Monitoring Stress-Recovery in athletes

Introduction

Performance has become increasingly significant not only in sports but also across various domains, such as individual and organizational productivity. This growing emphasis on optimizing performance raises important questions: How can peak performance be achieved? To what extent does a competitive environment affect mental health? And how do mental and physical well-being influence performance outcomes? This preliminary study situates itself within this framework, aiming to evaluate the effectiveness of different psychometric instruments designed to monitor the balance between stress and recovery. Specifically, the objective is to identify the most effective tools for monitoring athletes' well-being and preventing mental and physical burnout.

Objectives

This study evaluates the performance of the RESTQ and the Hooper Index, two widely used questionnaires in sports psychology in tracking athletes' responses to training. Questionnaires are practical, cost-effective tools that address relevant challenges in the field; however, there are trade-offs associated with their use. Specifically, there is a question of whether it is preferable to use a more comprehensive tool (which may take longer to complete) or a simpler, more time-efficient assessment that minimizes "questionnaire fatigue" and reduces the likelihood of participant drop-out.

This study also aims to assess individual-specific patterns in the processes regulating stress-recovery balance and dynamics. Emerging psychological literature highlights heterogeneity in individual mechanisms, and we seek to determine whether similar diversity exists in this field, aiming to uncover new insights.

Sample

Our data was collected from a sample of 16 elite athletes (10 male, 6 female) during the pre-season training phase. One female participant was excluded from the analysis due to a high proportion of missing data. The athletes belong to three distinct groups: a predominantly male group (M1), an all-male group (M2), and an all-female group (F).

Method

We conducted a six-week training monitoring period, structured into two mesocycles. Each mesocycle consisted of two weeks of high-intensity, high-volume training (HI), followed by one week of low-intensity recovery. We assessed various constructs using three questionnaires:

1. *RPE-Method for Training Load*: Administered after each training session to capture perceived exertion and overall training load.
2. *Hooper Index (Wellness)*: Completed every morning to monitor daily fluctuations in well-being.
3. *Recovery-Stress Questionnaire* (RESTQ-S): Administered six times throughout the study. Specifically, *date_1* and *date_4* mark the end of the first HI week of the mesocycles, *date_2* and *date_5* indicate the end of the HI periods, and *date_3* and *date_6* represent the start of a new HI block following the recovery week.

The Hooper Index assesses multiple constructs, with four single-item subscales for Sleep, Stress, Fatigue and Delayed Onset Muscle Soreness (DOMS). While lacking psychometric validation, we chose to evaluate this tool due to its frequent application in both practice and literature.

The RESTQ-S measures the recovery-stress balance with general and specific subscales assessing social, emotional, mental, and physical stress, offering a comprehensive view of an athlete's recovery and stress states over time.

Analysis

The analysis is divided into five main stages:

Preprocessing: We began by selecting relevant data and excluding one female participant due to a high proportion of missing values. During this phase, we also calculated the Acuto measure, an Exponential Moving Average with a 7-day window applied to the Training Load data to account for the diminishing effects of training over time. Given the athletes' experience and optimized recovery strategies, we considered a 7-day window effective for capturing the impact of past training on the target day's perceived load.

Imputation Strategy Evaluation: We need to address the missing data, which we assume is primarily due to forgetfulness. Therefore, we can confidently categorize the missing data as Missing Completely At Random (MCAR). We evaluated different imputation methods (KNN, MICE, and Median) by first building a reference dataset using mean imputation. Next, we modelled and predicted the imputed data using KNN, MICE, and Median, and evaluated errors against the reference dataset using RMSE and MAE as error metrics. To account for sampling error, we bootstrapped the data to establish confidence intervals for RMSE and MAE for each imputation method. This step identified KNN imputation as the most effective strategy, which we then applied for the final dataset. However, due to the small sample size, the RMSE and MAE values are uncertain, making it difficult to definitively determine which technique is superior.

Psychometric Analysis: We applied the Min-Max normalization to our dataset to ensure that all features are scaled to a common range. Then, we conducted statistical analyses to assess the performance of the two psychometric instruments, RESTQ and Hooper Index. We calculated the mean squared error (MSE) and generated a correlation matrix that included Acuto as a reference construct.

Clustering Modeling: We applied various machine learning models to explore patterns within the data. For unsupervised modeling, we used PCA and an Affinity Propagation to identify potential groupings. As a supervised model, we employed Decision Trees, using the different athlete groups as target labels to train the model and assess which features were most predictive of group membership.

*GIMME-MS Analysis**:*** We used Group Iterative Multiple Model Estimation for Multiple Solutions (GIMME-MS) to conduct a within-person analysis, leveraging the longitudinal assessments we collected. This approach constructs structural equation models (SEM) to evaluate both same-day and one-day lagged relationships between variables. Additionally, GIMME-MS can identify clusters within the sample and generate a general SEM that represents the entire sample. Only Wellness and TL could be included in the GIMME model because RESTQ was not measured daily like the other variables.

Results

The missing data ratio, defined as the proportion of missing values to the total number of values in the dataset, is 0.20 for the Wellness data and 0.044 for the RESTQ data.

For Wellness, the confidence interval (CI) using KNN imputation is [3.8446, 4.6469] for RMSE and [2.6188, 3.2795] for MAE. In contrast, the CI using Median imputation is [3.6295, 4.5075] for RMSE and [2.4452, 3.1164] for MAE. Similar patterns were observed for RESTQ: with KNN, the CI is

[4.4109, 4.8170] for RMSE and [3.1896, 3.5628] for MAE; for Median, the CI is [4.3799, 4.8212] for RMSE and [3.1904, 3.5291] for MAE. KNN performed better in terms of RMSE, although it was not superior for MAE; nonetheless, we chose KNN for imputation.

Regarding the mean squared error (MSE), the MSE for Wellness is 5.28, while for RESTQ it is 9.89. This finding is further illustrated in the plots: the general trend of the Hooper Index (*Figure 2b*) fits more closely with the trend of the Acuto measure (*Figure 1*) compared to the trend of the RESTQ (*Figure 2a*).

In *Figure 1*, the red line represents the fraction of zeros per Acuto. During the setup of our data collection, we overlooked that a single value of zero could convey two distinct meanings (“zero’s double meaning”): it could either indicate "I didn’t train today" or signify "I forgot to fill out the questionnaire", in which case it should be treated as a "NaN" value. As a result, we were unable to impute the "NaN" values for Training Load. The presence of a significant proportion of zeros leads to a reduction in standard deviation (SD). However, this decrease cannot be interpreted as a reduction in noise; rather, it reflects a lack of signal in the data. This phenomenon is further illustrated in the accompanying scatterplot, which demonstrates the negative correlation between SD and the proportion of zeros (*Figure 3*).

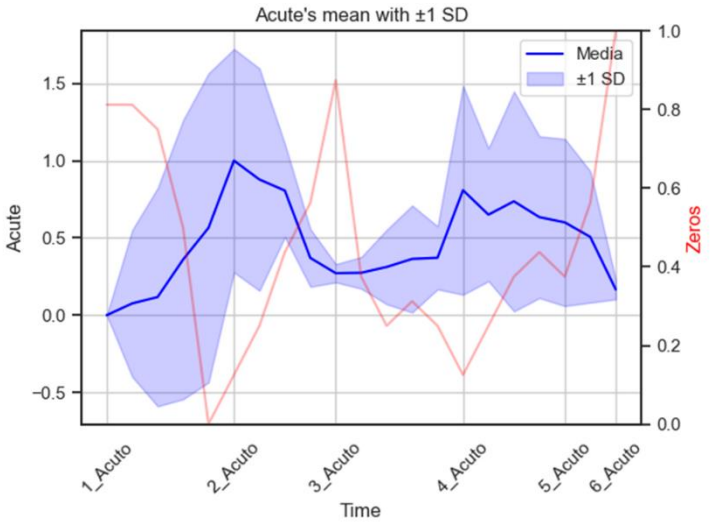

Figure 1. Trend of Acuto over time, with standard deviation shaded. The right axis displays the proportion of zeros, represented by the red line across the days. This line is as a rule of thumb for adjusting variance in relation to signal dispersion.

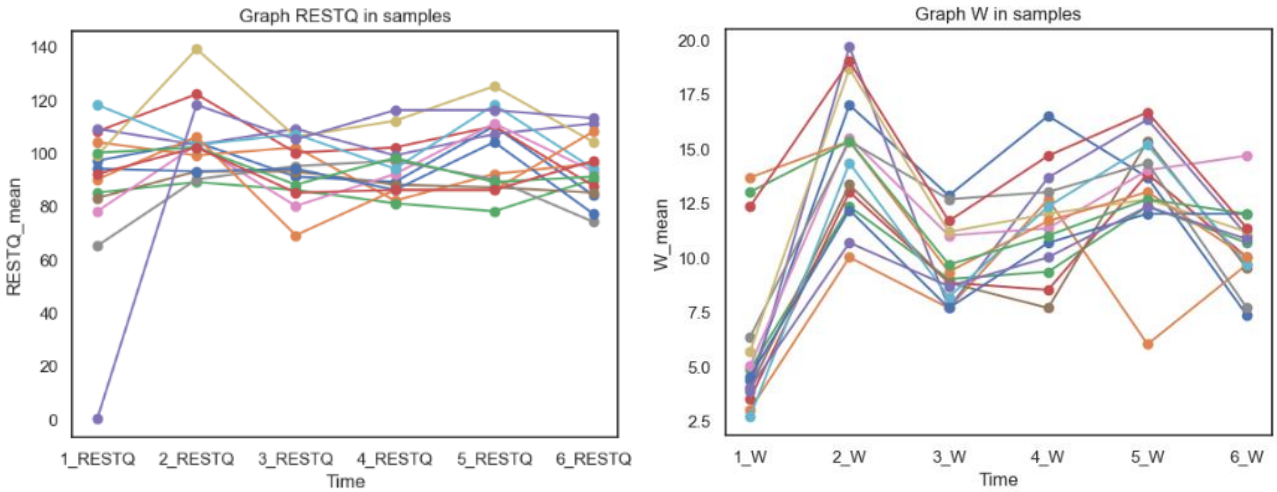

Figure 2. (a) Trend of the Hooper Index (W) across samples over time. (b) Trend of RESTQ across samples over time. Both y-axes are absolute values of Hooper Index and RESTQ.

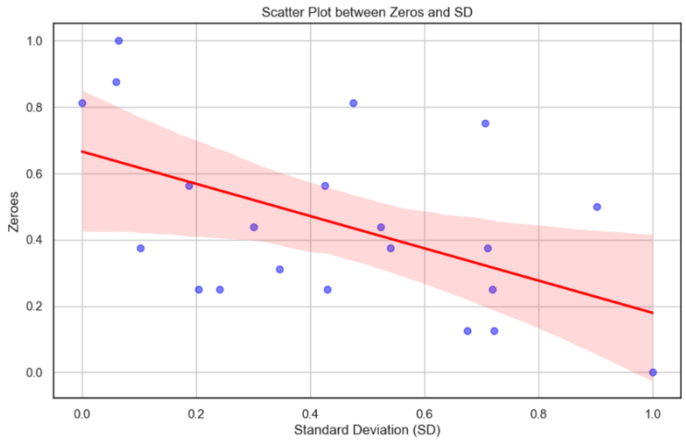

Figure 3. Scatterplot and regression analysis of standard deviation (SD) against the proportion of zeros. The negative correlation suggests a lack of signal as the proportion of zeros increases.

We also analyzed the correlation between the various constructs measured during training monitoring (*Figure 4*). As anticipated from our trend analysis and MSE, the correlation between W and Acuto was stronger ($r = 0.22$) than that between RESTQ and Acuto ($r = 0.15$). Interestingly, the correlation between W and RESTQ was almost zero ($r = -0.065$), indicating that while both questionnaires are frequently utilized in similar contexts with overlapping goals, they assess different constructs. This observation aligns with recent literature trends, which suggest that single-item scales can be both efficient and reliable in capturing distinct dimensions of measurement. However, the trade-off is that single-item scales may not provide a nuanced assessment of the construct, limiting their ability to capture the full complexity of the underlying phenomenon.

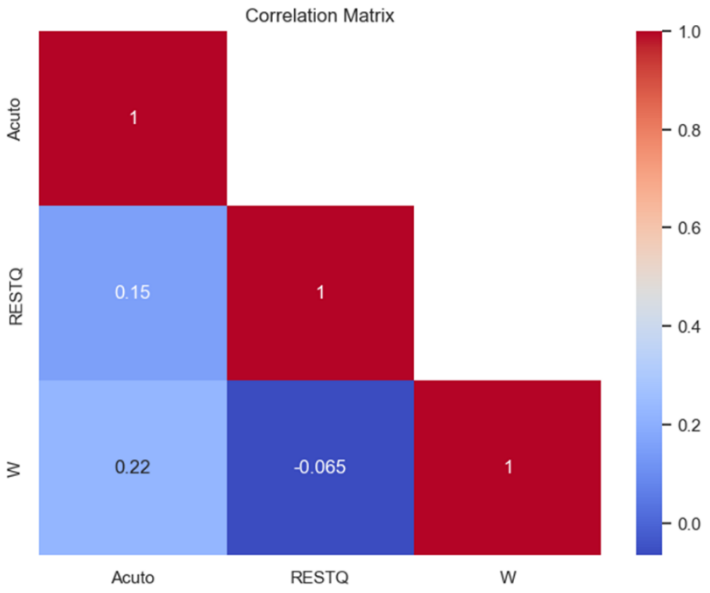

Figure 4. Correlation matrix of the measures.

We also attempted to run some machine learning models to uncover potential hidden non-linear trends in the data. We acknowledged that, given the small sample size, this was merely exploratory, and that the analysis might not yield reliable insights.

Affinity Propagation identified four clusters within the data and highlighted the sample “Skier-15” as an outlier, a finding corroborated by the PCA analysis (*Figure 5b*). This outlier designation arose from her data missingness being close to the cutoff. However, we decided to include this sample in our analysis to avoid excessive sample reduction. The clusters did not provide additional insights into the data, likely due to the small sample size.

The PCA provided intriguing insights: the principal components (PCs) of within-sample PCA (*Figure 5a*) were aligned with different measures, where PC1 represented the variance of RESTQ and PC2 represented the variance of the Hooper Index (W). Notably, PC1 and PC2 were bipolar concerning sex (“Genere_M”), potentially separating male and female participants.

In the PCA between samples (*Figure 5b*), we observed a slight alignment of the PCs along sex: the female (last four skiers in the image) showed higher loadings on PC1 compared to PC2. Interestingly, “Skier-1,” a female sample who trains with the male-predominant group (M1), also loaded more on PC1 compared to male samples. This observation opens avenues for theoretical speculation that we will address further. As previously mentioned, “Skier-15” was identified as an outlier. She exhibited strong loadings on the PCs (especially PC2), and this unusual pattern may be explained by her data missingness.

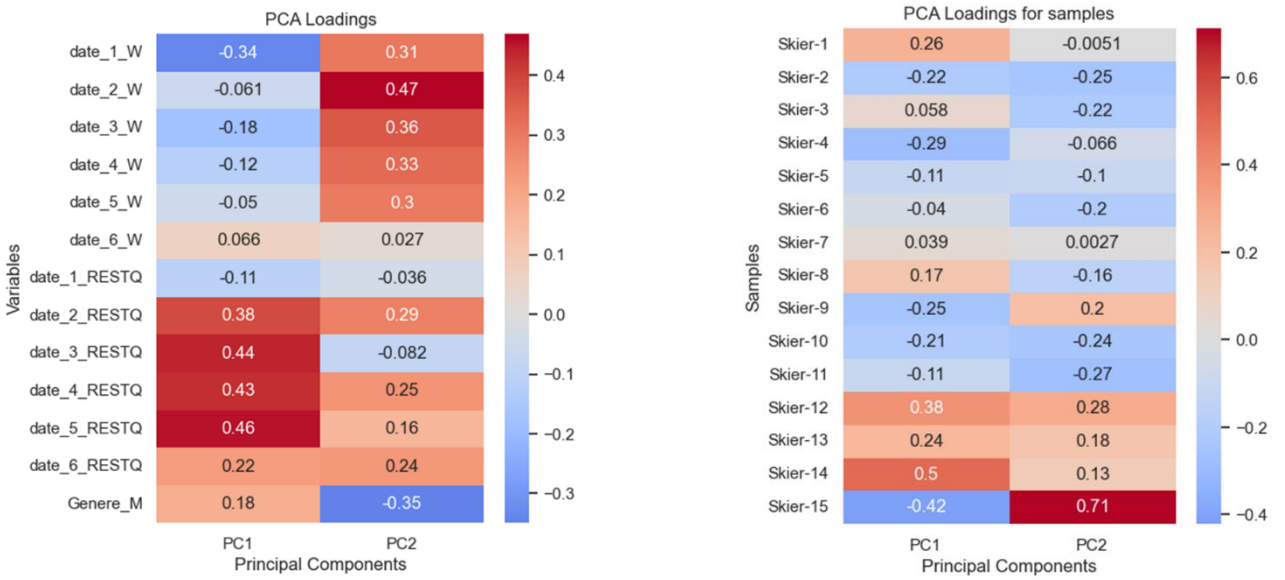

Figure 5. (a) PCA analysis within samples. PC1 aligns with the variance of RESTQ, while PC2 is more aligned with the variance of the Hooper Index. (b) PCA analysis between samples. PC1 effectively distinguishes the sex of the samples, with “Skier-15” exhibiting a unique pattern due to its data missingness.

We also employed a supervised model (Decision Tree Classifier), using group division as the target label. Our objective was to identify which variables had the most significant impact on the model's predictions (*Table 1*). Unsurprisingly, sex emerged as the most influential variable by far. This finding aligns with the patterns observed in the unsupervised models and reflects the fact that the training groups (M1, M2, F) were primarily defined by sex division.

Feature	Importance
Genere	0.395833
date_2_W	0.281250
date_4_W	0.166667
date_6_W	0.156250
date_1_W	0.000000
date_3_W	0.000000
date_5_W	0.000000
date_1_RESTQ	0.000000
date_2_RESTQ	0.000000
date_3_RESTQ	0.000000
date_4_RESTQ	0.000000
date_5_RESTQ	0.000000
date_6_RESTQ	0.000000

Table 1. Feature importance in the trained Decision Tree Classifier. It shows how much each feature reduces impurity across all trees in the model, calculated by summing impurity reductions at each split where the feature is used.

The final stage of the analysis involved a time-series approach using the GIMME-MS model. This model did not reveal any meaningful general patterns across the entire sample, nor did the identified subgroups provide significant insights. More revealing, however, were the individual-specific patterns: as anticipated, these models highlight the substantial heterogeneity characterizing stress-recovery dynamics.

Figure 6 displays several individual-specific models, immediately highlighting the dominance of heterogeneity, as each model is unique. When considering general observations, we note that TL (Training Load) has a lagged effect on other variables, particularly on DOMS (Delayed Onset Muscle Soreness), which aligns with its construct. This suggests that the body's response to training unfolds over time, eliciting varied reactions related to different perceptions, such as sleep quality, stress, DOMS, and more. We also observe that TL does not consistently elicit DOMS. This may be due to two factors: (1) insufficient statistical power and a lot of noise, which make it challenging to detect a uniform trend; (2) varying response levels among athletes, where those with a higher lean mass index or greater strength might experience less muscle damage (DOMS) compared to athletes with different body compositions and performance. Further studies could investigate this mechanism in greater depth.

Sleep often links to DOMS (directly or indirectly) and Fatigue, with the relationship sometimes showing Fatigue as a cause of poor sleep, or the reverse. This underscores the importance of Sleep as a recovery indicator.

The GIMME-MS model can enhance our understanding of functioning mechanisms within a specific sample, making it particularly valuable for individualized insights into each athlete. This approach allows both speculative exploration and practical optimization. For example, in *Figure 6-f*, we see that Sleep contributes to increased perceptions of Fatigue and Stress. This suggests that a single-item Sleep scale could effectively monitor athlete recovery, potentially reducing the number of items in the questionnaire. It's important to note, however, that this approach is only advisable when the questionnaire lacks psychometric validation; modifying rigorously validated questionnaires can compromise their psychometric properties, reducing validity and reliability.

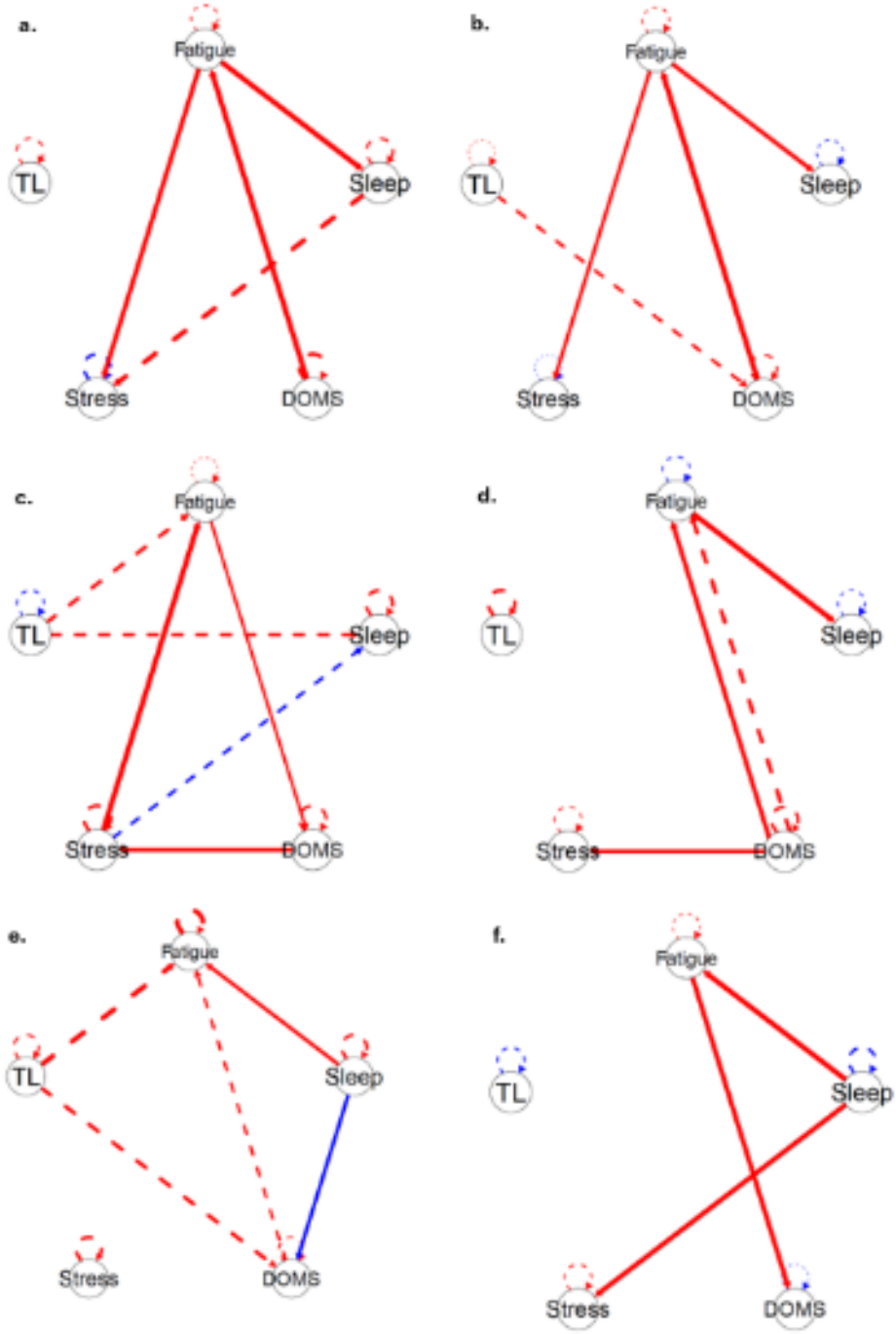

Figure 6. Sample-specific diagrams illustrate functional associations within athletes. Red arrows represent positive correlations, while blue arrows indicate negative correlations. Solid arrows denote same-day associations, and dashed arrows indicate lagged associations. Arrow thickness reflects the magnitude of these connections.

Discussion

This preliminary study suggests that the Hooper Index may be more effective than RESTQ for monitoring the stress-recovery balance: it has a lower mean-squared error (MSE) than RESTQ and shows a stronger correlation with Acuto. Although lacking formal statistical validation, the Hooper Index's simplicity as a single-item scale supports its practical value. Notably, this simplicity is advantageous only in contexts where social desirability bias is minimized, as the explicit content of single-item scales can heighten this bias. In our setting, social desirability may be low due to the athletes' experience and the collaborative environment of the team.

Machine learning models identified additional outliers within the sample set. Principal Component Analysis (PCA) managed to separate samples by sex, even revealing a female component within the M1 group. This finding raises an intriguing question: how could PCA detect "sex" differences despite a male-oriented training plan? To explore this, we removed sex from the feature set and recalculated the PCA (*Figure 7*): since sex was among the features, it was possible that the PCs captured the variance in this variable. After excluding sex, the pattern became less distinct, confirming that the PCs had initially leveraged sex-related variance. However, a slight pattern remained, as PC1 continued to align with female samples, suggesting some underlying difference between sexes. Two potential mechanisms may explain this outcome: (1) there are sex differences in training response, or (2) M1's coach adapted the training plan similarly to the approach used by F's coach. This could indicate that both coaches tailor their general training schedules according to sex, resulting in a resemblance between Skier-1's (female in M1) plan and the F group's training plan (Skier-12, Skier-13, Skier-14).

The GIMME model provided an additional perspective to the analysis, allowing for person-specific evaluations that enhance our understanding of individual athletes. This type of analysis facilitates personalized training and assessment, which traditional approaches often struggle to achieve. The heterogeneity of the models underscores the necessity of a tailored understanding of each athlete. Due to the different measurement resolutions, RESTQ could not be included in the model. Future scientific projects could incorporate additional variables into GIMME for a more comprehensive understanding of the phenomena. It's important to note that the person-specific plans are influenced by missing data and noise, along with the relatively short longitudinal measurement period, which may contribute to a lack of confidence in the results.

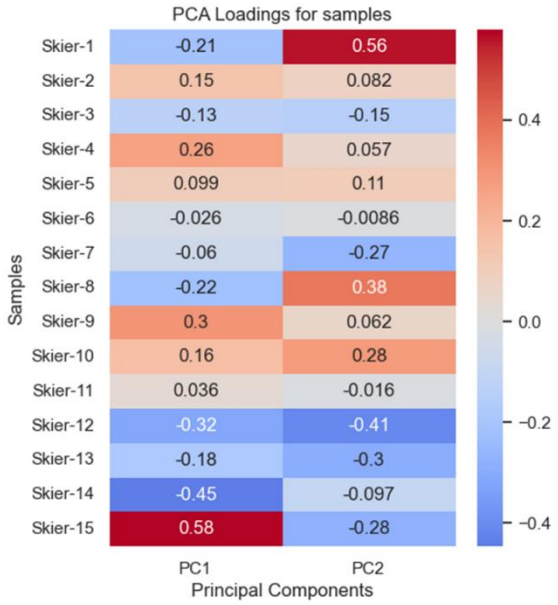

Figure 7. PCA analysis between samples excluding sex from the feature set. While the sex-related trend in PC1 diminishes significantly, it remains partially detectable.

Strengths and Limitations

Several factors could influence the results of this preliminary study. First and foremost, the sample size is relatively small, which limits our ability to detect fine patterns and achieve stable statistical evaluations, particularly concerning imputation. Another significant limitation is the issue of “zero’s double meaning”; having two interpretations for a single numerical value compromised the precision of our analysis. Additionally, our analysis assumes that Acuto is an unbiased and reliable measure of training impact; we used it as a reference for evaluating both the Hooper Index and RESTQ based on this premise. However, it’s not certain that Acuto is entirely unbiased; we chose it primarily due to its widespread use. Generally, every measure has its own errors, strengths, and limitations. From our perspective, the most effective way to achieve a comprehensive assessment is through a holistic approach that monitors multiple aspects of the athlete’s well-being.

Additionally, a more homogeneous group in terms of training schedules would likely improve the results. It would also be beneficial to measure the constructs with greater frequency (especially for RESTQ). However, this presents a trade-off between extensive longitudinal monitoring and detailed measurements; attempting to maintain fine-grained measurements over a long duration could lead to participant drop-out.

This evaluation also has several strengths. The sample consists of elite athletes, where the impact of training on their lives is substantial, resulting in more pronounced effects that are easier to detect. Additionally, the study demonstrates good ecological validity: the athletes trained as planned, and no adjustments were made during the data collection. Furthermore, the heterogeneity of the samples highlights the challenges of effectively monitoring the stress-recovery balance outside of experimental and controlled settings.

Supplementary Materials

The analysis’s code will be uploaded to my GitHub page in "Sport-analysis" directory (https://github.com/Luca-Sclisizzo).

Acknowledgements

We would like to thank all the athletes and staff who enthusiastically welcomed our project and were open to exploring new perspectives.