Predicting daily recovery during long-term endurance training using machine learning analysis

Authors: Jeffrey A. Rothschild[1*], Tom Stewart[1], Andrew E. Kilding[1], and Daniel J. Plews[1]

Affiliations:
[1] Sports Performance Research Institute New Zealand (SPRINZ), Auckland University of Technology, Auckland, New Zealand

*Corresponding author - Jeffrey Rothschild – Jeffrey.Rothschild@aut.ac.nz

Jeffrey Rothschild: ORCID: 0000-0003-0014-5878
Andrew Kilding: ORCID: 0000-0002-5334-8831

All authors have read and approved this version of the manuscript for pre-print.

Please cite as: Rothschild JA, Stewart T, Kilding AE, and Plews DJ. 2022. Predicting daily recovery during long-term endurance training using machine learning analysis. SportRχiv.

Abstract

Purpose: The aim of this study was to determine if machine learning models could predict the perceived morning recovery status (AM PRS), training feeling during exercise (exercise TF), and daily change in heart rate variability (HRV change) of endurance athletes based on training, dietary intake, sleep, HRV, and subjective wellbeing measures.

Methods: Self-selected nutrition intake, exercise training, sleep habits, HRV, and subjective wellbeing of 40 endurance athletes was monitored daily for 12 weeks (3,325 days of tracking). Global and individualized models were constructed using nine machine learning techniques and combined into an ensemble model at the group level, and with a single best algorithm chosen for individualized models. Model performance was compared with a baseline intercept-only model.

Results: Prediction error (root mean square error [RMSE]) was lower than baseline for the group models (12.1 vs. 17.5, 13.1 vs. 14.7, and 0.25 vs. 0.30 for AM PRS, exercise TF, and HRV change, respectively). At the individual level prediction accuracy outperformed the baseline model but varied greatly across participants (RMSE range 5.5 to 23.6, 5.7 to 18.2, and 0.05 to 0.52 for AM PRS, exercise TF, and HRV change, respectively).

Conclusion: Daily recovery measures can be predicted based on commonly measured variables, with a small subset of variables providing most of the predictive power. However, at the individual level the key variables may vary, and additional data may be needed to improve prediction accuracy.

Keywords: training load monitoring, cycling, running, triathlon, nutrition, sleep, HRV

1. Introduction

Coaches and athletes routinely monitor a range of metrics with the hope of gaining insight into how an athlete is responding to their training. These can include measures of training load (duration and intensity), heart rate variability (HRV), sleep, diet, and daily measures of subjective wellbeing, among others.[1] Despite careful planning, there can still be large discrepancies between the training stimulus prescribed by coaches and experienced by athletes.[2] Improved understanding of an athlete's training response could allow a training plan to be better tailored to an individual's needs, and help minimize the risks of non-functional overreaching, illness, and/or injury.[3]

Training load refers to the combination of training volume and intensity, and can be measured and classified as either external or internal.[4] External training loads are characterized by measures such as distance, power, or speed, whereas internal loads reflect the relative physiological strain represented by heart rate (HR), blood lactate, and session rating of perceived exertion (sRPE).[3] Internal load has been recommended as the primary measure when monitoring athletes, as it plays a pivotal role in determining training outcomes and can reflect variations in the stress response to a given external load due to other stressors such as extreme temperature, or accumulated training fatigue.[4] Coaches often use subjective wellness ratings by athletes for monitoring purposes, which are sensitive to fluctuations in training load.[5] However, much of the research on the relationship between training load, sRPE, and wellness has been in team sports and not endurance sports, and has not accounted for potential interactions between training load, sleep, and dietary intake.

From a nutrition perspective, athletes and sports nutritionists are continually challenged to balance the nutritional demands of training while also optimizing body composition and promoting skeletal muscle adaptation. Increasing energy and carbohydrate intake during periods of intensified endurance training can attenuate symptoms of overreaching,[6] yet many athletes routinely train in an overnight-fasted state and/or restrict carbohydrate intake before exercise.[7] The interaction between dietary intake and training quality in the context of longer-term, self-selected training and nutrition intake has not been well characterized. Although logistically challenging, investigating longer-term dietary intake during endurance training would help elucidate the role of self-selected nutrition intake on daily recovery during endurance training. The increased availability of valid and user-friendly mobile food-tracking apps can help facilitate data collection while minimizing disruption to an athlete's training and lifestyle.

The relationship between training, diet, sleep, and other lifestyle factors is complex, as many factors converge which may have non-linear and/or temporal relationships, with one often influencing the other. This underscores the need for more advanced tools for understanding athlete readiness and wellbeing. Machine learning techniques have been increasingly used in sports science, particularly in the context of multi-factorial data such as predicting injuries,[8] training feeling scores,[9] and subjective wellbeing,[10] as well as in nutrition research to model complex nutrient interactions and address confounding variables.[11] However, to our knowledge machine learning has yet to be used to predict an endurance athlete's perceived recovery or HRV based on a combination of factors routinely monitored by athletes and coaches. Therefore, the

goal of this study was to predict perceived AM recovery status, wellbeing during exercise, and daily change in HRV based on training metrics, dietary intake, sleep, HRV, and subjective wellbeing. Secondary aims were to highlight the most important variables for accurate prediction, and to examine the influence of factors that can tangibly be manipulated by coaches and athletes. It is hoped that such information can allow coaches to focus on a subset of variables with the strongest predictive power.

2. Methods

2.1 Study design

This observational study monitored the daily self-selected nutrition intake, exercise training, sleep habits, HRV, and subjective wellbeing of endurance athletes for 12 weeks. Throughout the study period, participants were free to perform any type of exercise and consume any type of diet. Measures of diet, training, sleep, HRV, and subjective wellbeing were recorded daily. Models were created for three primary outcome variables — two subjective measures (AM Perceived Recovery Status (PRS) score, and Training Feeling (TF) during exercise score), and an objective measure of change in resting HRV from the previous day (HRV change). The study was open to male and females aged 18 or older who train at least seven hours per week, were using a smartphone app to track their dietary intake at least five days per week, capture HRV daily, and track sleep duration using a wearable device. All study protocols and materials were approved by the Auckland University of Technology Ethics Committee (22/7), and all participants provided

informed consent prior to starting the study. Data collected from the same athletes related to training load and carbohydrate periodization have been reported elsewhere.[12]

2.2 Participants

Fifty-five endurance athletes (61.8% male, aged 42.6 ± 9.1 years, training 11.6 ± 3.9 hours per week) took part in the study. The primary sports represented were triathlon (67.3%), running (20.0%), cycling (10.9%), and rowing (1.8%). The self-reported competitive level included professional (2.6%), elite non-professional (qualify and compete at the international level as an age-group athlete, 34.6%), high-level amateur (qualify and compete at National Championship-level events as an age-group athlete, 25.6%), and amateur (enter races but don't expect to win, or train but do not compete, 37.2%) athletes.

2.3 Assessment of self-reported exercise

All exercise was recorded in Training Peaks software (TrainingPeaks, Louisville, CO, USA). Each session was noted for modality (e.g., bike, run, swim), total time, and session rating of perceived exertion (sRPE[13]) using the Borg CR100® scale, which offers additional precision compared with the CR10 scale.[14] Participants were instructed to rate their perceived effort for the whole training session within 1-h of exercise, although sRPE scores are temporally robust from minutes to days following a bout of exercise.[13] As an indicator of the type of feedback that occurs between athletes and coaches on a daily basis, participants also rated a subjective (TF) score from 0–100 using a customized scale based on the Perceived Recovery Status (PRS) scale[15] (supplemental Fig. 1). Athletes were instructed to consider how they felt during the training session, which was

distinct from the sRPE. For example, someone could feel very good during a hard workout and very poor during an easy workout, or vice-versa. Additionally, participants noted the amount of carbohydrate (in grams) consumed within the 4-h pre-exercise window.

2.4 Assessment of self-reported dietary intake

Details of dietary assessment have been described elsewhere.[12] Briefly, participants were instructed to maintain their typical dietary habits and record all calorie-containing food and drink consumed for the duration of the 12-week study, using the MyFitnessPal application (www.myfitnesspal.com). Due to previous habitual use, three participants used the Cronometer application (www.cronometer.com) and one participant used the Carbon application (www.joincarbon.com). Incomplete days of tracking (2.2 ± 4.6% of days per participant) were removed from the data, and analysis of the calorie intake trend over time was performed for each participant as an additional check of compliance as previously described.[12] Four participants were excluded from the analysis due to the detection of a downward trend in daily calorie intake that could not be explained by changes in training load or body weight.

2.5 Assessment of resting HRV and sleep

Resting HRV was recorded daily, and analyzed using the natural logarithm of the square root of the mean sum of the squared differences (Ln rMSSD) between R–R intervals.[16] For participants using Oura ring (Oura Health, Oulu, Finland) or Whoop straps (Whoop, Inc., Boston, USA) nocturnal HRV was used, whereas measurements were taken upon waking for those using the HRV4Training (www.hrv4training.com), Elite HRV (Elite HRV, Inc., Asheville, USA), or ithlete (HRV

Fit Ltd. Southampton, UK) smartphone apps. High correlations have been reported between nocturnal and morning HRV measurements.[17] Nightly sleep duration was recorded using wearable devices, which included Oura ring, Whoop strap, Applewatch, Fitbit, and Garmin models. These consumer-grade devices offer adequate accuracy in detecting sleep-wake times, but not sleep staging.[18-21] Further details of participant devices used for HRV and sleep tracking are shown in supplemental figure 2.

2.6 Assessment of subjective wellbeing

Each morning participants answered four questions related to subjective wellbeing, which have been shown to respond consistently to training-induced stress.[22] The PRS scale[15] was used to measure overall recovery with athletes manually typing a number into Training Peaks software. The 100-point version of the scale was used, which has been shown discriminate between answers better than the 10-point scale.[14] In addition, ratings of life stress (1–7), sleep quality (1–7), and muscle soreness (1–10) were also recorded into the software each morning (Supplemental Fig. 3). Participants were familiarized with all scales prior to starting the study. In addition, participants were asked to record their body mass at least one time per week.

2.7 Data preparation

Training load was calculated for each workout as the product of sRPE and duration of exercise in minutes,[23] divided by 10 to account for the 100-point scale. Exercise was summed into daily totals for workout duration and training load, along with coded variables for modality of workout (e.g., swim, bike, run, strength, other) and if any training was performed in the fasted state. Because

dietary protein and fat ingestion have minimal effects on substrate oxidation,[24] fasted training was defined as consuming < 5 g of carbohydrate in the 4-h pre-exercise window. For multiple exercise sessions in a single day, a weighted mean based on the duration of each session was used to calculate a single daily value for pre-exercise carbohydrate ingestion and TF score. External load metrics such as HR, power, or pace were not collected because many athletes undertake activities that can't be quantified on a common scale such as strength training, yoga, or swimming without a HR monitor. This was deemed acceptable because sRPE is considered to be a valid and reliable method for calculating training load across modalities.[23] Seven-day rolling measures for training monotony (a measure of day-to-day variability in the weekly training load, calculated as average daily load divided by the standard deviation) and training strain (product of total weekly training load and training monotony) were calculated.[23] Exponentially weighted 7-d moving averages of training load, HRV, and resting HR were calculated to account for residual effects of recent training.[25] A sleep index value was calculated as the product of sleep duration and subjective sleep quality.[26] Daily training volume (hours per day) and training load for each participant is presented in supplemental Figures 4 and 5.

Participants were excluded from the analysis if they trained an average of less than 6 h per week (n = 8) or did not log at least 85% of the required data points (n = 3). Participants who did not complete the full 12 weeks due to illness, injury, or drop-out but completed at least 6 weeks of tracking were included in the analysis (n = 11). Among participants included in the analysis (n = 40), 2.5 ± 1.7 % of data points were missing. Missing values were imputed at the individual level

using multiple linear regression and nearest neighbor algorithms for diet and training measures, and using median values for other variables.[27]

To increase the available options for modeling and interpretation, the data were transformed from a time series into independent observations. A time series is a sequence of data points at equally spaced points in time and ordered chronologically. Time series data cannot be analyzed with common techniques such as linear modeling if the day-to-day observations are correlated with observations at previous time points (i.e., auto-correlated) and are not independent of each other, as key assumptions of linear regression are violated.[28] To account for this, a process of Markov unfolding[29] was used. This is based on the Markov assumption, whereby the values in any state are influenced only by the values of the immediately preceding or a small number of immediately preceding states.[30] Data were analyzed for autocorrelation, and it was determined that a maximum of seven previous days could have a relevant influence on a given day's data. This makes logical sense, as many behavioral and training schedules follow a weekly cycle. The process of Markov unfolding entails copying the columns of the original dataset, shifting them down by one row, and stacking them as new columns on the right of the dataset (labeled as lag 1). This is repeated with shifts of 2–n, where n is the number of previous days to be included. The first n rows from the beginning of the dataset are discarded, as there are missing values for some of the lags. This results in a dataset that is a few rows shorter, but n + 1 times wider than the original dataset and the observations can be treated as totally independent, allowing the use of any modeling approach that assumes independent data. This approach to making the dataset ~7x wider can result in the curse of dimensionality, whereby the test error tends to increase as

the dimensionality of the problem (i.e. the number of predictors) increases,[28] but this may be mitigated by the use of algorithms which use regularization to conduct feature selection.[27] It should be noted that the variables created as 7-d rolling averages would allow the previous 14 days of information to be provided to the model (i.e., a 7-d average from 7 days ago). All analyses were carried out with R version 4.0.3 (The R foundation for Statistical Computing, Vienna, Austria). Descriptive statistics are provided as mean ± SD.

2.8 Models

A series of models were built for the three outcomes of interest — AM PRS score, exercise TF score, and daily change in HRV, at both the group level (full dataset) and for each individual participant. To reduce multicollinearity, highly correlated predictors (Pearson correlation > 0.85) were removed from the dataset prior to training each model by removing the one with the largest mean absolute correlation with the rest of the data.[27] For each outcome, models were made using the primary subset of variables (MAIN), and a subset of variables that can tangibly be manipulated by athletes/coaches (ACTIONABLE). Included variables are shown in Table 1. Descriptive statistics of these variables are provided in supplemental table 1. At the group level, ensemble models (described below) for each outcome were made using the two variable subsets (MAIN and ACTIONABLE). For comparison with the ensemble models, three linear regression models were created — a least absolute shrinkage and selection operator (LASSO) regression model using the MAIN set of variables, a linear mixed model using the 5 variables with the highest importance scores from the MAIN ensemble as fixed effects and participant ID specified as a random effect (Top 5 from group MAIN), and an intercept-only model as a baseline comparison

that was simply predicting the mean value. The LASSO was chosen as a linear model that can be used with a large number of variables due to its built-in feature selection process, the mixed model was chosen to determine how a very limited subset of variables would perform, and the intercept-only baseline was used to establish a realistic upper bound for the root mean squared error (RMSE), as useful prediction models should have lower RMSE values. At the individual level, in addition to the MAIN and ACTIONABLE models, linear models were made consisting of the 5 variables with the highest importance from the MAIN group model (Top 5 from group MAIN), the 5 variables with the highest importance from their own respective MAIN model (Top 5 from individual MAIN), and an intercept-only model as a baseline comparison.

For group models, data were split into a training set (75%) and a testing set (25%). To avoid data leakage,[31] all observations from a given participant were assigned to either the training or testing set, and preprocessing steps such as standardization and removal of highly correlated variables were performed only using the training set. Nine different learning algorithms, including parametric and non-parametric methods, were trained for each model using the *Tidymodels* ecosystem in R. These included three linear regression models with regularization (Ridge, LASSO, and LASSO with interaction terms), three non-linear regression models (Multivariate Adaptive Regression Spline (MARS), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)), two ensembles of decision trees (XGBoost and Light GBM), and a single layer neural network (NNET). Ten-fold cross-validation was repeated five times for tuning parameter optimization, and the tuned models were combined into a stacked ensemble using the *Stacks* R package. Stacking is a method that takes the outputs of many models and combines them to generate a new ensemble

model.[32] Predictions from each candidate in the ensemble are weighted based on a stacking coefficient, generated by the betas of a LASSO regression model fitting the true outcome with the predictions given in the data stack. Model performance (RMSE and R-squared) was calculated using the hold-out (testing) dataset. Variable importance (a measure of the strength of the relationship between observed values of the variable and the observed response) for the group ensemble models was determined using a permutation-based approach, which measures a feature's importance by calculating the increase of the model's prediction error after permuting the feature.[33] In addition to variable importance, partial dependence profiles were created to aid model interpretation, which show how the expected value of a model prediction changes after accounting for the average effect of all other variables.[33] A model info sheet for detecting and preventing data leakage is provided as a supplemental file based on the recommendations from Kapoor and Narayanan.[31]

The individual models used the same algorithms mentioned above, except for the NNET. Ten-fold cross-validation was repeated ten times, with the best algorithm and parameter set chosen based on the lowest RMSE. Accuracy metrics were calculated using 500 bootstrap resamples. Variable importance was calculated using a model-based approach,[34] and scaled so the total importance summed to 1. Because individual models could use different algorithms for each participant, scaling the importance allowed a summarization of importance across different model types by taking the mean values. To compare performance among the five types of individual models (best model MAIN, best model ACTIONABLE, Top 5 from group MAIN, Top 5 from individual MAIN, and baseline intercept-only model), a linear mixed model was used, with

RMSE from each model used as the dependent variable, model type specified as a fixed effect, and participant ID specified as a random effect. Estimated means were calculated using the *Emmeans* R package, and comparisons made using the Tukey test.

Table 1 Overview of variables included in the modeling

Category	Variables
Training	Exercise duration (min), modality, fasted training (yes/no), number of workouts per day, number of consecutive training days, session rating of perceived exertion (sRPE, highest for a single session each day and a duration-based weighted average for the day), training load (TL; min x sRPE), 7-d exponentially weighted and non-weighted moving average of TL, 7-d highest single-day TL, training monotony (weekly mean TL/weekly SD), training strain (weekly load x monotony), training feeling (TF), day of the week
Dietary	Total kcal, carbohydrate (CHO, g/kg), fat (g/kg), protein (g/kg), pre-exercise CHO (g), 3-d and 7-d moving averages of CHO, fat, protein, and kcal intake, 7-d moving average standard deviation of daily CHO intake and CHO monotony (weekly mean intake/ weekly SD)
Sleep	Sleep duration (hours), *sleep index (sleep duration x quality),* 7-d moving average sleep duration and *sleep index*
Subjective measures	*Perceived Recovery Status (PRS)*, soreness, life stress, *sleep quality*
Non-exercise	*Resting HRV* and *resting HR* (daily, change from previous day, and 7-d moving averages of each)
Planned interactions	**AM PRS:** 7-d average TL * 3-d average CHO intake 7-d training monotony * 3-d average CHO intake **Exercise TF**: Pre-exercise CHO intake * TL Prior day CHO intake * prior day TL 7-d average CHO intake * 7-d average TL **HRV:** Prior day TL * sleep duration Prior day TL * prior day AM PRS score
Subject characteristics	Participant ID, age, HRV app, sleep app, percentage of missing data, competitive level, primary sport, training age, body weight
Top 5 Variables from group MAIN	**AM PRS:** AM PRS 1 and 2 days ago, soreness, life stress, sleep quality **Exercise TF**: AM PRS, AM PRS 2 days ago, pre-exercise CHO, training strain 7 days ago, exercise duration (min) **HRV:** 7-d avg HRV change 1 day ago, HRV change 1, 2, and 7 days ago, HRV 1 day ago

Italics indicate variables that were removed from the ACTIONABLE models. Descriptive statistics for these variables are provided in supplemental table 1.

3. Results

A total of 3,325 days of tracking were included in the analysis (83.1 ± 9.6 per participant). Average participant training volume was 11.9 ± 3.5 h per week. Mean daily dietary intake was 38.9 ± 8.6 kcal/kg, 4.0 ± 1.5 g/kg carbohydrate, 1.9 ± 0.4 g/kg protein, and 1.7 ± 0.5 g/kg fat. Average sleep duration was 7.5 ± 0.7 hours per night. Values for the three main outcomes were 61.7 ± 18.5, 62.2 ± 15.7, and 0.0 ± 0.3 for AM PRS, exercise TF, and HRV change, respectively. Density plots showing the distribution of the three main outcome variables for each participant are shown in supplemental Figure 6. MAIN group models demonstrated improved accuracy compared with the baseline model (Table 2). Accuracy of the individual models was improved compared with the baseline models (Table 3) but varied more than 5-fold across participants (Figure 1). Figures 2–4 show the ten variables with the highest importance from the group modeling for AM PRS (Figure 2), Exercise TF (Figure 3), and HRV change (Figure 4), as well as a scatterplot comparing predicted vs. actual values (inset into each figure), and partial dependence plots showing how the expected value of a model prediction changes based on these variables. Figure 5 shows the ten variables with the highest mean importance scores across all participants for the individual MAIN models.

Table 2 Accuracy of Group Models

Outcome	Model	Variables	RMSE [95% CI]	R^2
AM PRS	Ensemble	MAIN	12.1 [11.5, 12.7]	0.52
AM PRS	LASSO	MAIN	12.9 [12.3, 13.6]	0.45
AM PRS	LMM	Top 5 from MAIN Ensemble	13.6 [13.0, 14.3]	0.41
AM PRS	Ensemble	ACTIONABLE	16.4 [15.6, 17.3]	0.16
AM PRS	Baseline	Intercept only	17.5 [16.7, 18.4]	NA
Exercise TF	Ensemble	MAIN	13.1 [12.4, 13.8]	0.23
Exercise TF	LMM	Top 5 from MAIN Ensemble	13.1 [12.4, 13.8]	0.22
Exercise TF	LASSO	MAIN	13.2 [12.5, 13.9]	0.20
Exercise TF	Baseline	Intercept only	14.7 [14.0, 15.5]	NA
Exercise TF	Ensemble	ACTIONABLE	14.9 [14.1, 15.7]	0.02
HRV change	Ensemble	MAIN	0.25 [0.24, 0.26]	0.40
HRV change	LASSO	MAIN	0.25 [0.24, 0.26]	0.41
HRV change	LMM	Top 5 from MAIN Ensemble	0.26 [0.25, 0.27]	0.33
HRV change	Baseline	Intercept only	0.30 [0.29, 0.32]	NA
HRV change	Ensemble	ACTIONABLE	0.32 [0.31, 0.34]	0

AM PRS: AM Perceived Recovery Status, Exercise TF: Exercise Training Feeling score, LASSO: linear regression model with regularization, LMM: Linear Mixed Model with participant ID specified as a random effect, RMSE: Root Mean Squared Error, in units of the original measurement (0–100 for AM PRS and Exercise TF, and Ln rMSSD for HRV change).

Table 3 Accuracy of Individual Models

Outcome	Model	Variables	RMSE [a]	R^2
AM PRS	Linear	Top 5 from individual MAIN	12.1 ± 4.3[a]	0.31 ± 0.17
AM PRS	Linear	Top 5 from group MAIN Ensemble	12.4 ± 4.1[a]	0.28 ± 0.16
AM PRS	*	MAIN	12.8 ± 4.2[ab]	0.23 ± 0.13
AM PRS	*	ACTIONABLE	13.3 ± 4.4[b]	0.19 ± 0.11
AM PRS	Baseline	Intercept only	14.2 ± 4.3[c]	NA
Exercise TF	Linear	Top 5 from individual MAIN	11.9 ± 3.6[a]	0.23 ± 0.10
Exercise TF	*	ACTIONABLE	12.6 ± 3.3[b]	0.11 ± 0.07
Exercise TF	Baseline	Intercept only	12.7 ± 3.3[b]	NA
Exercise TF	*	MAIN	12.8 ± 3.8[b]	0.12 ± 0.07
Exercise TF	Linear	Top 5 from group MAIN Ensemble	12.8 ± 3.6[b]	0.14 ± 0.11
HRV change	Linear	Top 5 from individual MAIN	0.20 ± 0.09[a]	0.59 ± 0.15
HRV change	*	MAIN	0.23 ± 0.11[b]	0.41 ± 0.19
HRV change	Linear	Top 5 from group MAIN Ensemble	0.24 ± 0.09[b]	0.37 ± 0.09
HRV change	Baseline	Intercept only	0.30 ± 0.12[c]	NA
HRV change	*	ACTIONABLE	0.31 ± 0.12[c]	0.08 ± 0.05

AM PRS: AM Perceived Recovery Status, Exercise TF: Exercise Training Feeling score, RMSE: Root Mean Squared Error, in units of the original measurement (0–100 for AM PRS and Exercise TF, and Ln rMSSD for HRV change). *For MAIN and ACTIONABLE models, values are from the single best-performing algorithm (LASSO: 30%, SVM: 30%, XGBoost: 23%, Light GBM: 12%, KNN: 4%, Ridge: 1%, and MARS: 0.4% of models). All metrics were established using 500 Bootstrap resamples. [a] Within each outcome, models not sharing any letter are significantly different by the Tukey test at the 5% level of significance.

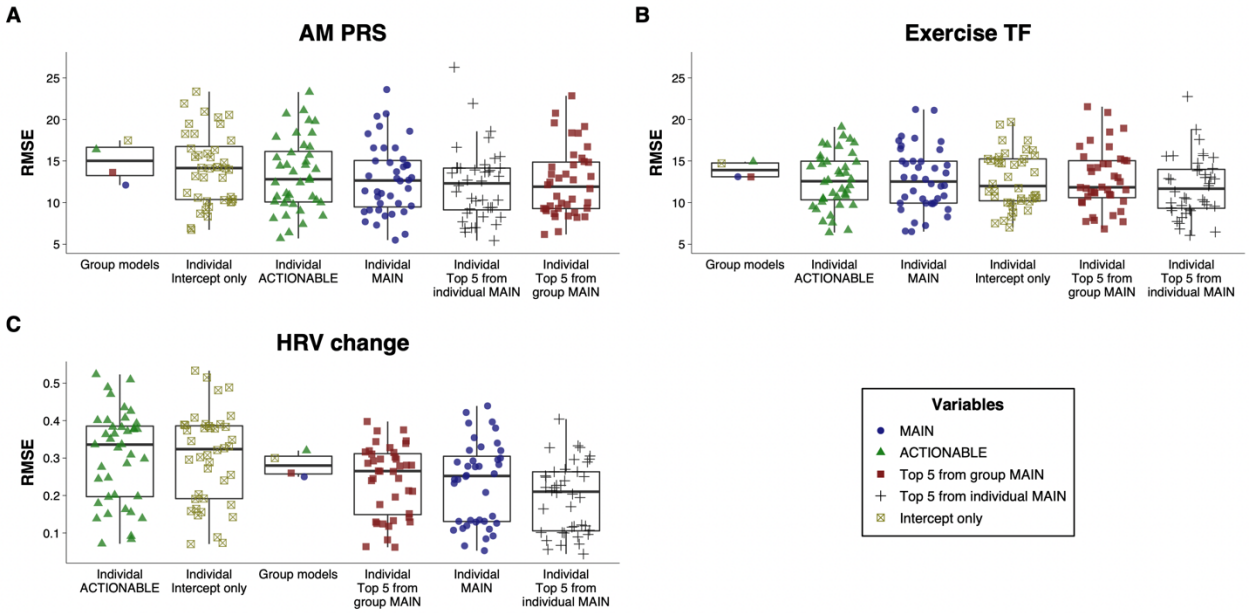


Figure 1. Root mean squared error (RMSE) of the group and individual models, separated by the variable set included in the model and ordered by mean RMSE values. For group models, MAIN and ACTIONABLE models represent the ensemble model, and "Top 5 from MAIN" represents a linear mixed model with the top 5 features from the MAIN group model based on variable importance scores. For individual models, "Top 5 from group MAIN" represents a linear model with the same top 5 features from the MAIN group model, and "Top 5 from individual MAIN" represents a linear model with the top 5 features from each participant's MAIN model. RMSE values were determined using out-of-sample data for group models and using 500 Bootstrap resamples for individual models.

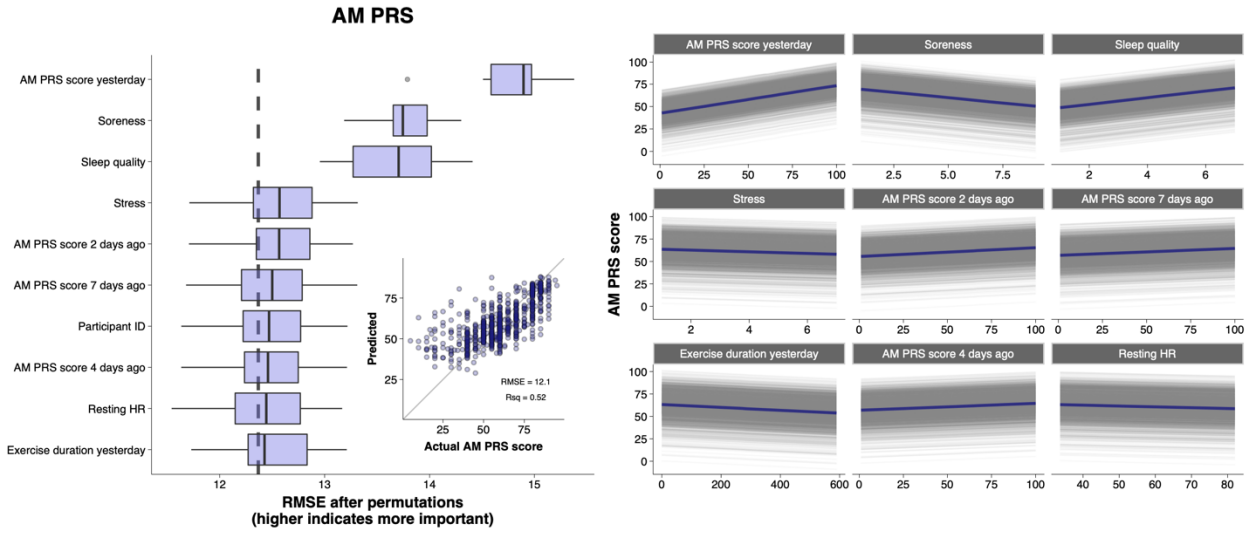


Figure 2. AM PRS group model results from the best performing model (ensemble using MAIN variables). Top 10 most important variables based on permutation-based feature importance are shown in a boxplot, along with a scatterplot of actual vs. predicted values on an out-of-sample dataset (inset), and partial dependence plots for the top 9 continuous variables (right), where colored lines represent the average of all observations shown individually as the grey lines. The vertical dashed line in the boxplot represents the full model RMSE from the training dataset.

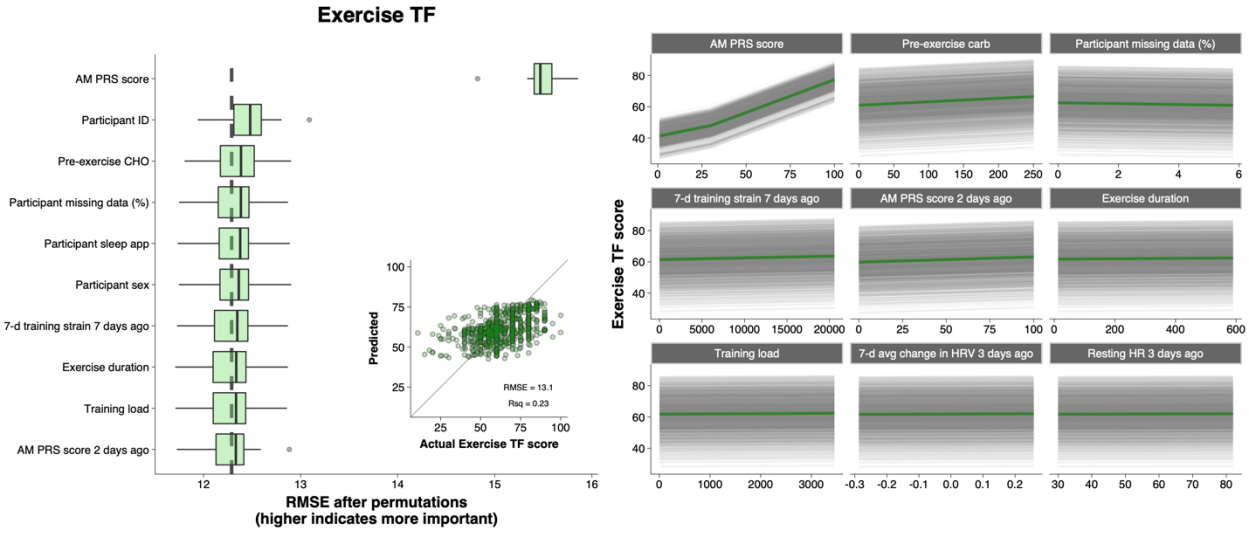


Figure 3. Exercise TF group model results from the best performing model (ensemble using MAIN variables). Top 10 most important variables based on permutation-based feature importance are shown in a boxplot, along with a scatterplot of actual vs. predicted values on an out-of-sample dataset (inset), and partial dependence plots for the top 9 continuous variables (right), where colored lines represent the average of all observations

shown individually as the grey lines. The vertical dashed line in the boxplot represents the full model RMSE from the training dataset.

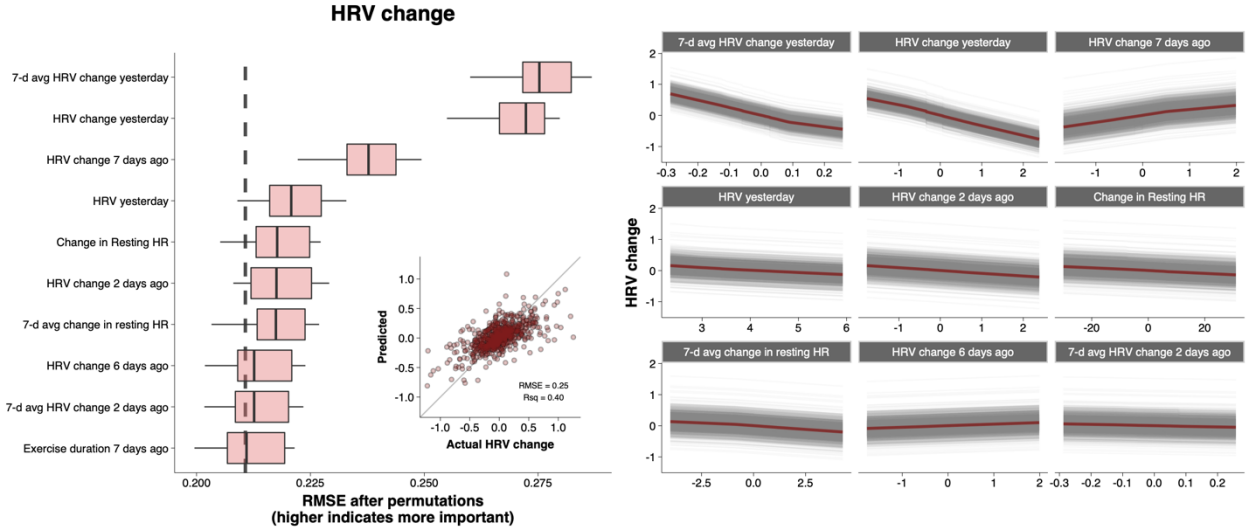


Figure 4. HRV change group model results from the best performing model (ensemble using MAIN variables). Top 10 most important variables based on permutation-based feature importance are shown in a boxplot, along with a scatterplot of actual vs. predicted values on an out-of-sample dataset (inset), and partial dependence plots for the top 9 continuous variables (right), where colored lines represent the average of all observations shown individually as the grey lines. The vertical dashed line in the boxplot represents the full model RMSE from the training dataset.

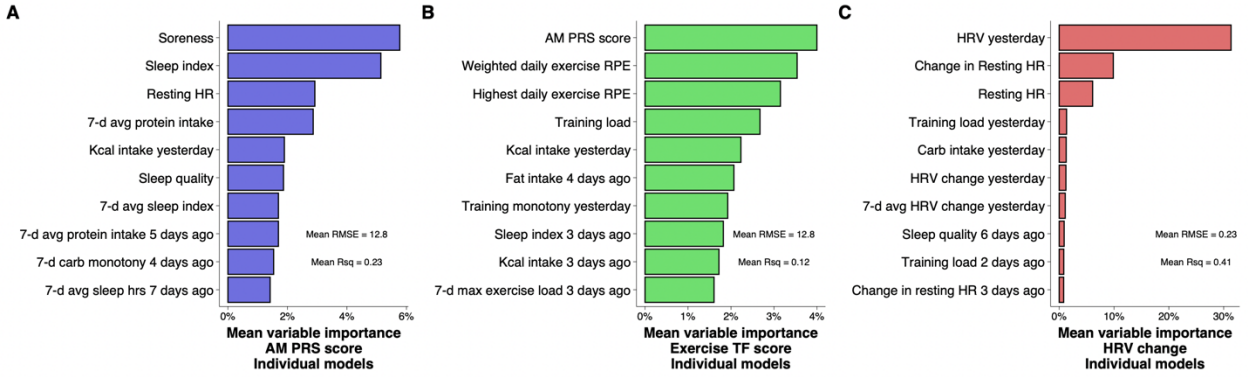


Figure 5. Variables with the highest mean variable importance scores from the individual MAIN models for AM PRS (A), Exercise TF (B), and HRV change (C) models. Root mean squared error (RMSE) and R-squared (Rsq) values shown are the average of the individual MAIN models.

4. Discussion

Athlete monitoring can help coaches better understand how an athlete is adapting to a training program, and minimize the risk of developing non-functional overreaching, illness, and/or injury.[3] This study utilized a novel approach to monitoring endurance athletes throughout 12 weeks of self-selected training to better understand the factors that can predict an athlete's day-to-day recovery and wellbeing. Key findings from this study are 1) day-to-day recovery measures can be predicted based on commonly measured variables, 2) a small subset of variables offers similar predictive capability as the full dataset, 3) predictive accuracy varies greatly at the individual level, and 4) remote monitoring of multiple training, diet, sleep, and recovery measures can be performed throughout longer-term training in real-world environments.

4.1 Model Performance

All models constructed using the MAIN variables outperformed the baseline model, demonstrating utility of the tracked variables. Unexpectedly, performance at the group level of the ACTIONABLE models was poor, indicating they alone do not offer any added value for predicting the recovery markers used in our study. As shown in the scatterplots in Figures 2–4, prediction accuracy was generally worse for scores on the upper and lower ends, likely due to the small number of extremely low or high values with which to train the models. At the individual level, the most striking finding was the large degree of variation in model performance (Figure 1). This suggests key variables may be missing from the models that could disproportionately affect some athletes more than others. For example, alcohol intake, acute illness, and the

menstrual cycle are known to influence HRV.[35] Moreover, participants spent only ~10% of their waking hours engaged in exercise, indicating the potential for many non-exercise factors to influence recovery and wellness such as walking, job-related physical activity/stress, massage, sauna, and/or ice baths. Future studies could expand on the current work by accounting for some or all of these additional factors.

At the group level, the ensemble models displayed the best overall performance, but these algorithms can be computationally expensive and slow to run. Because of this, two linear regression models were constructed as practical alternatives. The LASSO regression model is well suited to handle a large number of predictors because it uses regularization to reduce estimated coefficients towards zero,[28] essentially removing non-needed variables from the model. We also used a linear mixed model of the top 5 variables based on variable importance scores, with participant ID included as a random effect. Although the ensemble outperformed the other models for AM PRS score, the LASSO and linear mixed model performed well on out-of-sample data and the three models had roughly the same accuracy for exercise TF and HRV change (Table 2).

When constructing individual-level MAIN and ACTIONABLE models, the single best model from the suite of machine learning algorithms was chosen as the accepted model. Two linear regression models were then made, using the top 5 variables from the group and individual MAIN models. The best performance was achieved from the linear models with the top 5 individual variables (Table 3), highlighting the importance of a very small subset of variables that coaches

and practitioners could pay closer attention to. Importantly, the difference in performance between linear models of the top 5 from individual and top 5 from group models highlights the fact that the most important variables for each athlete will be different (Table 3). From a practical perspective, a prudent approach might be to start by monitoring a wide array of variables, and reduce the number based on feedback from initial models.

4.2 Variable Importance

The variable importance calculations of the group-level models revealed a small number of variables having a disproportionately large influence on prediction accuracy (Figures 2–4). This finding is corroborated by the generally good performance of the linear mixed models, which included only the top 5 variables, and implies the ability for coaches and practitioners to focus on just a few of the many variables that are routinely monitored. These include muscle soreness, life stress, and sleep quality for AM PRS scores, pre-exercise CHO, training strain, exercise duration, and AM PRS for exercise TF scores, and the changes in HRV over recent days for predicting the current day's change in HRV. However, when trying to predict at the individual level, the chosen variables should be specific to the individual. The aggregated importance scores from the individual models shown in Figure 5 are far more diverse than at the group level, supporting the notion that the most important variables vary among different athletes. For example, among two participants with the lowest RMSE values for the AM PRS models, the most important variables were muscle soreness, prior-day PRS scores, and prior-day protein intake for one athlete, while the top variables for another athlete were all related to sleep (prior night, previous nights, and 7-d rolling averages). The importance of the individual differences is further

evidenced by the improved performance (2–17% improvement in RMSE) of the individual linear models that used the individual's top 5 variables compared with the top 5 group variables (Table 3).

4.3 Explain vs. Predict

The priority of a statistical model can be to explain (i.e., test causal explanations), predict (new or future observations), or describe the data structure in a compact manner.[36] The focus of this analysis was on predictive power, for several reasons. The observational nature of our data from free-living environments is better suited to predictive modeling, whereas laboratory-controlled experimental data are better for explanatory modeling.[36] In the context of a large dataset with complex relationships, predictive modeling can help uncover potential new causal mechanisms and lead to the generation of new hypotheses.[36] This is reflected in the variable importance scores, particularly for HRV change, where few of the top predictors could be thought of as having any causal role. However, new hypotheses could be generated relating to a potential reversion to the mean effect for HRV, for example, based on the negative relationship between the top predictors and the daily change in HRV (Figure 4, partial dependence plot). From a practical perspective, use of these models should be limited to communicating the expected values for an athlete on a given day, rather than suggesting ways to modulate the variables of interest.

4.4 Athlete monitoring

Direct monitoring of training and fatigue responses is common in high-performance sport environments.[1] Better understanding of an athlete's response to training and recovery could help

coaches improve the effectiveness of a training program. However, it is challenging to control for, or even account for, the large number of variables potentially influencing an athlete's response to training, particularly over longer time frames. Observational studies can help to answer questions that would not be feasible to study in a controlled laboratory environment. A strength of this study design is the length of monitoring period, which allowed athletes to capture a range of daily and weekly training volumes. Advances in technology have also opened far more opportunities to gather valid and reliable data from athletes in their home training environments.[37,38] Although dietary intake can often be underreported, nearly all previous studies have used short-duration food records rather than smartphone apps. It has been suggested that familiarity with and interest in keeping food records may lead to more reliable estimates of energy intake,[39] and in our study all participants were already habitually recording dietary intake using a smartphone app. Although this approach to gathering data would not suit all athletes, many are accustomed to daily tracking of a wide range of data, and it is likely that a model-based analytical approach could offer valuable insight.

4.5 Machine Learning

Machine learning has been increasingly used in sports science, often for predicting injuries,[8] but also for predicting training feeling scores,[9] and subjective wellbeing.[10] Machine learning algorithms can be criticized for their lack of transparency, particularly when combined in an ensemble as we did in this study. This approach was chosen to optimize prediction accuracy, with linear models constructed as a transparent alternative. Indeed, the ensemble models achieved the best performance, but the linear models also performed nearly as well (Tables 2 and 3). This

finding is echoed by a systematic review showing no performance benefit of machine learning over logistic regression for clinical prediction models.[40] However, in our study the complex models played a critical role in being used to identify the top variables for the linear models.

4.6 Limitations

Limitations of this study relate to the observational and uncontrolled nature of the data collection, the large number of variables collected, and the potential for important factors to have not been collected. Participants were required to record their training, diet, sleep, HRV, and subjective wellbeing daily for 12 weeks. We specifically recruited people who were already doing this routinely, as this approach would not be practical for all athletes. Data integrity was checked based on the number of missing values, and by looking for trends in dietary reporting that could not be explained by changes in training load or body weight. Nonetheless, it is possible that participants did not always enter data as accurately as possible. There is also the risk of bias in reporting if an athlete is aware that their coach or a researcher will be seeing their data, answering based on what they think is desirable. Despite capturing a wide range of variables, we only had a single measure of internal training load and no measure of external load. This was done to accommodate athletes training across a variety of endurance and strength training modalities. Future research in single-sport athletes (e.g., cyclists or runners) would allow additional load metrics like HR, total work, or distance to be more easily factored into the modeling. In addition, energy availability, alcohol intake, and menstrual cycle tracking would be desirable metrics to include. Future work could also benefit from using continuous sliding scales for subjective wellbeing measures that would allow decimal places to be recorded, rather than

the 7- or 10-point integer scales built-in to the training monitoring software. This was the reason we used the 100-point, rather than 10-point PRS scale.[14] Finally, no performance measures were captured, leaving the ultimate utility of this approach unclear.

5. Perspective

To our knowledge, this is the first study of its kind to track this diverse range of self-selected and self-reported training of endurance athletes. Findings from this study, and the approach used, can enable coaches and athletes to better understand and focus on the few key measures which can offer an outsized amount of predictive capability. Although the prediction accuracy could likely be improved by capturing additional variables of interest, the current predictions offer information that is practically relevant. For example, an RMSE value of 12 from our model using the 100-point scale would translate to an average error of 1.2 when using a 10-point wellbeing scale, providing a coach with a useful gauge of an athlete's readiness. These data also reveal the importance of looking into factors affecting each athlete, rather than applying group-level findings to the individual. Importantly, use of these models should be limited to communicating the expected values for an athlete on a given day, rather than suggesting ways to modulate the variables of interest. This approach can also be combined with domain knowledge to individualize key metrics for athlete monitoring and evaluation.

Figure captions

Figure 1. Root mean squared error (RMSE) of the group and individual models, separated by the variable set included in the model and ordered by mean RMSE values. For group models, MAIN and ACTIONABLE models represent the ensemble model, and "Top 5 from MAIN" represents a linear mixed model with the top 5 features from the MAIN group model based on variable importance scores. For individual models, "Top 5 from group MAIN" represents a linear model with the same top 5 features from the MAIN group model, and "Top 5 from individual MAIN" represents a linear model with the top 5 features from each participant's MAIN model. RMSE values were determined using out-of-sample data for group models and using 500 Bootstrap resamples for individual models.

Figure 2. AM PRS group model results from the best performing model (ensemble using MAIN variables). Top 10 most important variables based on permutation-based feature importance are shown in a boxplot, along with a scatterplot of actual vs. predicted values on an out-of-sample dataset (inset), and partial dependence plots for the top 9 continuous variables (right), where colored lines represent the average of all observations shown individually as the grey lines. The vertical dashed line in the boxplot represents the full model RMSE from the training dataset.

Figure 3. Exercise TF group model results from the best performing model (ensemble using MAIN variables). Top 10 most important variables based on permutation-based feature importance are shown in a boxplot, along with a scatterplot of actual vs. predicted values on an out-of-sample dataset (inset), and partial dependence plots for the top 9 continuous variables (right), where colored lines represent the average of all observations shown individually as the grey lines. The vertical dashed line in the boxplot represents the full model RMSE from the training dataset.

Figure 4. HRV change group model results from the best performing model (ensemble using MAIN variables). Top 10 most important variables based on permutation-based feature importance are shown in a boxplot, along with a scatterplot of actual vs. predicted values on an out-of-sample dataset (inset), and partial dependence plots for the top 9 continuous variables (right), where colored lines represent the average of all observations shown individually as the grey lines. The vertical dashed line in the boxplot represents the full model RMSE from the training dataset.

Figure 5. Variables with the highest mean variable importance scores from the individual MAIN models for AM PRS (A), Exercise TF (B), and HRV change (C) models. Root mean squared error (RMSE) and R-squared (Rsq) values are the average of the individual MAIN models.

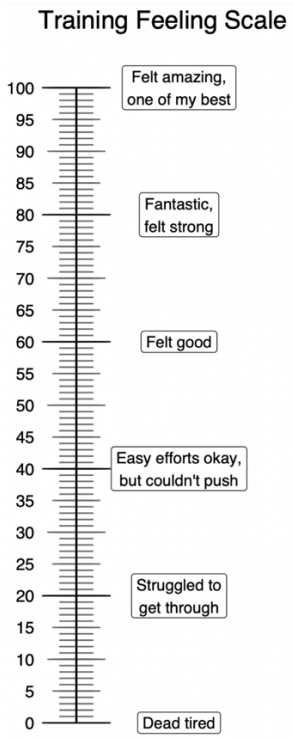
Availability of data and materials

The authors are willing to discuss data sharing under collaborative agreements. Please contact the corresponding author.

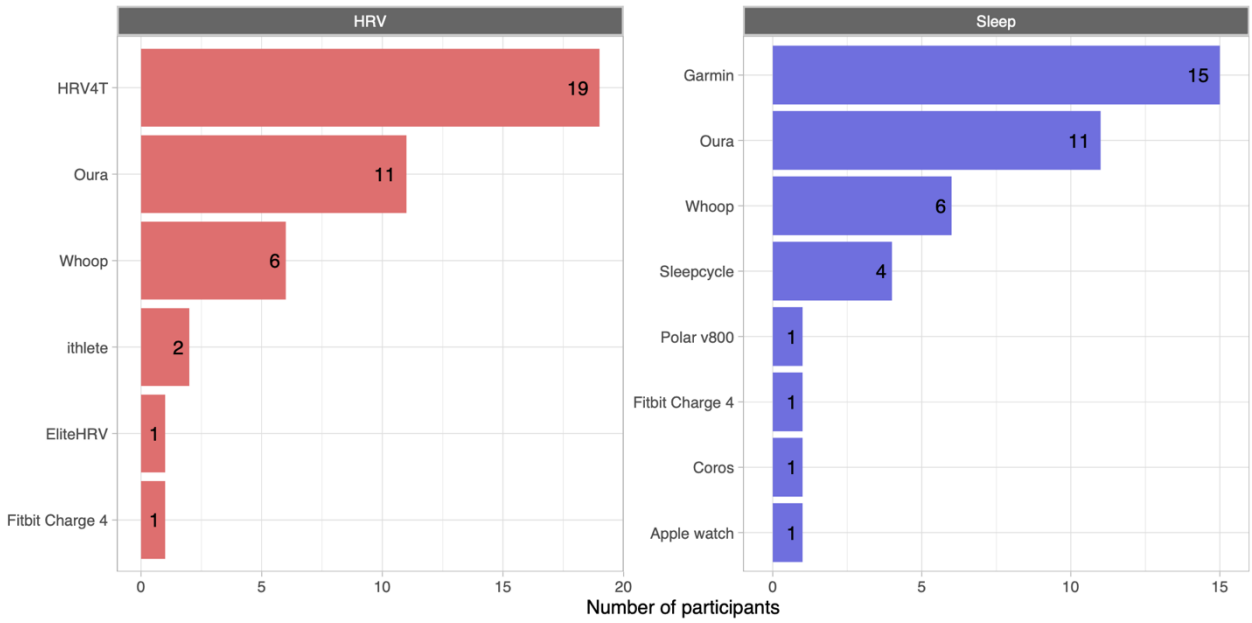
Code availability

The R code used in this analysis is publicly available, using a synthetic dataset mimicking the original data at https://github.com/Jeffrothschild/ML_predictions_code

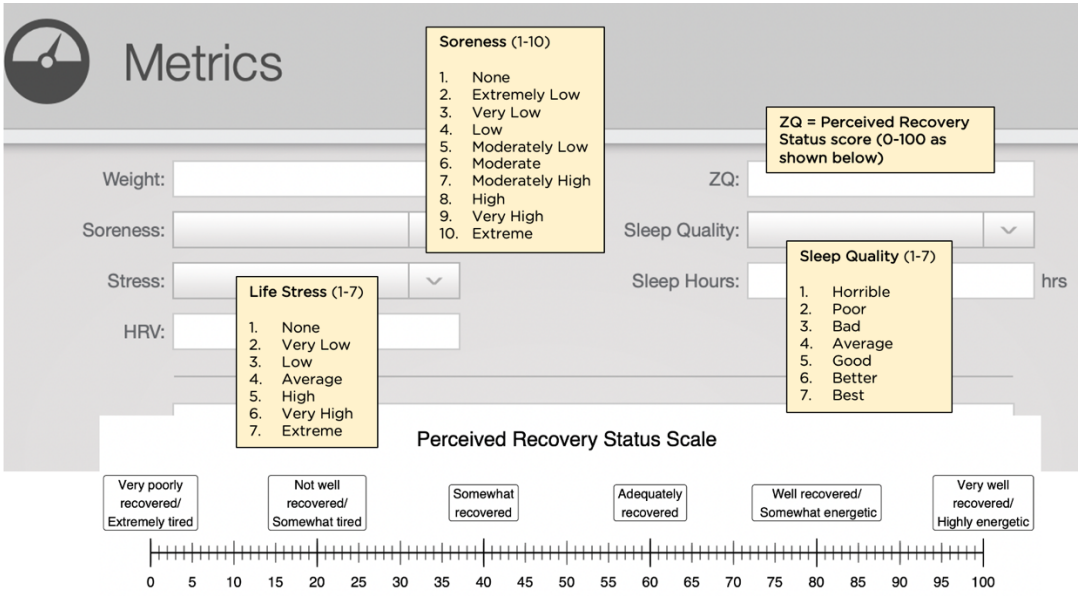
Supplemental Figures



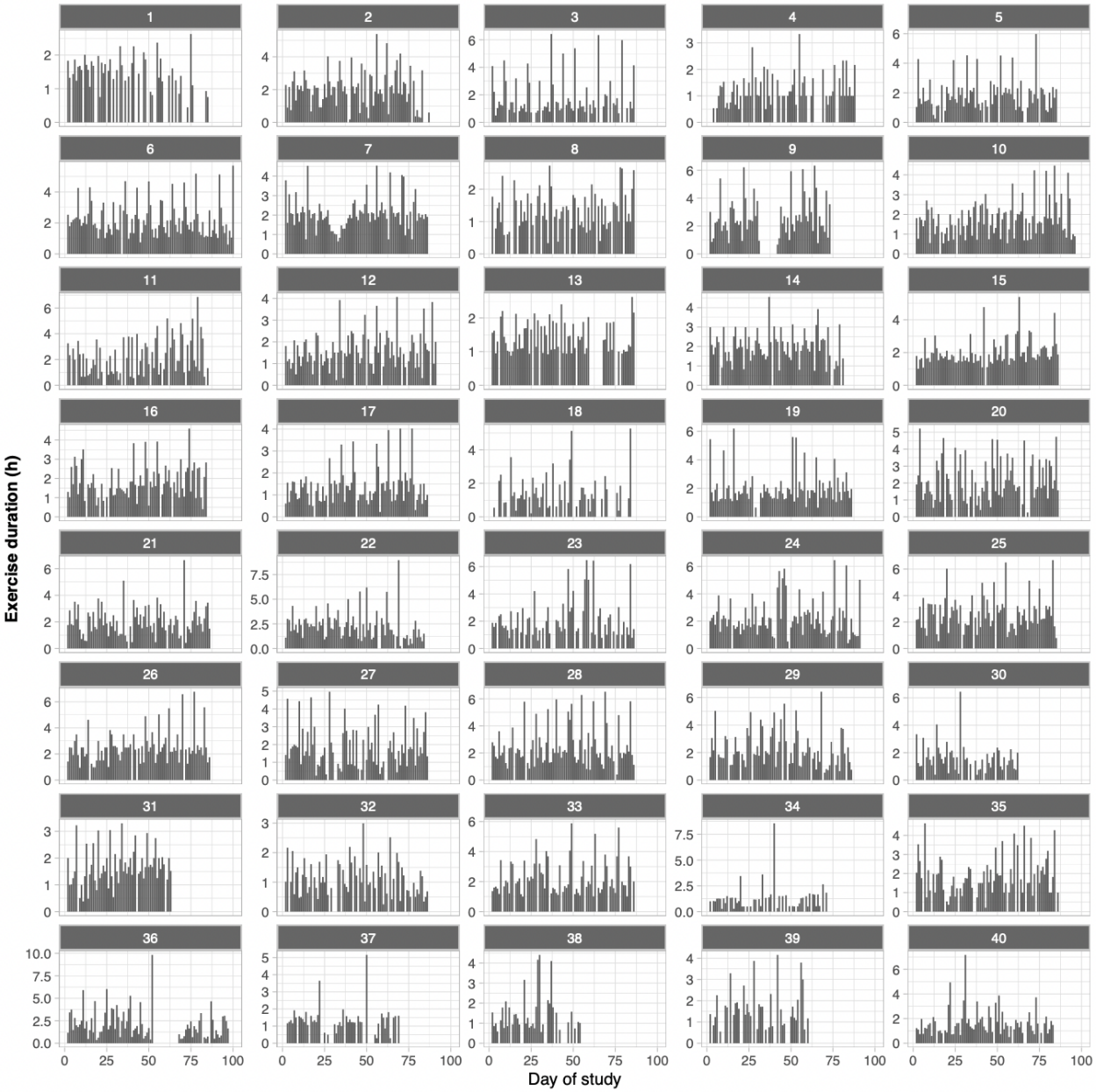
Supplemental figure 1 – Training Feeling (TF) scale



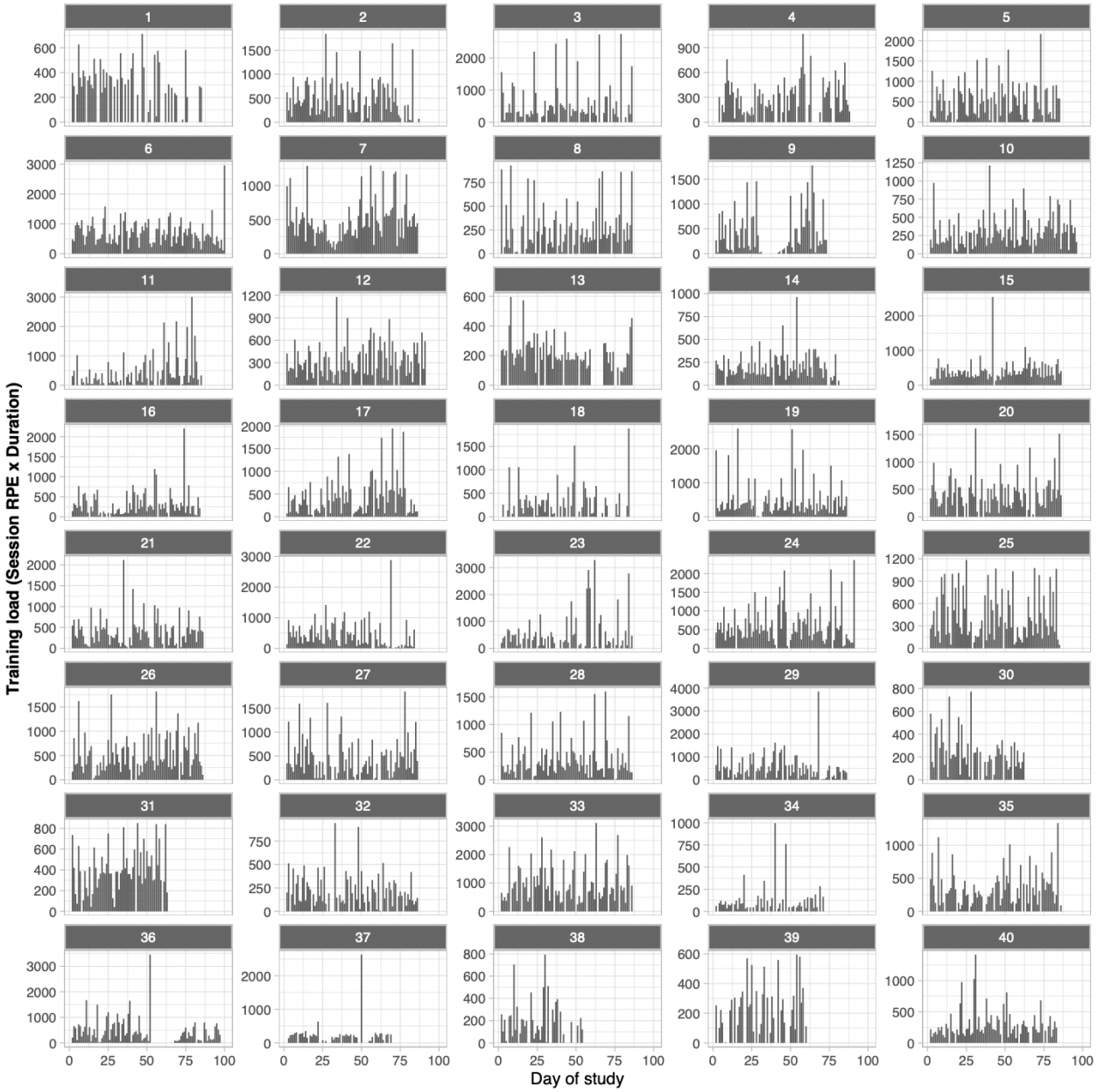
Supplemental figure 2. Participant devices used for sleep and HRV tracking.



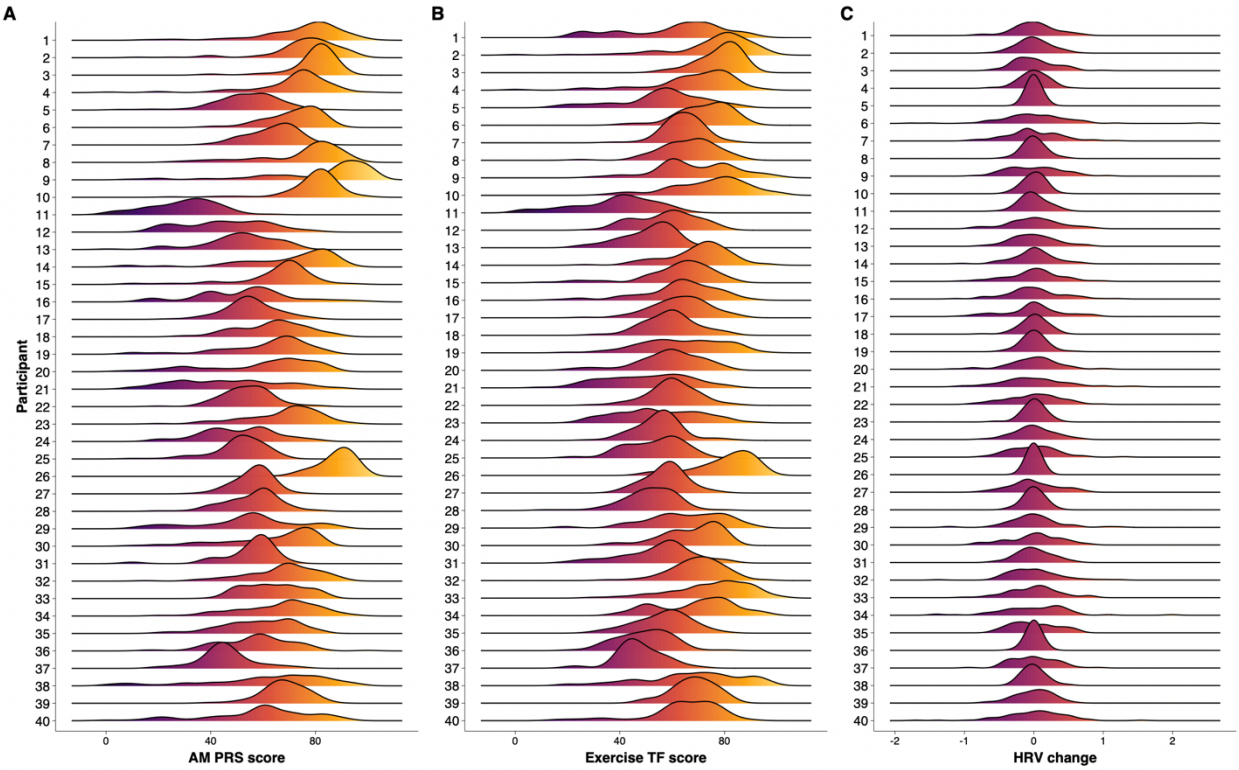
Supplemental figure 3 – 100-pt PRS scale and scale for other subjective measures



Supplemental figure 4 – Daily training volume (hours per day) for each participant for each day of the study.



Supplemental figure 5 – Daily training load (product of session RPE and exercise duration in minutes) for each participant for each day of the study.



Supplemental figure 6 – Density plot showing the distribution of the three main outcome variables for each participant.

References

1. Bourdon PC, Cardinale M, Murray A, et al. Monitoring Athlete Training Loads: Consensus Statement. *Int J Sports Physiol Perform.* 2017;12(Suppl 2):S2161-S2170.
2. Voet JG, Lamberts RP, de Koning JJ, de Jong J, Foster C, van Erp T. Differences in execution and perception of training sessions as experienced by (semi-) professional cyclists and their coach. *Eur J Sport Sci.* 2021:1-9.
3. Halson SL. Monitoring training load to understand fatigue in athletes. *Sports Med.* 2014;44 Suppl 2:S139-147.
4. Impellizzeri FM, Marcora SM, Coutts AJ. Internal and External Training Load: 15 Years On. *Int J Sports Physiol Perform.* 2019;14(2):270-273.
5. Clemente FM, Mendes B, Palao JM, et al. Seasonal player wellness and its longitudinal association with internal training load: study in elite volleyball. *J Sports Med Phys Fitness.* 2019;59(3):345-351.
6. Achten J, Halson SL, Moseley L, Rayson MP, Casey A, Jeukendrup AE. Higher dietary carbohydrate content during intensified running training results in better maintenance of performance and mood state. *J Appl Physiol (1985).* 2004;96(4):1331-1340.

7. Rothschild J, Kilding AE, Plews DJ. Prevalence and Determinants of Fasted Training in Endurance Athletes: A Survey Analysis. *Int J Sport Nutr Exerc Metab.* 2020;30(5):345-356.
8. Van Eetvelde H, Mendonca LD, Ley C, Seil R, Tischer T. Machine learning methods in sport injury prediction and prevention: a systematic review. *J Exp Orthop.* 2021;8(1):27.
9. Mezyk E, Unold O. Machine learning approach to model sport training. *Comput Human Behav.* 2011;27(5):1499-1506.
10. Perri E, Simonelli C, Rossi A, Trecroci A, Alberti G, Iaia FM. Relationship Between Wellness Index and Internal Training Load in Soccer: Application of a Machine Learning Model. *Int J Sports Physiol Perform.* 2021;16(5):695-703.
11. Morgenstern JD, Rosella LC, Costa AP, de Souza RJ, Anderson LN. Perspective: Big Data and Machine Learning Could Help Advance Nutritional Epidemiology. *Adv Nutr.* 2021;12(3):621-631.
12. Rothschild JA, Morton JP, Stewart T, Kilding AE, Plews DJ. The quantification of daily carbohydrate periodization among endurance athletes during 12 weeks of self-selected training: presentation of a novel Carbohydrate Periodization Index. *medRxiv.* 2022.
13. Foster C, Boullosa D, McGuigan M, et al. 25 Years of Session Rating of Perceived Exertion: Historical Perspective and Development. *Int J Sports Physiol Perform.* 2021;16(5):612-621.
14. Clemente FM, Rabbani A, Araujo JP. Ratings of perceived recovery and exertion in elite youth soccer players: Interchangeability of 10-point and 100-point scales. *Physiol Behav.* 2019;210:112641.
15. Laurent CM, Green JM, Bishop PA, et al. A practical approach to monitoring recovery: development of a perceived recovery status scale. *J Strength Cond Res.* 2011;25(3):620-628.
16. Plews DJ, Laursen PB, Stanley J, Kilding AE, Buchheit M. Training adaptation and heart rate variability in elite endurance athletes: opening the door to effective monitoring. *Sports Med.* 2013;43(9):773-781.
17. Mishica C, Kyrolainen H, Hynynen E, Nummela A, Holmberg HC, Linnamo V. Evaluation of nocturnal vs. morning measures of heart rate indices in young athletes. *PLoS One.* 2022;17(1):e0262333.
18. Chinoy ED, Cuellar JA, Huwa KE, et al. Performance of seven consumer sleep-tracking devices compared with polysomnography. *Sleep.* 2021;44(5).
19. Roberts DM, Schade MM, Mathew GM, Gartenberg D, Buxton OM. Detecting sleep using heart rate and motion data from multisensor consumer-grade wearables, relative to wrist actigraphy and polysomnography. *Sleep.* 2020;43(7).
20. Miller DJ, Lastella M, Scanlan AT, et al. A validation study of the WHOOP strap against polysomnography to assess sleep. *J Sports Sci.* 2020;38(22):2631-2636.
21. Zaffaroni A, Coffey S, Dodd S, et al. Sleep Staging Monitoring Based on Sonar Smartphone Technology. *Annu Int Conf IEEE Eng Med Biol Soc.* 2019;2019:2230-2233.
22. Saw AE, Main LC, Gastin PB. Monitoring the athlete training response: subjective self-reported measures trump commonly used objective measures: a systematic review. *Br J Sports Med.* 2016;50(5):281-291.

23. Haddad M, Stylianides G, Djaoui L, Dellal A, Chamari K. Session-RPE Method for Training Load Monitoring: Validity, Ecological Usefulness, and Influencing Factors. *Front Neurosci.* 2017;11:612.
24. Rothschild J, Kilding AE, Stewart T, Plews DJ. Factors influencing substrate oxidation during submaximal cycling: a modelling analysis. *Sports Med.* 2022;In Press.
25. Plews DJ, Laursen PB, Kilding AE, Buchheit M. Evaluating training adaptation with heart-rate measures: a methodological comparison. *Int J Sports Physiol Perform.* 2013;8(6):688-691.
26. Sawczuk T, Jones B, Scantlebury S, Till K. Influence of Perceptions of Sleep on Well-Being in Youth Athletes. *J Strength Cond Res.* 2021;35(4):1066-1073.
27. Kuhn M, Johnson K. *Applied predictive modeling.* Vol 26: Springer; 2013.
28. James G, Witten D, Hastie T, Tibshirani R. *An introduction to statistical learning : with applications in R.* New York: Springer; 2021.
29. Truda G. Quantified Sleep: Machine learning techniques for observational n-of-1 studies. *arXiv preprint arXiv:2105.06811.* 2021.
30. Gudivada VN, Rao D, Raghavan VV. Big data driven natural language processing research and applications. *Handbook of Statistics.* Vol 33: Elsevier; 2015:203-238.
31. Kapoor S, Narayanan A. Leakage and the Reproducibility Crisis in ML-based Science. *arXiv preprint arXiv:2207.07048.* 2022.
32. Yang P, Hwa Yang Y, B Zhou B, Y Zomaya A. A review of ensemble methods in bioinformatics. *Current Bioinformatics.* 2010;5(4):296-308.
33. Biecek P, Burzykowski T. *Explanatory model analysis: Explore, explain and examine predictive models.* Chapman and Hall/CRC; 2021.
34. Greenwell BM, Boehmke BC, Gray B. Variable Importance Plots-An Introduction to the vip Package. *R J.* 2020;12(1):343.
35. Altini M, Plews D. What is behind changes in resting heart rate and heart rate variability? A large-scale analysis of longitudinal measurements acquired in free-living. *Sensors.* 2021;21(23):7932.
36. Shmueli G. To explain or to predict? *Statist Sci.* 2010;25(3):289-310.
37. Kinnunen H, Rantanen A, Kentta T, Koskimaki H. Feasible assessment of recovery and cardiovascular health: accuracy of nocturnal HR and HRV assessed via ring PPG in comparison to medical grade ECG. *Physiol Meas.* 2020;41(4):04NT01.
38. Plews DJ, Scott B, Altini M, Wood M, Kilding AE, Laursen PB. Comparison of Heart-Rate-Variability Recording With Smartphone Photoplethysmography, Polar H7 Chest Strap, and Electrocardiography. *Int J Sports Physiol Perform.* 2017;12(10):1324-1328.
39. Champagne CM, Bray GA, Kurtz AA, et al. Energy intake and energy expenditure: a controlled study comparing dietitians and non-dietitians. *J Am Diet Assoc.* 2002;102(10):1428-1432.
40. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol.* 2019;110:12-22.

Supplemental Table 1. Descriptive statistics of main variables

stat	AM_sleep_qualilty	AM_soreness	AM_stress
Mean	4.5	4.3	3.2
SD	1.2	1.9	1.2
Min	1	1	1
Max	7	9	8

stat	diet_carb_g	diet_carb_g_kg	diet_fat_g	diet_fat_g_kg	diet_kcal	diet_kcal_kg	diet_protein_g	diet_protein_g_kg
Mean	281.7	4	116.6	1.6	2759.6	39	136.7	1.9
SD	146.2	2	52.8	0.7	866.9	11.9	46	0.6
Min	0	0	9.2	0.1	626	8	1	0
Max	1351	17.6	386	6.1	7378	122.9	386	5.3

stat	exercise_CARB_before	exercise_duration_min	exercise_load	exercise_RPE_max	exercise_RPE_weighted	exercise_wrkts_per_day
Mean	43.4	102.7	388.3	41.5	36.1	1.5
SD	36.3	74.6	415.5	23.2	20.1	0.9
Min	0	0	0	1	1	0
Max	250.3	589.7	3852.9	120	120	8

stat	roll_monotony	roll_strain	sleep_hours	sleep_index	pulse
Mean	1.4	3734.3	7.5	33.9	49.5
SD	0.7	2671.7	1.1	10.6	7.1
Min	0.4	7.6	2.1	3.5	30
Max	12.7	20571.9	11.5	77	82

For explanation of variables refer to Table 1 of main paper

Model Info Sheet for Detecting and Preventing Leakage in ML-based Science

Section 1: Information about paper or report

1) Author(s): Jeffrey A. Rothschild, Tom Stewart, Andrew E. Kilding, Daniel J. Plews

2) Title of the paper or report which introduces the model

Predicting daily recovery during long-term endurance training using Machine Learning analysis

3) DOI or permanent link to the paper or report (for example, link to arxiv.org webpage)

4) License: Under which license(s) are the data and/or model shared?

5) Email address of the corresponding author

Jeffrey.Rothschild@aut.ac.nz

Section 2: Scientific claim(s) of interest

6) Does your paper make a generalizable claim based on the ML model?

Our models aid the prediction of day-to-day recovery status among endurance athletes

7) Is the scientific claim made about a distribution or population from which you can sample?

Yes
If yes: (a) what is the population or distribution about which the scientific claim is being made?

Endurance athletes training ≥ 6 hours per week

(b) What is the sample used for the study?

Male and female endurance athletes aged 18 or older who train at least seven hours per week and use a smartphone app to track their dietary intake, heart rate variability (HRV), and sleep.

8) Does the scientific claim only apply to certain subsets of the distribution mentioned in Q6?

Our model might not generalize to athletes not currently tracking the various metrics.

Section 3: Train-test split is maintained across all steps in creating the model

9) Train-test split type: How was the dataset split into train and test sets?

For group models, data were split into a training set (75%) and a testing set (25%). To avoid data leakage, all observations from a given participant were assigned to either the training or testing set, and preprocessing steps such as standardization and calculations for removal of highly correlated variables were performed only using the training set.

For individual models there were not enough data points for a test-train split, so cross-validation was used for hyperparameter tuning and accuracy metrics were calculated using 500 bootstrap resamples.

10) Are there duplicates in the dataset? If yes, explain how duplicates are handled to ensure the train-test split.

No.

11) In case the dataset has dependencies (e.g., multiple rows of data from the same patient), describe how the dependencies were addressed (for example, using block-cross validation).

All data splits (test-train and cross-validation) were performed so that a given participant could not be in both the training and testing sets. To address potential issues with autocorrelation, a process of Markov unfolding was used (described in the manuscript).

12) List all the pre-processing steps used in creating your model. For example, imputing missing data, normalizing feature values, selecting a subset of rows from the dataset for building the model.

Missing data was imputed at the individual level, prior to joining data into a grouped dataset. Multiple linear regression and nearest neighbor algorithms were used for diet and training measures, and median values were used for other variables

Normalization, the removal of variables with zero variance, and the removal of highly correlated variables was performed on the training data, using the recipes R package in the Tidymodels framework.

13) How was the train-test split maintained during each pre-processing step? If applicable, use a separate line for each step mentioned in Q14.

The test-train split is maintained throughout the process in the Tidymodels framework.

14) List all the modeling steps used in creating your model. For example, feature selection, parameter tuning, model selection.

- Highly correlated variables were removed from the training set
- Parameter tuning was performed on the 9 different algorithms using a workflow set in the workflowsets R package.
- Models were selected in two main ways:
    - Group models – the tuned models were fed into the Stacks R package, which then created an ensemble using bootstrap resampling
    - Individual models – the model (and hyperparameter set) with the lowest RMSE was selected as the chosen model

15) How is the train-test split is observed during each modeling step? If applicable, use a separate line for each step mentioned in Q16.

The test-train split is maintained throughout the process in the Tidymodels framework.

16) List all the evaluation steps used in evaluating model performance. For example, cross-validation, out-of-sample testing.

Group models – Accuracy metrics were established using the previously unseen test dataset
Individual models – Accuracy metrics were established using 500 Bootstrap resamples

17) How is the train-test split observed during each evaluation step? If applicable, use a separate line for each step mentioned in Q18.

Group models – Accuracy metrics were established using the previously unseen test dataset.
Individual models – Bootstrap resamples were used in lieu of a test-train split.

Section 4: Test set is drawn from the distribution of scientific interest.

18) Why is your test set representative of the population or distribution about which you are making your scientific claims?

Data were collected and analyzed using 40 athletes from this population.

19) Explain the process for selecting the test set and why this does not introduce selection bias in the learning process.

The test set was randomly selected from the initial dataset.

20) In case your model is used to predict a future outcome of interest using past data, detail how data in the training set is always from a date earlier than the data in the test set.

Although lagged values were included in the test set, the test-train split was performed at the participant level (meaning no participants were in both sets).

Section 5: Each feature used in the model is legitimate for the task

21) List the features used in the model, alongside an argument for their legitimacy. A legitimate feature is one that would be available when the model is used in the real world and is not a proxy of the outcome being predicted. You can also include this list in an appendix and reference the relevant section of your Appendix here.

Category	Variables
Training, dietary intake, sleep duration	These are variables that are routinely tracked by athletes and coaches and would be expected to have the largest influence on day-to-day recovery
Engineered features	We created 7-d moving averages of several factors relating to training load and dietary intake, with the assumption that accumulated training load (and/or energy deficit/surplus) could influence day-to-day recovery. We also calculated a sleep index (sleep duration x quality) based on other published findings.
Subjective measures	Athletes daily perceptions of soreness, life stress, and sleep quality were considered an important reflection of real-world practices by coaches.
Non-exercise	As objective recovery measures, we included resting HRV and resting HR (daily, change from previous day, and 7-d moving averages of each)
Subject characteristics	Participant ID, age, HRV app, sleep app, percentage of missing data, competitive level, primary sport, training age, body weight, day of the week. These could help to explain variation between athletes.
For outcome measures recorded in the morning (AM PRS, HRV change), any variables that occurred later the same day were excluded from the modeling (e.g., how someone felt during exercise that day, or their dietary intake that day).	