $\dot{V}O_2$, and $\dot{V}CO_2$.

due to phase variation (MSE_{phase}/MSE_{total}) as 9%. The registered curves (with phase variance removed) were utilized for all subsequent analysis.

Fig. 2.8 depicts the smoothed and registered heart rate FD objects as an example. Plots of the RR, $\dot{V}O_2$, and $\dot{V}CO_2$ functions are available in Appendix B.

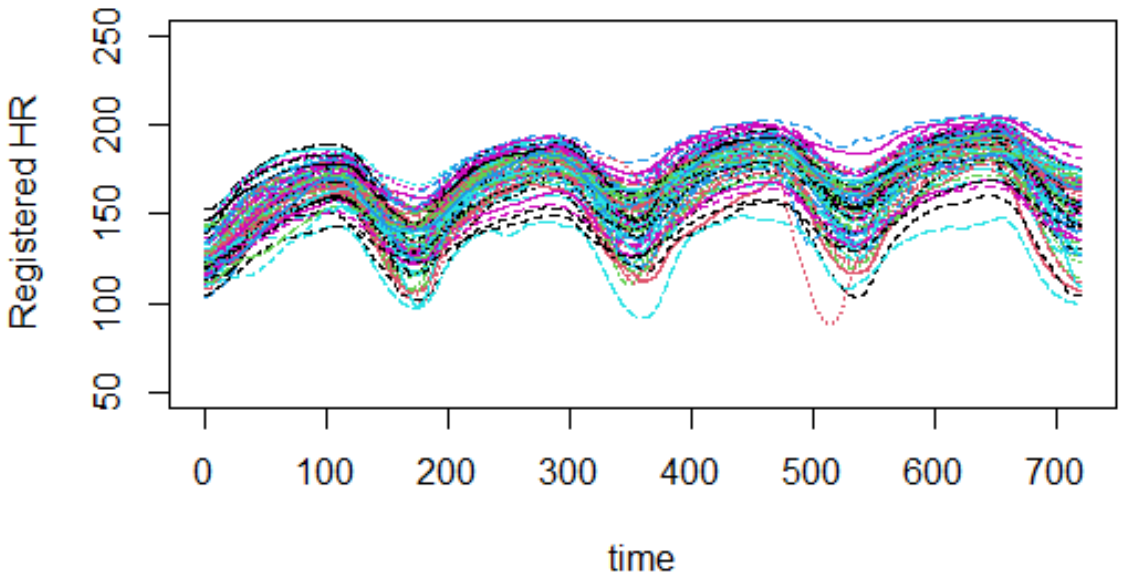
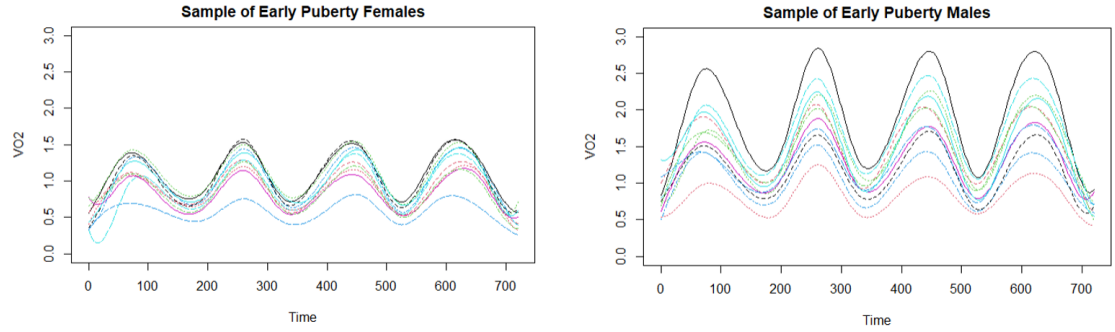


Figure 2.8: Four bouts of Heart Rate (one of the four MBEB signals) after converting the discrete time series to 78 smoothed and continuously registered functional data objects. Each participant's time series is represented as an individually colored function. The smoothing procedure was controlled by a roughness penalty, which resulted in reasonably smooth functions without wild variations in the approximating function. In other words, the curves were deemed appropriately smooth for this particular application. The FD was then continuously registered to remove phase variation.

Functional *t*-Tests

To address Research Question 2a (RQ2a), we investigated the null hypothesis (H_0) that there exists no statistically significant difference in the functional means of participants in contrasting sub-groups. To test for differences between *gender groups*, we sampled 11 each of males and females at the early-puberty level, to which we applied permutation t-Tests on their functional means. To test for differences between *puberty groups*, we compared 11 samples of early puberty males vs. late puberty males. For

time periods where the t-statistic exceeded the critical value (0.05), we could reject (H_0). Interestingly, this procedure revealed distinct puberty and gender differences throughout the time series of $\dot{V}O_2$ and $\dot{V}CO_2$, but no such significant mean functional difference existed for HR and RR. Fig. 2.9 and 2.10 highlight one result of this experiment.



(a) Early pubertal females (b) Early pubertal males

Figure 2.9: A sample of 11 early pubertal females' (a) and 11 early pubertal males (b) $\dot{V}O_2$ functional data. The purpose of creating these random samples was to conduct statistical testing on the functional means of two sub-groups; in this example, we hold puberty level constant to test the effects of gender on the response variable, $\dot{V}O_2$.

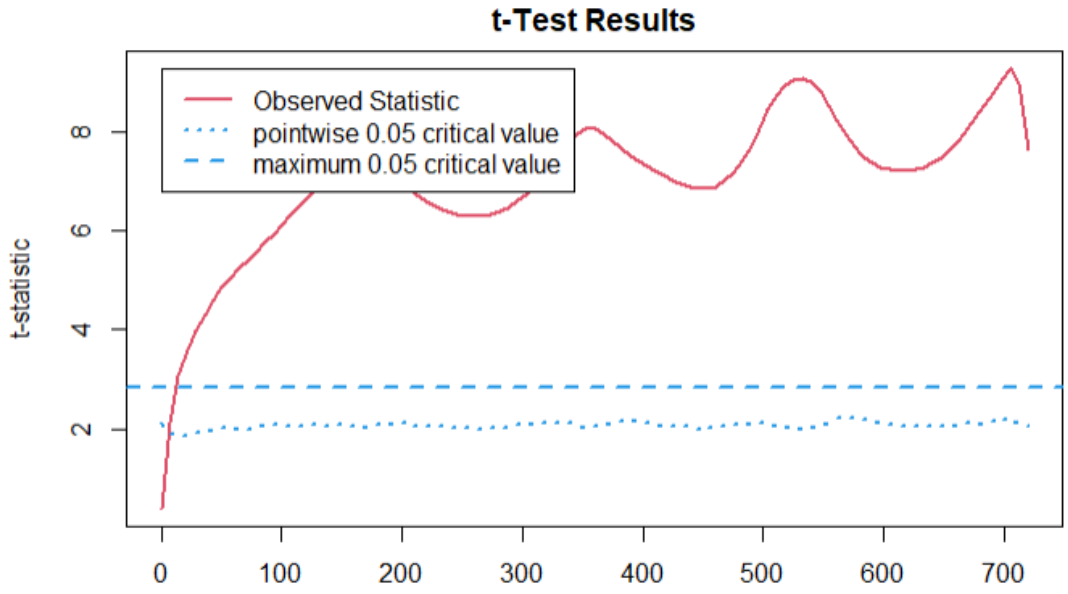


Figure 2.10: Visual output of the functional permutation t-test between Early- and Late-puberty males. This particular chart shows that, for $\dot{V}O_2$, a statistically significant difference in the means exists between our random sample of early puberty males and late puberty males. The t-statistic is calculated to be greater than the critical value across the entire time interval. This could signify a fundamental difference in the physiology between puberty groups when holding gender status constant.

Outlier Detection

Functional depth is a measure of centrality for a given curve within a group of trajectories. (Febrero, Galeano, and González-Manteiga 2008). Several approaches are covered in the literature, and we chose to explore one of these methods. Fraiman and Muniz introduced functional data depth in 2001, which allows for the ranking of distributed functional curves by their Fraiman-Muniz (FM) depth (Fraiman and Muniz 2001). This center-outward ranking of FD curves can be interpreted as the degree of "outlyingness" for a particular curve of interest.

Outliers may occur in a functional dataset by gross errors such as measurement mistakes, but often times they are not the result of gross error but perhaps indicate a noteworthy difference in patterns from the rest of the curves. In other words, outliers in our dataset could be considered curves that are not compatible with the assumption that the entire FD set is drawn from the same stochastic process. The outliers appear dissimilar in either their overall shape or their distance from the expected central function of the sample at some or all intervals of interest. The outlier detection algorithm used here weighs the data according to depth after a bootstrap smoothing function is applied to the relevant curves.

By assigning depth scores to our functional data, we could identify (and remove) outlier participants. Outliers in this sense would be those whose overall time series deviates furthest from the centrality measure – whether it be the mean, median, mode, or otherwise. This could be beneficial in classifying new data, allowing researchers to determine how well a particular participant's signal curve fits with the known central function in our existing MBEB dataset. Additionally, we could infer that a newly introduced trajectory with relatively high depth will be classified accurately given the

model structures presented in this research.

For the results presented in this paper, outlier curves were visually inspected but not removed from the classification models. Fig. 2.11 is the visual output of one outlier detection procedure.

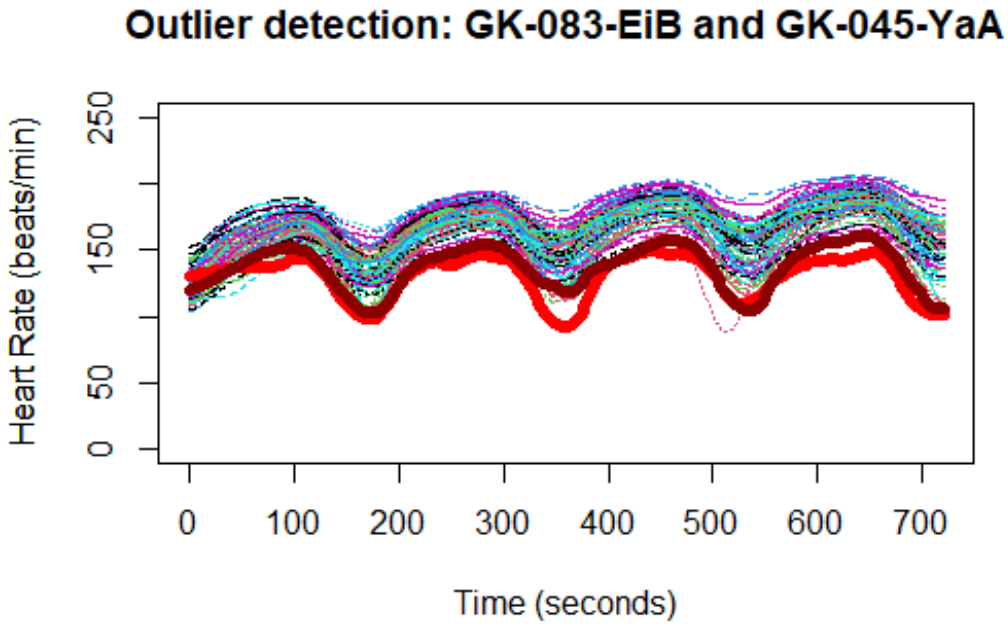


Figure 2.11: Example of outlier curves in the heart rate functional data set. The two curves (bold red) were identified as potential outliers based on their low Fraiman-Muniz depth ranking amongst the 78 participants.

Supervised Functional Classification

Research Question 2 called for investigation into FDA's ability to discriminate between classes of quitters and non-quitters. We applied various classification models to the FD object set. The goal was to find a classifier with the minimum error rate. Our first approach was to predict 'quit' status from combinations of the demographic and functional data covariates. Demographic variables were gender (binary), puberty

level (binary), and body mass (continuous).

The flexible nature of FDA allowed us to test seven unique classification models: generalized spectral additive models (GSAM), linear discriminant analysis (LDA), recursive partitioning and regression trees (RPART), RandomForest (RF), support vector machines (SVM), neural network (NNet), and k-Nearest Neighbors (KNN). Tuning parameters were optimized for each classifier/response variable combination; for example, the number of k-neighbors that produced the lowest error was accepted as the final parameter. The probability value for binary discriminant (i.e. classification threshold) was optimized within each GSAM model; we searched across a range between 0.3 and 0.8, and the threshold which produced the highest F1 score was selected. Ten-fold cross-validation was built into each classification model.

Functional representation of HR alone was the first FD covariate we tested: $quit = s(HR_{[0,720]})$ where $HR_{[0,720]}$ is the smoothed HR function over the first four bouts. The function $s(\cdot)$ denotes an additive effect over the variable. After this approach proved fruitful on the cleanest physiological signal, we applied the classifiers to RR, $\dot{V}O_2$, and $\dot{V}CO_2$ FD objects with the same model parameters. This allowed us to compute model performance and compare the results side-by-side. Overall model accuracy was calculated as the number of correct classifications divided by the total number of attempts. The F1 score was computed as $(2*(precision*recall)/(precision+recall))$. Finally, we combined all demographic and functional covariates for HR, RR, $\dot{V}O_2$, and $\dot{V}CO_2$ into a 'full' multivariate model and tested the classification rate. The dteailed structure of each model is described in Appendix C.

The final step in answering RQ2 was to conduct one-way functional analysis of variance (FANOVA) over our 78 independent samples. The purpose was to empirically determine whether 'quitters' and 'non-quitters' display differences in their functional

means. The HR, RR, $\dot{V}O_2$, and $\dot{V}CO_2$ data were bootstrap resampled 500 times, plotted, and analyzed. A p-value ≤ 0.05 was considered significant to reject the null hypothesis (H_0) of equality of mean functions between participants labelled 'quit' and 'not quit.'

Chapter 3

Results

The first section of this chapter describes the results of our classification technique using FDA on four physiological signals. Section 2 presents the results of FANOVA.

3.1 Classification using FDA

The results of the ten best performing models for our physiological response variables are presented in Table 3.1. All models performed better when the continuous variable 'body mass' was omitted. The GSAM structure generally performed best among the tested classifiers. The highest F1 score (93.5%) for classifying quitters and non-quitters was achieved using smoothed functional data *and* demographic covariates as predictors in a GSAM. Providing functional data alone (with no demographic covariates) resulted in a maximum classification F1 score of 91.1%. In Table 3.1, the bottom row contains the maximum performance achieved using sample entropy as a predictor variable with elastic net logistic regression; this is provided for purposes of comparison.

After testing each individual cardiovascular signal, we constructed a 'full' model. This model used all functional data of HR, RR, $\dot{V}O_2$, and $\dot{V}CO_2$ together, along with gender, puberty level, and body mass. The results are shown in Table 3.2. The 'Full GSAM' model performed best (F1 score 93.5%, accuracy 93.6%).

Table 3.1: Individual Classification Model Performance

Model	Accuracy (%)	F1 Score (%)
$\dot{V}O_2$ GSAM + Covariates	93.6	93.5
$\dot{V}O_2$ GSAM	91.0	91.1
$\dot{V}CO_2$ GSAM	87.2	87.5
$\dot{V}CO_2$ GSAM + Covariates	87.2	86.8
$\dot{V}O_2$ NNet	84.6	83.8
HR GSAM + Covariates	82.1	82.1
HR GSAM	78.2	80.5
HR SVM	79.5	80.0
$\dot{V}CO_2$ NNet	80.8	80.0
HR LDA	79.5	79.5
Logistic Regression with entropy values	72.2	73.7

Table 3.2: Full Classification Model Performance

Model	Accuracy (%)	F1 Score (%)
Full GSAM	93.6	93.5
Full LDA	87.2	87.2
Full RPART	83.3	84.0
Full SVM	77.0	78.0
Full NNet	70.5	72.3
Full KNN	65.4	69.0
Full RandomForest	66.7	66.7

3.2 Functional Analysis of Variance

The statistic of interest for drawing conclusions from FANOVA was the probability of a true difference in functional means over the bootstrapped observations. A p-value ≤ 0.05 indicated that we could reject (H_0) and conclude that a significant difference in functional means was present.

Fig. 3.1 is the visual depiction of heart rate functional means for quitters and non-quitters and compares the estimated HR curves after bootstrap resampling. Detailed FANOVA results for RR, $\dot{V}O_2$, and $\dot{V}CO_2$ are included in Appendix D. Table 3.3 shows the resulting p-values and conclusions from FANOVA. To summarize, we found that children in the 'quit' and 'not-quit' groups have significantly different functional means for three signals: heart rate, oxygen uptake rate, and carbon dioxide uptake rate. Each of these variables display higher mean functions across the four bouts for those who ultimately quit the MBEB session.

Table 3.3: FANOVA Results

Response Variable	p-value	Conclusion
Heart Rate	0.000	S.S. difference in means
Respiratory Rate	0.186	not S.S. difference in means
$\dot{V}O_2$	0.000	S.S. difference in means
$\dot{V}CO_2$	0.000	S.S. difference in means

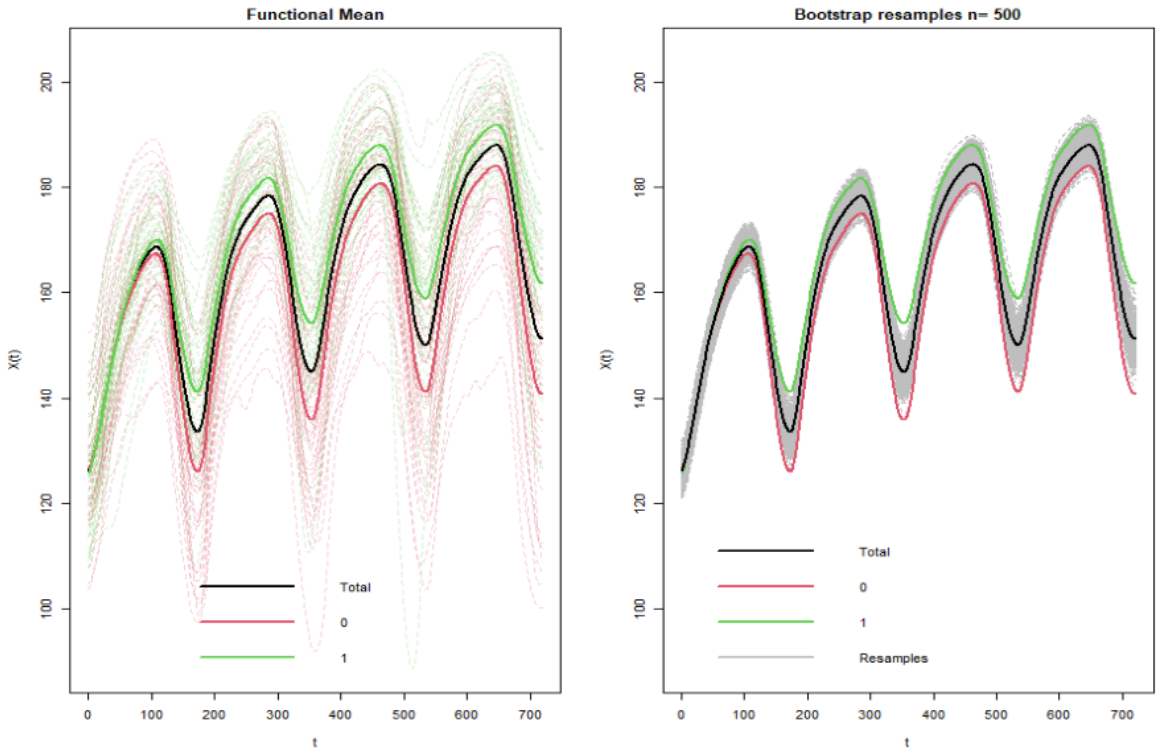


Figure 3.1: Comparison of functional means for the Heart Rate signal [X(t)] during the first four exercise bouts. Participants who quit exercise during MBEB are labelled as '1' and colored green. The red line depicts the functional mean for non-quitters. The black line indicates the mean trajectory for all 78 participants. The plot on the right shows Heart Rate curves for MBEB quitters (green) and non-quitters (red) with 500 bootstrap resamples. The black line represents the bootstrapped mean function.

Chapter 4

Discussion

4.1 Physiological Implications

The differences that we found between gender and puberty subgroups are generally in agreement with historical findings. There is evidence, for example, that healthy, early pubertal children have substantially faster HR and $\dot{V}CO_2$ exercise responses than healthy late-pubertal or adult individuals (Baraldi et al. 1991; Dan M. Cooper, Kaplan, et al. 1987). $\dot{V}O_2$ kinetics appear to be less dependent on puberty status, but children typically have higher oxygen uptake per work performed than do late pubertal or adult individuals (Zanconato, D. M. Cooper, and Armon 1991; Armon et al. 1991). These differences were identified via t-test after our second-by-second observations were transformed into functional data. Statistically significant functional differences between males and females were more difficult to discern in our data set and require further study. Additional analysis is also necessary to confirm that the conclusions in this research are valid for the low-intensity exercise setting.

4.2 Functional Data Analysis as a Strategy for Processing Exercise Time Series

Recall that RQ1 asked about FDA's applicability in the exercise data environment. Based on model performance alone, FDA seems to be a highly useful systematic strategy for processing exercise-induced physiological signals. By transforming the raw data into appropriately smoothed functions, the standard multivariate statistical approaches are quite useful for understanding the underlying physiology and highlighting differences between population subgroups. In addition to the promising predictive capability we presented here, the general benefits of FDA were apparent. As exercise response signals are inherently noisy and non-linear, exploration of the data as smoothed functions was instrumental in our statistical analysis.

Many conventional statistical techniques are useful for "ramp" style exercise time series, as the on- and off-transient structure does not exist; participants exercise at a progressively increasing work rate until exhaustion. However, these methods struggle to capture the patterns when considering multiple repeated exercise intervals which are found in MBEB. In comparison to FDA, other statistical procedures are more sensitive to the assumption of independence between repeated measurements. (In particular, cardiovascular time series observations are highly dependent on previous measurements by the nature of the underlying processes.) Additionally, some techniques fail to utilize all information present in the response variable during repeated bouts; for example, the multi-level models described in Section 2.2.1 are only useful if we simplify the time series into average values of each bout. Whereas FDA consistently resulted in accuracy and F1 greater than 90%, the other classification technique using signal entropy (section 2.2.3) was not able to predict the 'quit' group

with an accuracy better than 80%; this may be explained by a loss of information when converting from the raw signal.

FDA allows for similar handling of sparse datasets and those in which the exercise perturbance intervals of the given protocol are not so cleanly aligned. (Matabuena, Vidal, et al. 2019) argues that FDA's ability to reduce predictive error could be beneficial for clinical practitioners and exercise prescription, especially in settings where a maximal stress test is not feasible. We validated the shorter-duration exercise test as a viable alternative. FDA is useful even when the exercise and rest durations are inconsistent between participants.

Our results show that the FDA technique provides highly interpretable results for the clinician. Clear patterns emerge after transforming the noisy, non-linear discrete observations into smooth functions. An appropriate amount of signal noise and phase variation are reduced in the process, and the interpretation of these functions is straightforward. Additionally, we contributed a method of data representation that does not require substantial understanding of "black box" machine learning techniques. This benefits the clinician as an end-user of our functional data processing and allows for a wide variety of follow-on analyses (see Chapter 5). As an example, suppose we were interested in the trajectory differences between healthy patients and those with a chronic disease (instead of predicting which participants quit exercise). The graphical depictions of sub-group mean functions (Fig. 3.1 and Appendix D) can aid a clinician with determining whether a patient's trajectory more closely aligns with that of a healthy or non-healthy subject.

We presented a number of conclusions about our sample data set to address RQ2. One-way functional ANOVA answered our statistical hypotheses about equivalence of means in the two groups; this was a robust test of functional variance across the

exercise test. We found that children who quit during MBEB were characterized by a statistically significant higher functional mean heart rate, $\dot{V}O_2$, and $\dot{V}CO_2$ as compared to non-quitters across the four bouts.

Our models included gender, maturational status, and body mass as scalar covariates alongside functional MBEB signals to predict which children would quit exercise voluntarily. Oxygen uptake, carbon dioxide uptake, and heart rate were especially informative signals for predicting quitters based on the first four exercise bouts. Incorporating gender and puberty level was beneficial for several models. The top performing model ('$\dot{V}O_2$ GSAM + Covariates') classified quitters with 93.5% F1 score; by adding the demographic features to the functional covariate, we improved the classification rate by several points. We also showed the ability to sample from subgroups and conduct permutation t-tests of the functional means, testing for differences between gender and maturational status. This particular comparative method is more challenging with discrete data, as evidenced by our difficulties in implementing MLMs (section 2.2.1) and estimating time constant changes across bouts (2.2.2). These modeling techniques require more careful consideration of the confounding and interaction effects of demographic variables.

To assess our methodology, we compared the performance to separate baseline studies of functional data in the field of medicine. In 2002, fetal heart rates were studied to predict the risk categories of infants at birth (Ratcliffe, Heller, and Leader 2002). This research achieved 94% correct predictions using logistic regression, but the parameter estimates were based on the same data on which predictions were made. The first paper to utilize FDA in the prediction of maximal heart rate during exercise was published in 2019 (Matabuena, Vidal, et al. 2019). This study employed machine learning for "ramp"-style exercise, and measured the utility of FDA by comparing

predictive error of several models. Functional regression produced a root mean square error that was lower than each of the non-functional regression methods. The accuracy reported in that paper is not directly comparable to our functional classification metrics, but conclusively demonstrated FDA's efficacy for processing cardiovascular signals.

Chapter 5

Future Work

FDA is currently a very active research topic. With the FDA approach being validated on this small sample of youth during MBEB, there is plenty of opportunity for further research. First, there exist other important frequency and gas exchange variables as output from CPET and MBEB; work output (watts), minute ventilation ($\dot{V}E$), respiratory quotient (RQ), and the ratio of $\dot{V}E$ to $\dot{V}CO_2$ ($\dot{V}E/\dot{V}CO_2$ slope) could be handled with FDA and may prove useful for diagnostic purposes. Specific cardiovascular conditions may be identified by transforming all 15 time series depicted in the 9-panel plot (Fig. 1.2) into functional objects. Furthermore, some physiological signals are correlated with body mass; it would be interesting and prudent to test theories about the dynamics of gas exchange variables while specifically accounting for (i.e. normalizing by) body mass or lean body mass.

The binary classification methods used here assume that whether or not a child quits an exercise test is an appropriate proxy for his or her physical fitness. Though outside our scope, there are undoubtedly other factors at play when a child makes the decision to quit during intense exercise.

In conversion from discrete to functional data, the selection of smoothing parameters and basis representation are subjective. A complex B-spline basis was chosen for this dataset due to the popularity and flexible nature of splines as well as the ability to capture the on- and off-transient signal patterns that resulted from MBEB. Other

basis transformations should be investigated for their goodness of fit on a given data set. The R packages used in this research are capable of creating Fourier expansions, regression splines, and kernel smoothing bases (among others). The classic multivariate procedure of principal component analysis can be useful in the initial step of converting data to functional form and producing a reasonable summary of the data (J. O. Ramsay, Hooker, and Graves 2009). Regression analysis of the functional principal components themselves should be investigated further. FDA provides for functional clustering if the researcher is interested in non-supervised assignment of signal curves into similar groups. Functional regression with a functional response is also possible, whereby one could model the expected functional response of a physiological signal from the known function of another signal (Beyaztas and Shang 2020). A reasonable implementation would be to estimate gas exchange variable functions from the participant's observed work output time series.

Classification was conducted with the full set of 78 functional curves and the corresponding demographic variables. Outlier analysis revealed that some curves may not fit well with the population's distribution, thereby possibly reducing classification accuracy. Machine learning models may benefit from fine-tuning the outlier detection procedure and removing a percentile of low-depth curves.

Additionally, we hope to validate our methodology with other population groups, such as pediatric patients with chronic diseases or obesity, or a group of young adults. Our participant pool, while homogeneous and reflective of the local community at the site, was not representative of the population as a whole and further studies will be necessary to gauge the effect of racial, ethnic, and other social determinants on exercise responses as children grow and develop. The data analyzed in this paper may be useful as a baseline to which we can compare the signals of diseased individuals. If

new data is converted to functional form with the parameters used here, it would be possible to measure the relative depth score of the new curves. One might infer that the new curves would perform well in quitter prediction if the functions demonstrate high depth scores (i.e. they are not outliers from the data utilized in this research).

A natural extension of FDA is forecasting via stochastic methods, which seems highly relevant to exercise testing. By converting discrete data to functional data, we have transitioned to a continuous space where the functions themselves can be tested in terms of their forecasting error on unseen observations at future time intervals. This has been used in the context of mortality rate and pollution rate prediction (see chapter 8 of Kokoszka and Reimherr 2021). Forecasting models could be used to predict a patient's output in future bouts for any of the measured variables, to include work rate (watts). If feasible, this could mean that MBEB data sets are useful in assessing fitness with only four exercise bouts (or fewer). It should be explored further to determine if there is a minimum number of consecutive exercise bouts that can be transformed to FD and used effectively for fitness classification. This would be of great value in the medical community, especially for fitness testing in symptom-limited patients who are unable to maintain continuous exercise for more than a few brief bouts.

Bibliography

Gilliam, Thomas B. et al. (1981). "Physical Activity Patterns Determined by Heart Rate Monitoring in 6–7 Year-Old Children". In: *Medicine & Science in Sports & Exercise* 13.1, pp. 65–67. ISSN: 0195-9131.

Cooper, Dan M., Martin R. Kaplan, et al. (June 1987). "Coupling of Ventilation and CO2 Production during Exercise in Children". In: *Pediatric Research* 21.6, pp. 568–572. ISSN: 1530-0447. DOI: 10.1203/00006450-198706000-00012.

Armon, Y. et al. (Feb. 1991). "Oxygen Uptake Dynamics during High-Intensity Exercise in Children and Adults". In: *Journal of Applied Physiology* 70.2, pp. 841–848. ISSN: 8750-7587, 1522-1601. DOI: 10.1152/jappl.1991.70.2.841.

Baraldi, Eugenio et al. (June 1991). "Heart Rate Recovery from 1 Minute of Exercise in Children and Adults". In: *Pediatric Research* 29.6, pp. 575–579. ISSN: 0031-3998, 1530-0447. DOI: 10.1203/00006450-199106010-00011.

Ramsay, J. O. and C. J. Dalzell (1991). "Some Tools for Functional Data Analysis". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 53.3, pp. 539–561. ISSN: 2517-6161. DOI: 10.1111/j.2517-6161.1991.tb01844.x.

Zanconato, S., D. M. Cooper, and Y. Armon (Sept. 1991). "Oxygen Cost and Oxygen Uptake Dynamics and Recovery with 1 Min of Exercise in Children and Adults". In: *Journal of Applied Physiology* 71.3, pp. 993–998. ISSN: 8750-7587, 1522-1601. DOI: 10.1152/jappl.1991.71.3.993.

Bailey, Robert C. et al. (July 1995). "The Level and Tempo of Children's Physical Activities: An Observational Study". In: *Medicine & Science in Sports & Exercise* 27.7, pp. 1033–1041. ISSN: 0195-9131. DOI: 10.1249/00005768-199507000-00012.

Potter, C. R. et al. (June 1999). "Breath-to-Breath "Noise" in the Ventilatory and Gas Exchange Responses of Children to Exercise". In: *European Journal of Applied Physiology and Occupational Physiology* 80.2, pp. 118–124. ISSN: 0301-5548. DOI: 10.1007/s004210050567.

Fraiman, Ricardo and Graciela Muniz (Dec. 2001). "Trimmed Means for Functional Data". In: *Test* 10.2, pp. 419–440. ISSN: 1133-0686, 1863-8260. DOI: 10.1007/BF02595706.

Costa, Madalena and Ary L Goldberger (2002). "Multiscale Entropy Analysis of Complex Physiologic Time Series". In: *PHYSICAL REVIEW LETTERS* 89.6, p. 4.

Ratcliffe, Sarah J., Gillian Z. Heller, and Leo R. Leader (2002). "Functional Data Analysis with Application to Periodically Stimulated Foetal Heart Rate Data. II: Functional Logistic Regression". In: *Statistics in Medicine* 21.8, pp. 1115–1127. ISSN: 1097-0258. DOI: 10.1002/sim.1068.

Ratcliffe, Sarah J., Leo R. Leader, and Gillian Z. Heller (Apr. 2002). "Functional Data Analysis with Application to Periodically Stimulated Foetal Heart Rate Data. I: Functional Regression". In: *Statistics in Medicine* 21.8, pp. 1103–1114. ISSN: 0277-6715. DOI: 10.1002/sim.1067.

"ATS/ACCP Statement on Cardiopulmonary Exercise Testing" (Jan. 2003). In: *American Journal of Respiratory and Critical Care Medicine* 167.2, pp. 211–277. ISSN: 1073-449X, 1535-4970. DOI: 10.1164/rccm.167.2.211.

Stirling, J, M Zakynthinaki, and B Saltin (Sept. 2005). "A Model of Oxygen Uptake Kinetics in Response to Exercise: Including a Means of Calculating Oxygen Demand/Deficit/Debt". In: *Bulletin of Mathematical Biology* 67.5, pp. 989–1015. ISSN: 00928240. DOI: 10.1016/j.bulm.2004.12.005.

Guazzi, Marco, Rosa Raimondo, et al. (July 2007). "Exercise Oscillatory Ventilation May Predict Sudden Cardiac Death in Heart Failure Patients". In: *Journal of the*

American College of Cardiology 50.4, pp. 299–308. ISSN: 07351097. DOI: 10.1016/j.jacc.2007.03.042.

Febrero, Manuel, Pedro Galeano, and Wenceslao González-Manteiga (2008). "Outlier Detection in Functional Data by Depth Measures, with Application to Identify Abnormal NOx Levels". In: *Environmetrics* 19.4, pp. 331–345. ISSN: 1099-095X. DOI: 10.1002/env.878.

Flynn, Kathryn E. et al. (Apr. 2009). "Effects of Exercise Training on Health Status in Patients With Chronic Heart Failure: HF-ACTION Randomized Controlled Trial". In: *JAMA* 301.14, pp. 1451–1459. ISSN: 0098-7484. DOI: 10.1001/jama.2009.457.

Peng, C.-K., Madalena Costa, and Ary L. Goldberger (Jan. 2009). "Adaptive Data Analysis of Complex Fluctuations in Physiologic Time Series". In: *Advances in Adaptive Data Analysis* 01.01, pp. 61–70. ISSN: 1793-5369, 1793-7175. DOI: 10.1142/S1793536909000035.

Ramsay, J. O., Giles Hooker, and Spencer Graves (2009). *Functional Data Analysis with R and MATLAB*. Use R! Dordrecht ; New York: Springer. ISBN: 978-0-387-98184-0 978-0-387-98185-7.

Røislien, Jo et al. (Nov. 2009). "Simultaneous Estimation of Effects of Gender, Age and Walking Speed on Kinematic Gait Data". In: *Gait & Posture* 30.4, pp. 441–445. ISSN: 09666362. DOI: 10.1016/j.gaitpost.2009.07.002.

Ramsay, James O. and Bernard W. Silverman (2010). *Functional Data Analysis*. 2. ed. Springer Series in Statistics. New York, NY: Springer. ISBN: 978-1-4419-2300-4 978-1-4419-7761-8.

Febrero-Bande, Manuel and Manuel Oviedo de la Fuente (2012). "Statistical Computing in Functional Data Analysis: The R Package fda.usc". In: *Journal of Statistical Software* 51.4, pp. 1–28.

Older, Paul (Dec. 2013). "Anaerobic Threshold, Is It a Magic Number to Determine Fitness for Surgery?" In: *Perioperative Medicine* 2.1, p. 2. ISSN: 2047-0525. DOI: 10.1186/2047-0525-2-2.

Ullah, Shahid and Caroline F. Finch (May 2013). "Applications of Functional Data Analysis: A Systematic Review". In: *BMC Medical Research Methodology* 13.1, pp. 1–12. ISSN: 14712288. DOI: 10.1186/1471-2288-13-43.

Myers, Jonathan et al. (Feb. 2014). "A Neural Network Approach to Predicting Outcomes in Heart Failure Using Cardiopulmonary Exercise Testing". In: *International Journal of Cardiology* 171.2, pp. 265–269. ISSN: 01675273. DOI: 10.1016/j.ijcard.2013.12.031.

Marron, J. S. et al. (Nov. 2015). "Functional Data Analysis of Amplitude and Phase Variation". In: *Statistical Science* 30.4. ISSN: 0883-4237. DOI: 10.1214/15-STS524. arXiv: 1512.03216.

Zakynthinaki, Maria S. (Apr. 2015). "Modelling Heart Rate Kinetics". In: *PLoS ONE* 10.4. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0118263.

Mazzoleni, Michael et al. (Jan. 2016). "Modeling and Predicting Heart Rate Dynamics across a Broad Range of Transient Exercise Intensities during Cycling". In: *Sports Engineering* 19, pp. 117–127. DOI: 10.1007/s12283-015-0193-3.

Mirizzi, Gianluca et al. (Apr. 2016). "Prediction of the Chemoreflex Gain by Common Clinical Variables in Heart Failure". In: *PLOS ONE* 11.4, e0153510. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0153510.

Cardiopulmonary Exercise Testing (CPET) (July 2017). http://www.geh.nhs.uk/directory-of-services/specialties-and-services/c/cardio-respiratory-unit-cru/cardiopulmonary-exercise-testing-cpet/.

Guazzi, Marco, Francesco Bandera, et al. (Sept. 2017). "Cardiopulmonary Exercise Testing: What Is Its Value?" In: *Journal of the American College of Cardiology* 70.13, pp. 1618–1636. ISSN: 0735-1097. DOI: 10.1016/j.jacc.2017.08.012.

Hox, Joop, Mirjam Moerbeek, and Rens van de Schoot (Sept. 2017). *Multilevel Analysis: Techniques and Applications*. 3rd. New York, NY, USA: Routledge.

Matabuena, Marcos, Mario Francisco-Fernández, and Ricardo Cao (2017). "Predicting the Physiological Limits of Sport Stress Tests with Functional Data". In: *Functional Statistics and Related Fields*. Ed. by Germán Aneiros et al. Cham: Springer International Publishing, pp. 179–187. ISBN: 978-3-319-55845-5 978-3-319-55846-2. DOI: 10.1007/978-3-319-55846-2_24.

Baril, Jonathan-F. Benjamin Jason Jeremy (2018). "The Use of Activity Monitoring and Machine Learning for the Functional Classification of Heart Failure". M.H.Sc. Canada: University of Toronto (Canada). ISBN: 9780438673595.

Hearn, Jason et al. (Aug. 2018). "Neural Networks for Prognostication of Patients With Heart Failure: Improving Performance Through the Incorporation of Breath-by-Breath Data From Cardiopulmonary Exercise Testing". In: *Circulation: Heart Failure* 11.8. ISSN: 1941-3289, 1941-3297. DOI: 10.1161/CIRCHEARTFAILURE.118.005193.

Armstrong, Neil (2019). "Youth Aerobic Fitness". In: *Pediatric Exercise Science* 31.2, pp. 137–143. ISSN: 0899-8493, 1543-2920. DOI: 10.1123/pes.2019-0039.

Cooper, Dan M., Ronen Bar-Yoseph, et al. (Jan. 2019). "Exercise and Lung Function in Child Health and Disease". In: *Kendig's Disorders of the Respiratory Tract in Children (Ninth Edition)*. Ed. by Robert William Wilmott et al. Philadelphia: Elsevier, 212–230.e7. ISBN: 978-0-323-44887-1. DOI: 10.1016/B978-0-323-44887-1.00012-2.

Matabuena, Marcos, Juan Vidal, et al. (Aug. 2019). "Application of Functional Data Analysis for the Prediction of Maximum Heart Rate". In: *IEEE Access* PP, pp. 1–1. DOI: 10.1109/ACCESS.2019.2938466.

Beyaztas, Ufuk and Han Lin Shang (Mar. 2020). "On Function-on-Function Regression: Partial Least Squares Approach". In: *Environmental and Ecological Statistics* 27.1, pp. 95–114. ISSN: 1352-8505, 1573-3009. DOI: 10.1007/s10651-019-00436-1. arXiv: 1912.06995.

Febrero-Bande, Manuel Febrero et al. (Feb. 2020). *Fda.Usc: Functional Data Analysis and Utilities for Statistical Computing*.

Schaefer, Julia et al. (June 2020). "The Use of Machine Learning in Rare Diseases: A Scoping Review". In: *Orphanet Journal of Rare Diseases* 15.1, p. 145. ISSN: 1750-1172. DOI: 10.1186/s13023-020-01424-6.

Sietsema, Kathy E et al. (2020). *Wasserman & Whipp's: Principles of Exercise Testing and Interpretation: Including Pathophysiology and Clinical Applications*. ISBN: 978-1-975136-45-1 978-1-975136-43-7.

Bartels, Matthew N. and David Z. Prince (2021). "Acute Medical Conditions: Cardiopulmonary Disease, Medical Frailty, and Renal Failure". In: *Braddom's Physical Medicine and Rehabilitation*. Elsevier, 511–534.e5. ISBN: 978-0-323-62539-5. DOI: 10.1016/B978-0-323-62539-5.00027-8.

Bazoukis, George et al. (Jan. 2021). "Machine Learning versus Conventional Clinical Methods in Guiding Management of Heart Failure Patients—a Systematic Review." In: *Heart Failure Reviews* 26.1, pp. 23–34. ISSN: 13824147.

Kokoszka, Piotr and Matthew Reimherr (2021). *Introduction to Functional Data Analysis*. First issued in paperback. Texts in Statistical Science Series. Boca Raton London New York: CRC Press. ISBN: 978-1-03-209659-9 978-1-4987-4634-2.

Liguori, Gary et al., eds. (2021). *ACSM's Guidelines for Exercise Testing and Prescription*. Eleventh edition. Philadelphia: Wolters Kluwer. ISBN: 978-1-975150-22-8 978-1-975150-21-1.

Neder, J. Alberto et al. (2021). "Clinical Interpretation of Cardiopulmonary Exercise Testing: Current Pitfalls and Limitations". In: *Frontiers in Physiology* 12. ISSN: 1664-042X.

Ramsay, J. O., Spencer Graves, and Giles Hooker (Nov. 2021). *Fda: Functional Data Analysis*.

Bar-Yoseph et al. (2022). "Heart Rate and Gas Exchange Dynamic Responses to Multiple Brief Exercise Bouts (MBEB) in Early and Late Pubertal Boys and Girls". In: *unpublished*.

Appendices

Appendix A

Raw Data Plots

The figures below are the second-by-second observations of our four variables of interest: HR, RR, $\dot{V}O_2$, and $\dot{V}CO_2$.

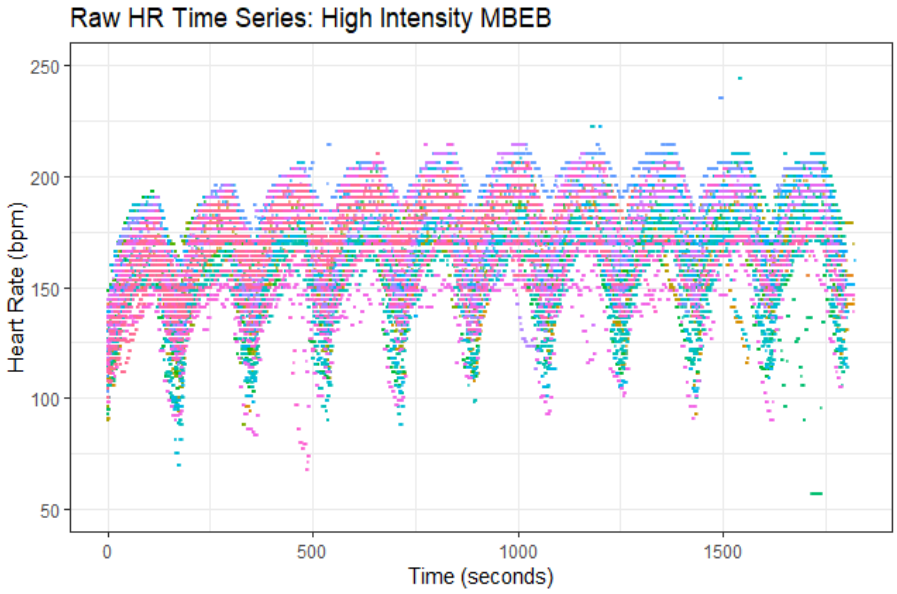


Figure A.1: Ten bouts of Heart Rate, originally observed breath-by-breath and time interpolated to a second-by-second representation. Each participant's observations are shown with a unique color.

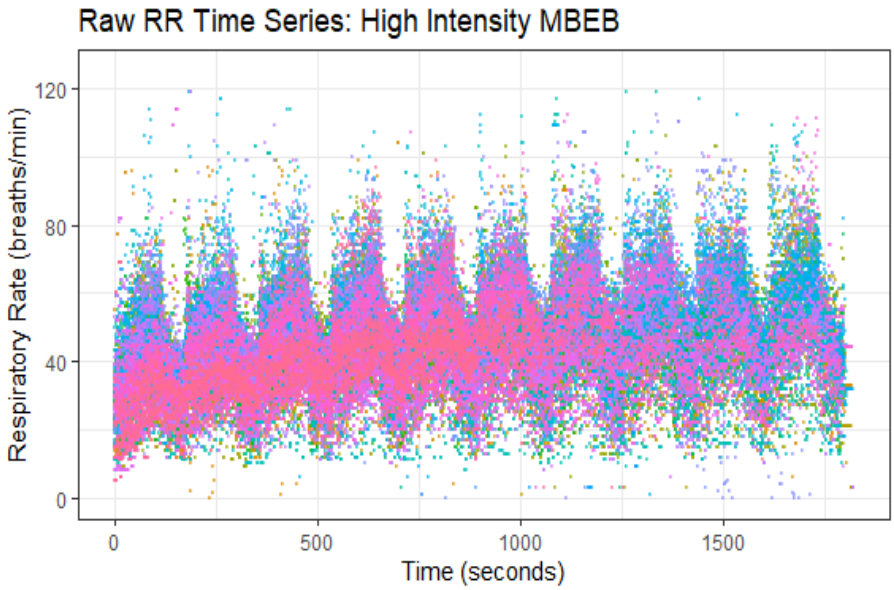


Figure A.2: Ten bouts of Respiratory Rate, originally observed breath-by-breath and time interpolated to a second-by-second representation. Each participant's observations are shown with a unique color.

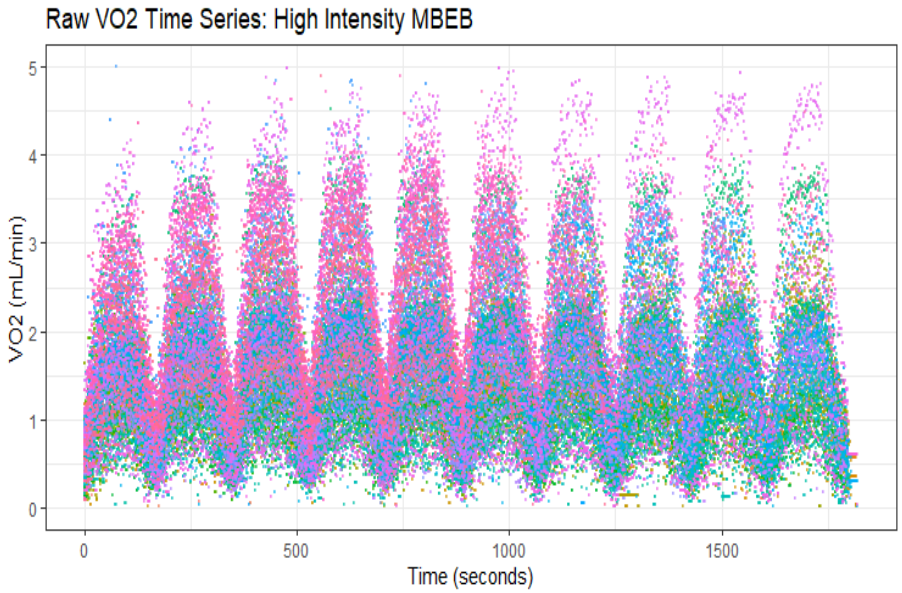


Figure A.3: Ten bouts of O_2 Uptake Rate, originally observed breath-by-breath and time interpolated to a second-by-second representation. Each participant's observations are shown with a unique color.

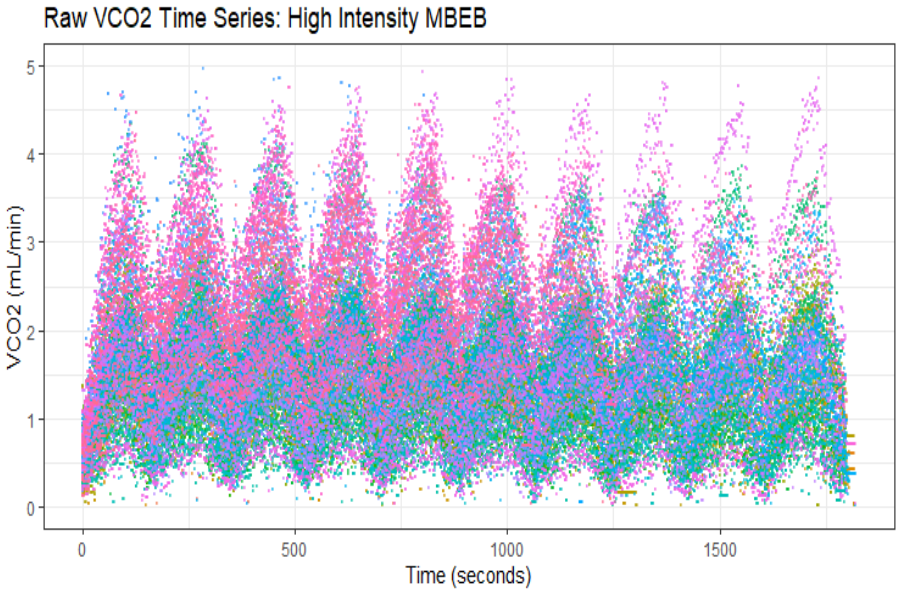


Figure A.4: Ten bouts of CO_2 Uptake Rate, originally observed breath-by-breath and time interpolated to a second-by-second representation. Each participant's observations are shown with a unique color.

Appendix B

Smoothed & Registered Functional Data Plots

The figures below are the functional data representations of our four variables of interest: HR, RR, $\dot{V}O_2$, and $\dot{V}CO_2$.

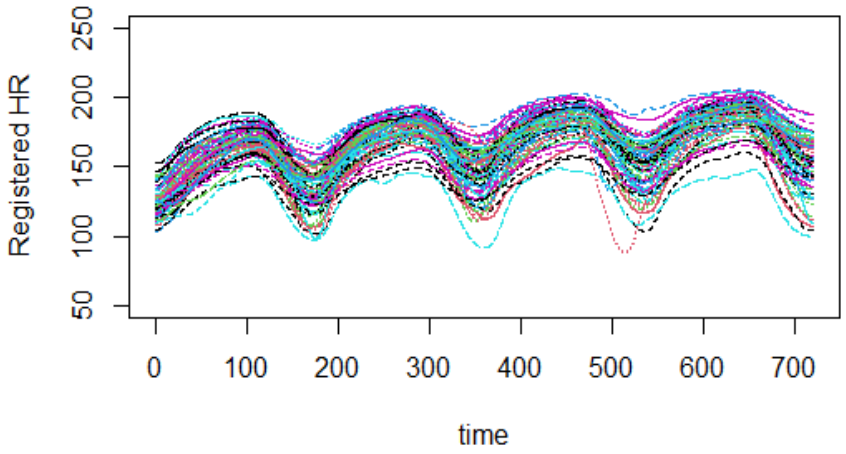


Figure B.1: Four bouts of Heart Rate after converting the discrete time series to 78 smoothed and continuously registered functional data objects.

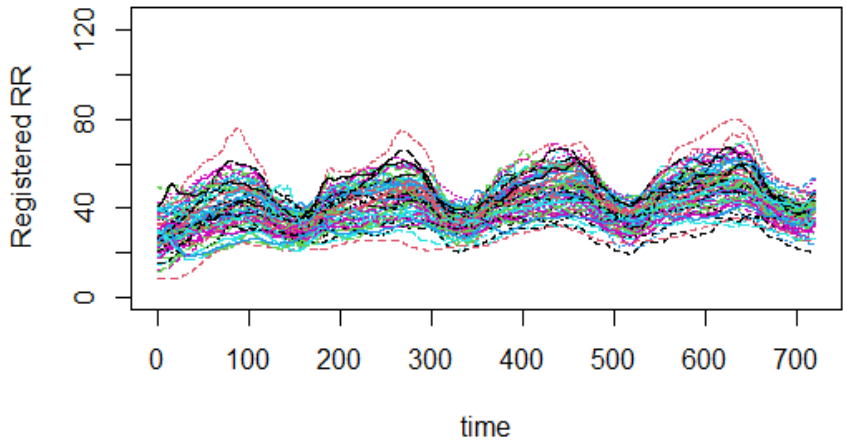


Figure B.2: Four bouts of Respiratory Rate after converting the discrete time series to 78 smoothed and continuously registered functional data objects.

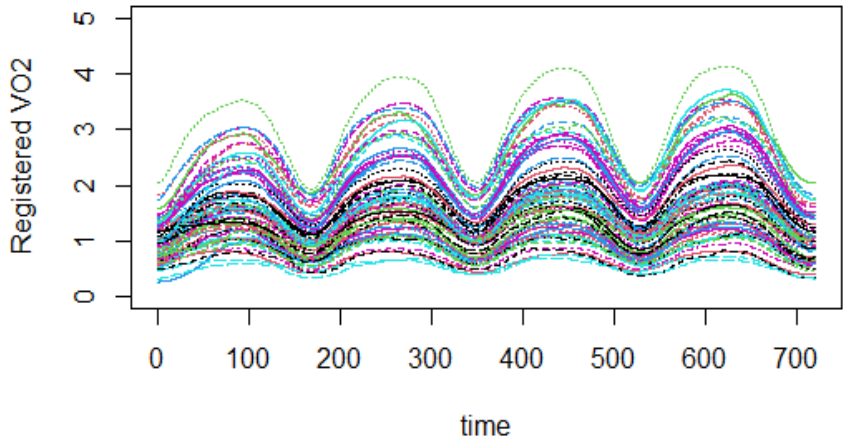


Figure B.3: Four bouts of O_2 uptake rate after converting the discrete time series to 78 smoothed and continuously registered functional data objects.

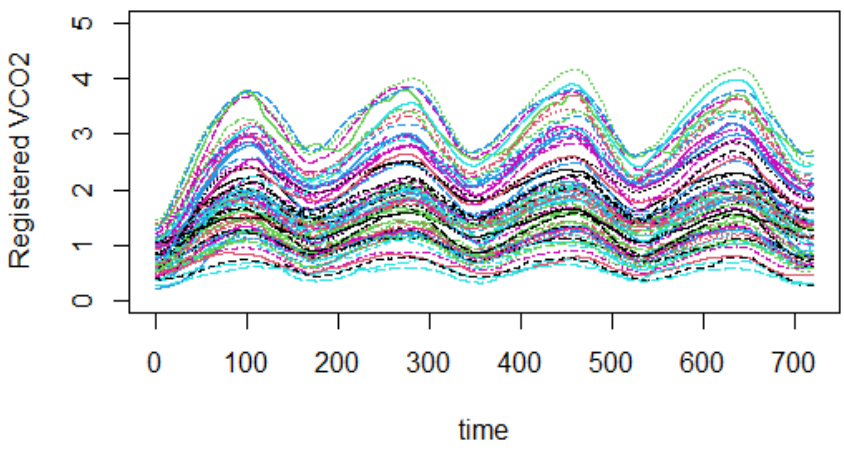


Figure B.4: Four bouts of CO_2 uptake rate after converting the discrete time series to 78 smoothed and continuously registered functional data objects.

Appendix C

Classification Model Descriptions

This appendix details the structure of each classification model. Models were built with consistent parameters to allow for performance comparison. Note that the individual models use only functional data from **one** physiological signal, and the multivariate models use functional data coefficients from **all four** signals. Also, the full multivariate models include $BodyMass$ as a third demographic scalar variable.

Classification was performed in the R package *'fda.usc'*. Wrapper versions of the following packages were called within the *'fda.usc'* functions:

- RPART: *rpart* package
- RandomForest: *randomForest* package
- SVM: *e1071* package
- LDA: *MASS* package
- Neural Network: *nnet* package

The binary class 'quit' (1 or 0) was predicted with the following covariates ($X_{[0,720]}$ represents the response variable and the function $s(\cdot)$ denotes an additive effect over the variable):

- GSAM: $s(X_{[0,720]})$
  - equal weights (1) were used for all observations in GSAM models

  - The probability value for binary discriminant (i.e. classification threshold) was optimized within each GSAM model; we searched across a range between 0.3 and 0.8, and the threshold which produced the highest F1 score was selected.

- GSAM + Covariates: $s(X_{[0,720]}) + Gender + PubertyLevel$ ($+ BodyMass$ for the full model)

- RPART: $s(X_{[0,720]}) + Gender + PubertyLevel$ ($+ BodyMass$ for the full model)

  - the value of prior probabilities was set to the default for *rpart*

- K-Nearest Neighbors: $X_{[0,720]} + Gender + PubertyLevel$ ($+ BodyMass$ for the full model)

  - the k number of nearest neighbors was chosen based on trial and error, to determine which k resulted in the lowest classification error. Therefore, k varies between 12 and 14 among the models.

- RandomForest: $X_{[0,720]} + Gender + PubertyLevel$ ($+ BodyMass$ for the full model)

  - we used the default value for the number of trees to grow (500) and the number of variables available for splitting at each tree node (square root of total number of variables)

- Support Vector Machines: $X_{[0,720]} + Gender + PubertyLevel$ ($+ BodyMass$ for the full model)

  - default values were used for the C parameter (1) and γ parameter (1/data dimension) in the radial basis function kernel

- Linear Discriminant Analysis: $X_{[0,720]} + Gender + PubertyLevel$ ($+ BodyMass$ for the full multivariate model)

  - the important parameter was the prior probabilities of class membership; with our balanced data, we used the class proportions for the training set

- Neural Network: $X_{[0,720]} + Gender + PubertyLevel$ ($+ BodyMass$ for the full model)

  - we used the default value for weights (1) in the neural net

Appendix D

FANOVA Results

Functional ANOVA results for the HR, RR, $\dot{V}O_2$, and $\dot{V}CO_2$ guided the investigation of the null hypothesis. For p-values ≤ 0.05, we reject the null hypothesis and conclude that there is a statistically significant difference in the functional means for quitters and non-quitters.

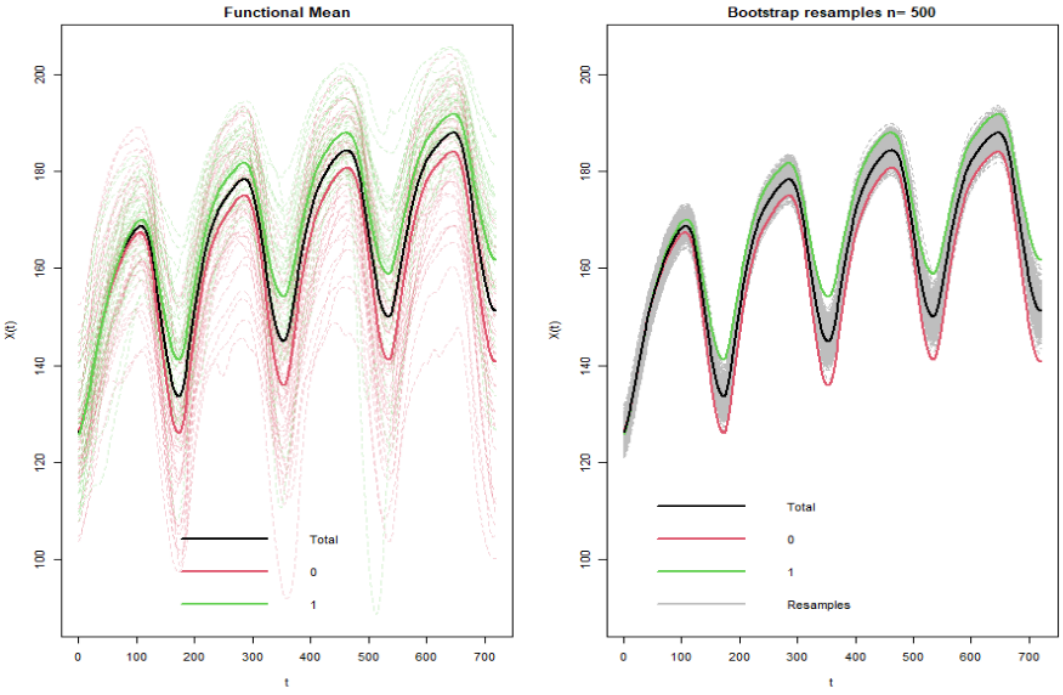


Figure D.1: Comparison of functional means for the Heart Rate signal during the first four exercise bouts. The reported p-value = **0.000**. Participants who quit exercise during MBEB are labelled as '1' and colored green. The red line depicts the functional mean for non-quitters. The black line indicates the mean trajectory for all 78 participants.

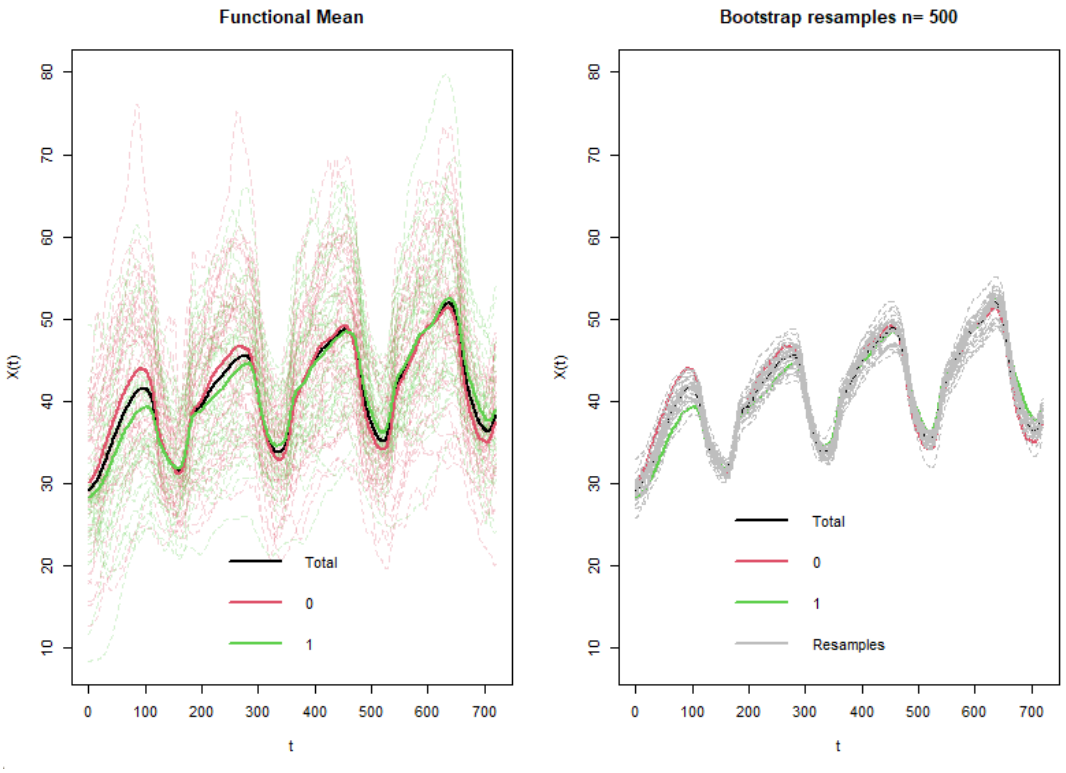


Figure D.2: Comparison of functional means for the RR signal during the first four exercise bouts. The reported p-value = **0.186**. Participants who quit exercise during MBEB are labelled as '1' and colored green. The red line depicts the functional mean for non-quitters. The black line indicates the mean trajectory for all 78 participants. Notice that there is substantial overlap between the two groups' signals; quitters and non-quitters have virtually indistinguishable respiratory rate trajectories.

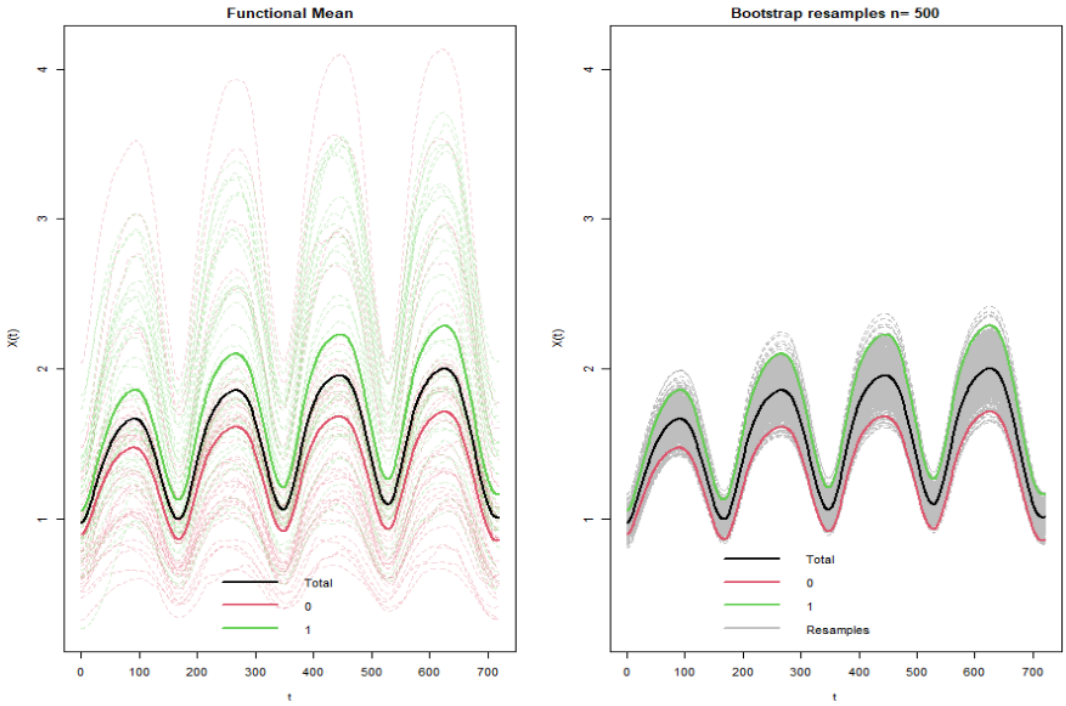


Figure D.3: Comparison of functional means for the $\dot{V}O_2$ signal during the first four exercise bouts. The reported p-value = **0.000**. Participants who quit exercise during MBEB are labelled as '1' and colored green. The red line depicts the functional mean for non-quitters. The black line indicates the mean trajectory for all 78 participants.

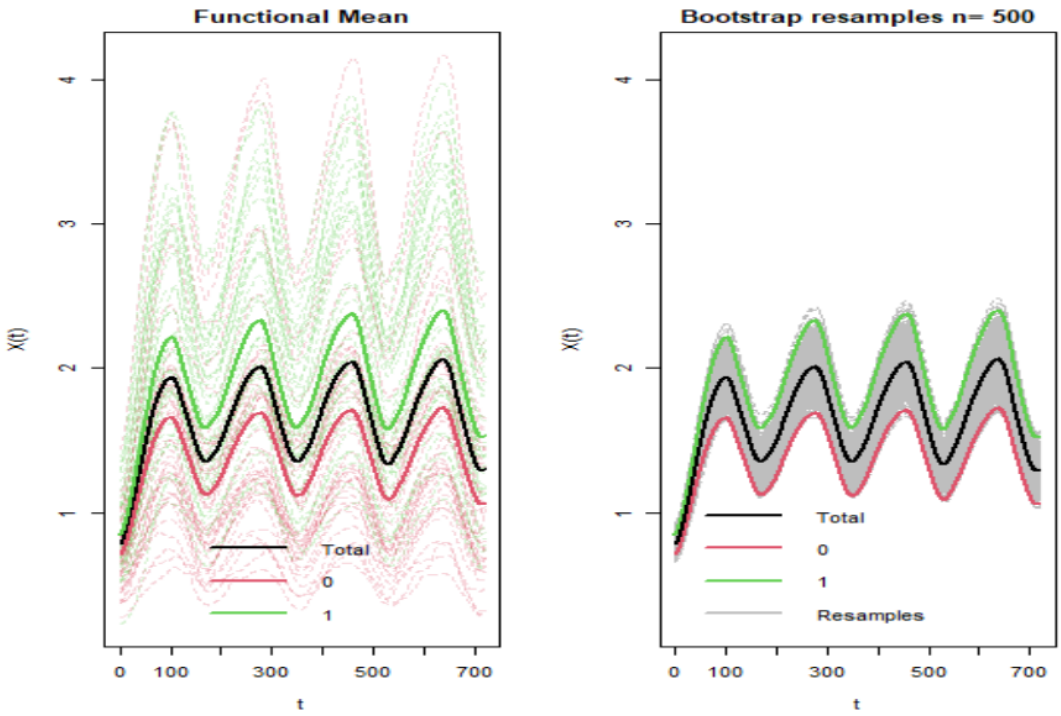


Figure D.4: Comparison of functional means for the $\dot{V}CO_2$ signal during the first four exercise bouts. The reported p-value = **0.000**. Participants who quit exercise during MBEB are labelled as '1' and colored green. The red line depicts the functional mean for non-quitters. The black line indicates the mean trajectory for all 78 participants.