International Journal of Clinical and Health Psychology 26 (2026) 100727

Contents lists available at ScienceDirect

International Journal of Clinical and Health Psychology

journal homepage: www.elsevier.es/ijchp

Limited prediction of self-rated performance in elite female footballers

Ailikute Aikeremu [a,#], Xiangqin Song [c,#], Siyu Pan [b], Shubo Xu [c], Vahid Farrahi [d], David Moreau [e], Matthew Heath [f,g,h], A. Mark Williams [a], Markus Raab [i], Ale Quartiroli [j,k], Alyx Taylor [l], Fred Paas [m,n], Gangyan Si [b,*], Liye Zou [a,**]

[a] Body-Brain-Mind Laboratory, School of Physical Education & Sports Science, South China Normal University, Guangzhou, China
[b] School of Psychology, Wuhan Sports University, Wuhan, China
[c] School of Physical Education, Beijing Normal University, Beijing, China
[d] Research Group of Data Analytics in Sport Science, Institute for Sports and Sport Science, TU Dortmund University, Dortmund, Germany
[e] School of Psychology and Centre for Brain Research, University of Auckland, Auckland 1030, New Zealand
[f] School of Kinesiology, Faculty of Health Sciences, University of Western Ontario, 1151 Richmond Street, London, ON N6A 3K7, Canada
[g] Graduate Program in Neuroscience, University of Western Ontario, 1151 Richmond Street, London, ON N6A 3K7, Canada
[h] Canadian Centre for Activity and Aging, University of Western Ontario, 1201 Western Road, London, ON N6G 1H1, Canada
[i] Institute of Psychology, Department of Performance Psychology, German Sport University Cologne, Cologne, Germany
[j] Psychology Department, University of Wisconsin–La Crosse, La Crosse, WI 54601, USA
[k] School of Psychology, Sport, and Health Sciences, University of Portsmouth, Portsmouth PO1 2EG, United Kingdom
[l] Health Sciences University, United Kingdom
[m] Department of Psychology, Education, and Child Studies, Erasmus University Rotterdam, Rotterdam, The Netherlands
[n] School of Education, University of New South Wales, Sydney, NSW, Australia

ARTICLE INFO

Keywords:
Athlete monitoring
Individual baseline
Self-rated performance
Women's football
Prediction modelling

ABSTRACT

Background: Routine athlete monitoring captures psychological and physical states across training and match cycles. Whether these indicators distinguish subsequent self-rated match performance below an athlete's own prior ratings remains uncertain.

Objective: This retrospective exploratory prediction study examined whether match-day minus one (MD-1) monitoring indicators identified below-baseline self-ratings of overall team performance and of each player's own offensive and defensive performance among elite female footballers competing in Norway.

Methods: SoccerMon provided 207 player-match records from 30 elite female footballers in two Toppserien clubs during 2020–2021. Requiring at least three strictly earlier match reports retained 124 records from 25 players. Psychological-behavioural, physical-status and joint predictor sets were evaluated using logistic regression and XGBoost. Below-baseline ratings were more than 0.50 points below the historical mean. Player-grouped validation used nested threshold selection. Average precision (AP) was the primary metric.

Results: Below-baseline prevalence was 31.5%, 26.6% and 32.3% for team, offensive and defensive ratings. The joint logistic reference model yielded AP values of 0.450 (95% CI 0.346–0.570), 0.308 (0.245–0.409) and 0.280 (0.241–0.363), respectively; ROC-AUCs were 0.590, 0.496 and 0.401. High false-discovery proportions limited classification utility.

Conclusions: Discrimination was limited and inconsistent. These findings do not support clinical assessment, athlete selection or applied decision use. Individual historical references provide a transparent exploratory outcome definition; objective performance validity and transportability require direct investigation.

* Corresponding author: Gangyan Si, School of Psychology, Wuhan Sports University, Wuhan, China
** Corresponding author: Liye Zou, Body-Brain-Mind Laboratory, School of Physical Education & Sports Science, South China Normal University, Guangzhou, China
E-mail addresses: gangyansi@outlook.com (G. Si), liyezou123@gmail.com (L. Zou).
# These authors contributed equally to this work.

https://doi.org/10.1016/j.ijchp.2026.100727
Received 30 July 2026; Accepted 30 September 2026
Available online 30 October 2026
1697-2600/© 2026 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Introduction

Routine athlete monitoring is used to track training load, recovery, readiness and subjective well-being across training and match cycles (Bourdon et al., 2017; Halson, 2014; Kellmann et al., 2018; Saw et al., 2016). An important performance question is whether these pre-match indicators distinguish subsequent match appraisals that fall below an athlete's usual level. In this study, mood, stress, fatigue, sleep and readiness are candidate predictors of self-rated match performance, not mental-health outcomes. They cannot diagnose mental disorder or substitute for validated mental-health assessment (Reardon et al., 2019).

Self-reports are generated within athlete–staff and organisational relationships. Psychological safety can affect willingness to raise concerns (Edmondson, 1999; Edmondson & Lei, 2014), and research with young elite athletes identifies stigma, confidentiality, prior experiences and relationships with coaches or providers as relevant to help-seeking (Gulliver et al., 2012). Athlete monitoring also raises privacy, confidentiality and coercion concerns (Karkazis & Fishman, 2017). These literatures justify reporting the source setting and treating reporting conditions as possible influences; they do not establish that such mechanisms operated in this cohort. The observed setting comprises two clubs competing in Norway's Toppserien during the 2020–2021 seasons. These are sample descriptors, not measured cultural constructs, and they do not establish athletes' nationality, beliefs or disclosure norms.

This framing raises a measurement problem for raw monitoring and performance ratings. Athletes may use scales differently, and stable between-player differences may obscure within-athlete change (Baumgartner & Steenkamp, 2001; Curran & Bauer, 2011; Wang & Maxwell, 2015). Group-level associations cannot automatically be assumed to describe individual processes in repeated self-report data (Fisher et al., 2018; Molenaar, 2004; Zentgraf et al., 2024). Person-specific centring shifts the performance target from a between-athlete rating level to a deviation from the athlete's own reference. For prediction, that reference must use only ratings dated before the target match. This aligns with individualised monitoring and reduces reliance on a common raw-score threshold (Curran & Bauer, 2011; Hamaker et al., 2015; Hecksteden et al., 2017). It does not, by itself, remove response bias or demonstrate an individual causal process. Accordingly, the present analysis combines historically defined outcomes with player-grouped validation.

Three considerations motivate the present analysis. First, female athletes remain underrepresented in sport and exercise science research (Costello et al., 2014; Cowley et al., 2021; Emmonds et al., 2019), while load and readiness monitoring in elite women's football requires population-specific evidence (Luteberget et al., 2021; Beato et al., 2024). Second, this study addresses self-rated match performance relative to a player's earlier ratings, rather than a common threshold applied to all players. Related SoccerMon work examined next-day readiness decline (Long et al., 2026), a different target from post-match performance appraisal. Third, evaluating logistic regression and XGBoost against simple benchmarks can establish whether these routine reports contain useful predictive information, including when performance is weak (Collins et al., 2024). The empirical gap is therefore prediction of historically defined match appraisals, with explicit reporting of the sample setting and limits of transferability.

To this end, we examine whether match-day minus one (MD-1) monitoring indicators, configured as psychological-behavioural, physical-status, and joint predictor sets, could identify below-baseline self-ratings of overall team performance and of each player's own offensive and defensive performance among elite female footballers competing in Norway. L2-regularised logistic regression and XGBoost were used as complementary modelling approaches. Because algorithm comparison was not a substantive aim, interpretation focused on the discriminative information contained in the available indicators rather than on which classifier performed better (Christodoulou et al., 2019). All model interpretation was descriptive and hypothesis-generating. The immediate contribution is to quantify discrimination, uncertainty and false-alert burden before any applied use is considered. Given the exploratory design, no directional hypotheses were specified for predictor-set or classifier performance; the analysis instead assessed whether MD-1 indicators contained discriminative information for below-baseline performance across the three outcome domains.

Materials and methods

Study design

This study was a secondary retrospective observational prediction-modelling analysis of the open-access SoccerMon dataset. It examined predictive discrimination for below-baseline player-match performance, defined relative to each athlete's own performance history. The design did not estimate causal effects and was not intended as clinical screening, intervention evaluation, or applied decision support (Bouwmeester et al., 2012; Collins et al., 2024; Steyerberg et al., 2010).

The source study obtained written informed consent covering participation and open publication of anonymised data and reported approval by the Norwegian Privacy Data Protection Authority (reference 296155), together with exemption from regional medical and health research ethics committee review (Midoglu et al., 2024). The released data permit secondary research use. The present analysis used deidentified records without participant recontact.

Data Source and Study Context

Data were drawn from SoccerMon, a dataset from two elite women's football teams in Norway's Toppserien during 2020–2021 (Midoglu et al., 2024). The analysis used subjective monitoring, derived internal-load summaries and post-match ratings, not GNSS measurements. Psychological safety, trust, perceived surveillance, disclosure climate, age, playing position, nationality, ethnicity, language and migration background were unavailable in the deidentified analytic files. The setting descriptors therefore delimit the source sample and do not supply unmeasured participant characteristics. Table 1 summarises the variables, domains, scales and operational definitions.

Table 1
SoccerMon variables used in the present analysis: definitions and scales.

Variable	Domain	Scale	Operational definition
Mood	Psychological	1–5	MD-1 self-reported affective state
Stress	Psychological	See note	MD-1 perceived stress; source scale discrepancy noted below
Fatigue	Recovery	1–5	MD-1 self-reported general fatigue
Readiness to train	Recovery	1–10	MD-1 readiness to train
Sleep duration	Behavioural	Hours	MD-1 reported total sleep duration
Sleep quality	Recovery	1–5	MD-1 self-reported sleep quality
Muscle soreness	Physical	1–5	MD-1 self-reported muscle soreness
Acute load	Load	AU	7-day rolling average of daily load at MD-1
Chronic load	Load	AU	28-day rolling average of daily load at MD-1
Team performance	Outcome	1–10	Player's appraisal of the team overall
Offensive performance	Outcome	1–10	Player's appraisal of their own offensive performance
Defensive performance	Outcome	1–10	Player's appraisal of their own defensive performance

Note. MD-1 = day before the recorded match date; AU = arbitrary units. Acute and chronic training load are SoccerMon-derived 7-day and 28-day rolling averages of daily session-RPE load. The source narrative describes stress on a 1–5 scale, whereas its parameter table lists 1–10; observed stress values were 2–5. We retained the supplied numeric values and do not infer clinical severity or a favourable/adverse direction. Outcomes are subjective 1–10 ratings, without an objective performance calibration.

Temporal alignment and date variables

Records were aligned by anonymised player identifier and calendar date. The match date was the date attached to the match-performance report; the predictor date was one day earlier. This establishes ordering of recorded dates, not independently verified submission times. MD-1 represents the immediately preceding day and may capture a state close to match participation, but proximity does not establish superior prediction. To evaluate the alternative window, we calculated MD-7 to MD-1 means for each wellness variable and compared them with MD-1 on the same eligible records. We required at least four available daily values for every wellness variable; acute and chronic load retained their existing MD-1 rolling summaries. Complete seven-day wellness coverage was available for 125 of the original 207 records and 72 of the 124 historical-reference records; the paired four-day-minimum comparison included 118 records.

Analytic sample and player-match construction

Daily monitoring records were reorganised into a player-match table linking MD-1 variables to match-day post-match reports by anonymised player identifier and date. Complete records across the nine core MD-1 predictors yielded 215 player-match observations. Players with fewer than three complete match-performance records were excluded in the submitted cohort, leaving 207 observations from 30 elite female footballers. For the revised historical-baseline analysis, a target match additionally required at least three reports on strictly earlier match dates, retaining 124 records from 25 players. Earlier outcome reports could contribute to the baseline even if their MD-1 predictors were incomplete. Complete-case analysis avoided imputing unobserved daily reports but does not remove selection bias when missingness relates to player state or reporting practice (Carpenter & Kenward, 2013; Sterne et al., 2009; White & Carlin, 2010). Reliability depends on within-player variability and the number of observations (Hopkins, 2000). Three prior reports were a pragmatic minimum, not a validated stability criterion; a five-prior-report sensitivity analysis retained 84 targets.

Monitoring variables and predictor configurations

Nine MD-1 predictors were retained in the three theory-informed configurations used in the submitted analysis (Table 2). The psychological-behavioural configuration comprised mood, stress and sleep duration, representing perceived affective state, stress and sleep behaviour in routine athlete monitoring (Beato et al., 2024; Fullagar et al., 2015; Saw et al., 2016). The physical-status configuration comprised acute training load, 28-day chronic training load and muscle soreness within athlete-monitoring frameworks (Bourdon et al., 2017; Halson, 2014; Impellizzeri et al., 2019). The joint configuration included all nine predictors, adding fatigue, readiness to train and sleep quality as recovery indicators (Kellmann et al., 2018; Saw et al., 2016; Beato et al., 2024). These were pragmatic conceptual groupings, not statistically derived factors, validated clinical subscales or mutually exclusive physiological domains. The joint set therefore tested a broader information set, not a psychological–physical interaction. A six-variable union of the first two sets was examined separately as a sensitivity analysis.

Table 2
A priori predictor configurations.

Configuration	Predictors	Theoretical rationale
Psychological-behavioural	Mood, stress, sleep duration	Affective state, perceived stress and sleep behaviour
Physical-status	Acute load, chronic load, muscle soreness	Recent training exposure and musculoskeletal status
Joint	All nine predictors	Adds fatigue, readiness to train and sleep quality to the first two sets

Individual-baseline outcome construction

The three 1–10 post-match ratings were modelled separately because the dataset did not establish that they form a single psychometric construct. Team performance was the reporting player's appraisal of the team overall; offensive and defensive performance concerned that player's own performance in each domain. Team ratings were not aggregated across players. The source did not supply a validated scoring rubric or independent match-event criterion for these ratings. For each player and domain, deviation was the current rating minus the mean of all deduplicated reports on strictly earlier match dates, with at least three earlier reports required. Current and future ratings were excluded from that baseline. Deviations below −0.50 were coded 1 for below-baseline prediction; typical and above-baseline observations were coded 0 (Table 3). The half-point cut-off was operational, not a validated clinical or psychometric threshold. We also predicted above-baseline ratings separately and evaluated three-category classification. Sensitivity analyses at −0.25 and −0.75 examined the below-baseline definition (Supplementary Figure S1).

Model development and validation

The analysis comprised 18 exploratory model specifications: three predictor configurations; three outcomes; and two classifiers. L2-regularised logistic regression was used as a regularised linear reference model. XGBoost was used as a flexible non-linear algorithm. Logistic predictors were standardised within training folds only, and modelling decisions were fixed across folds to reduce overfitting in a small repeated-measures dataset (Christodoulou et al., 2019; Vabalas et al., 2019; Varoquaux, 2018).

Logistic regression used L2 regularisation ($C = 1$), balanced class weights and the liblinear solver. XGBoost used 60 trees, depth 2, learning rate 0.03, subsampling and column sampling fractions of 0.80, minimum child weight 3 and L2 penalty 2; the training-set class ratio determined positive-class weighting. These settings were fixed before the revision fits, without searching the outer test results. The revision plan was recorded after the original analyses were known and was not preregistered.

Model performance was evaluated using repeated five-fold stratified grouped cross-validation with player identity as the grouping variable. Five repeats yielded 25 grouped test folds and ensured that records from the same player did not appear in both training and test partitions. Within each outer training set, three-fold grouped validation selected the probability threshold that maximised F1; ties selected the higher threshold. The model was then refitted to the outer training set and evaluated with that frozen threshold. Precision–recall performance was quantified by non-interpolated average precision (AP), because event prevalence differed across outcomes (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015). ROC-AUC, precision, recall, F1, false-positive rate, false-discovery proportion and Brier score were also reported. Held-out predictions were pooled within each repeat and the five repeat-level metrics averaged. We compared the selected threshold with

Table 3
Historical individual-baseline outcome classification rules.

Deviation from historical mean	Class label	Below binary	Above binary	Three-class code
Deviation < -0.50	Below baseline	1	0	0
$-0.50 \leq$ deviation $\leq +0.50$	Typical baseline	0	0	1
Deviation $> +0.50$	Above baseline	0	1	2

0.50, an all-positive classifier and training-prevalence probability predictions. Leave-one-player-out validation used each held-out player's own training-derived threshold. The procedure is illustrated in Fig. 1a.

Ninety-five per cent confidence intervals used 1,000 player-cluster bootstrap samples of the saved held-out predictions, retaining all records and repeats for each sampled player. These intervals condition on the fitted models; they do not include model-refitting uncertainty or variation across a population of clubs. The joint logistic model was the common reference for the abstract and sensitivity comparisons, while all 18 primary specifications were reported. Additional checks covered the seven-day window, five-report historical minimum, six-variable union, above-baseline and multinomial outcomes, club adjustment, leave-one-club-out validation and training in 2020 with testing in 2021.

Player-level variance decomposition used a one-way random-intercept ANOVA with an effective group-size correction for unequal record counts. Between-player variance was constrained to be non-negative; ICC was its proportion of total variance. Club comparisons used each player's mean as the independent unit, with Welch mean-difference confidence intervals and descriptive two-sided p values. These comparisons describe the two observed squads and do not estimate effects in a population of clubs. No multiple-testing-adjusted confirmatory inference was intended.

Model interpretation

Model interpretation was descriptive rather than mechanistic. Standardised logistic coefficients from full-sample models were examined on the standardised log-odds scale as fitted-model associations (Collins et al., 2024; Steyerberg et al., 2010). Positive coefficients indicated higher predicted odds of below-baseline performance and negative coefficients indicated lower predicted odds. For joint XGBoost models, Tree SHAP values were computed for the outcome defined by the primary −0.50 below-baseline cut-off to describe fitted-model feature attribution (Lundberg & Lee, 2017; Lundberg et al., 2020). Mean absolute SHAP values ranked attribution magnitude across the nine predictors. The native XGBoost contribution calculation used the raw log-odds output; additivity against the model margin was verified. To examine dependence on model fitting, we also summarised joint logistic coefficients from the outer training fits and Tree SHAP values for outer held-out records. These post-hoc summaries were not interpreted as causal effects, stable predictor importance or clinically meaningful mechanisms. Analyses used Python 3.12.14, pandas 3.0.3, scikit-learn 1.9.0, XGBoost 3.3.0, SciPy 1.18.0 and NumPy 2.5.0.

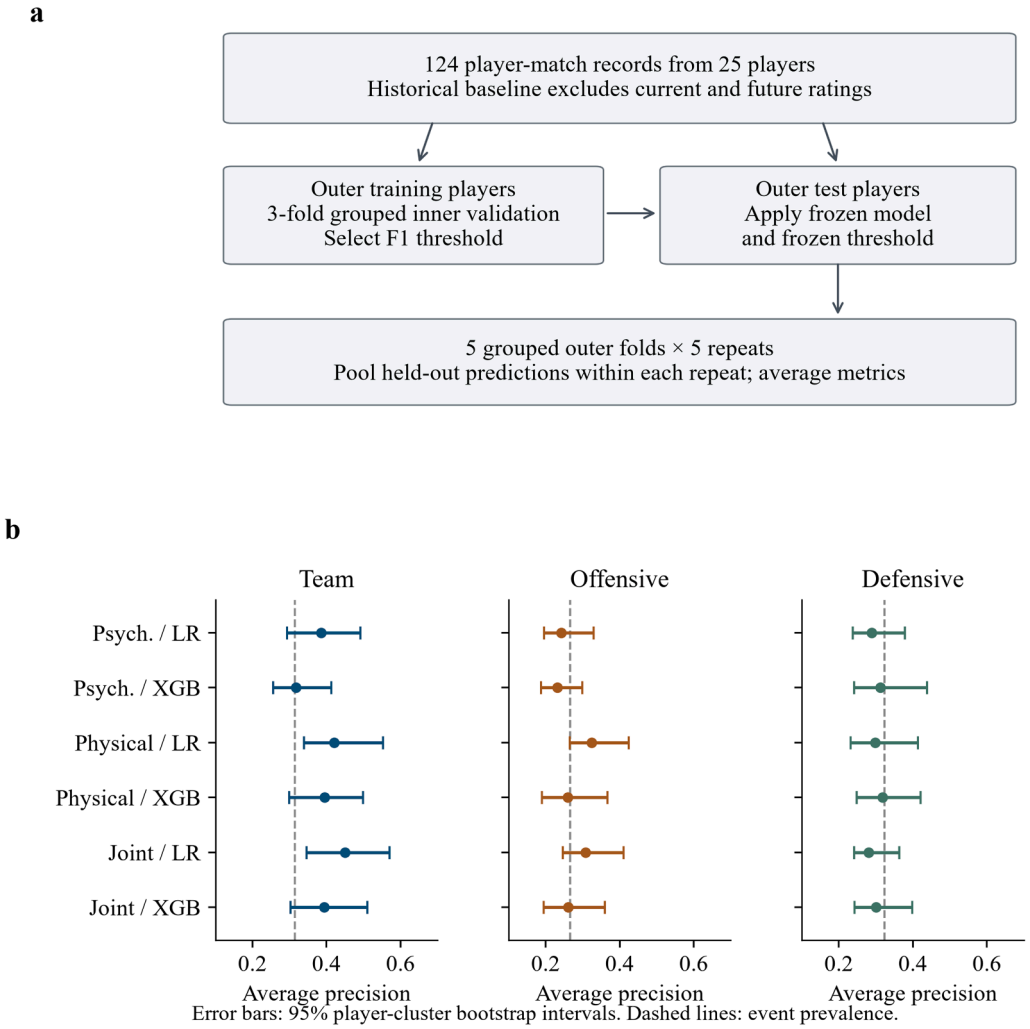


Fig. 1. Player-grouped validation design and primary-model average precision. Panel a shows outer player-grouped validation with inner grouped threshold selection. Panel b reports all six model configurations for each outcome. Psych. = psychological-behavioural; LR = logistic regression; XGB = XGBoost. Error bars are 95% player-cluster bootstrap intervals; dashed lines indicate event prevalence. Estimates average five repeats after pooling held-out predictions within each repeat. The intervals condition on the fitted models and do not represent independent-fold variability.

Results

Sample flow and analytic dataset

The raw match-performance file contained 248 records from 36 players. Removal of one exact duplicate left 247 unique player-date records. Linkage to MD-1 monitoring data and complete-case retention across nine predictors yielded 215 records from 35 players. Excluding players with fewer than three complete reports removed eight records from five players, leaving the submitted cohort of 207 observations. The additional requirement for three strictly earlier reports excluded 83 targets, producing 124 records from 25 players (Team A: 83 records from 17 players; Team B: 41 records from 8 players). Target dates spanned 11 July 2020 to 17 November 2021, with a median of five records per player (range 1–11). A retained player could have only one target record while still having three earlier baseline reports. Supplementary Figure S2 preserves the full sample-construction sequence.

Predictor descriptives and player-level variance structure

Tables 4 and 5 present the player-level variance decomposition and descriptive statistics for the nine MD-1 monitoring variables, respectively. Several psychological and recovery self-reports showed restricted dispersion. Mood (SD = 0.51) and stress (SD = 0.52) ranged from 2 to 5, and fatigue ranged from 2 to 4. Acute load (SD = 103.45; range 0–507.14) and chronic load (SD = 86.72; range 0–481.43) showed greater numerical variability on their different scales. These patterns are compatible with range restriction (Sackett & Yang, 2000), but cross-variable SDs on different units are not directly comparable. Possible explanations include stable states, restricted use of scale anchors, socially desirable responding or other reporting processes. The present data cannot distinguish these explanations, so the variance-restriction account remains a hypothesis.

Player-level variance decomposition showed substantial between-player structure in several variables (Table 4; Fig. 2a). The largest ICCs were observed for chronic load (0.663), mood (0.657), soreness (0.570), readiness (0.567) and fatigue (0.548). Sleep duration reached the zero boundary for the estimated between-player variance. These descriptive results support distinguishing between-player variation from within-player fluctuation (Curran & Bauer, 2011; Hamaker et al., 2015; Wang & Maxwell, 2015). They do not validate the baseline threshold.

Between-team heterogeneity

Using player means rather than treating repeated match records as independent, Team A had higher mean mood and stress than Team B, but the mean-difference intervals included zero (Table 6). The differences were 0.29 (95% CI −0.07 to 0.65) for mood and 0.18 (−0.18 to 0.55) for stress. The raw team-performance difference was −0.58 (−1.69 to 0.53). These descriptive contrasts illustrate the uncertainty around squad-level reporting differences and do not establish equivalent reporting between squads or identify a cause of the differences.

Table 4
ICC and within-/between-player variance decomposition for MD-1 monitoring variables.

Variable	ICC	Between-player variance	Within-player variance	Within-player share
Chronic load	0.663	5152.406	2619.946	33.7%
Mood	0.657	0.178	0.093	34.3%
Muscle soreness	0.570	0.239	0.180	43.0%
Readiness to train	0.567	0.502	0.384	43.3%
Fatigue	0.548	0.178	0.147	45.2%
Stress	0.456	0.127	0.151	54.4%
Acute load	0.442	4836.529	6101.199	55.8%
Sleep quality	0.177	0.082	0.383	82.3%
Sleep duration	0.000	0.000	0.925	100.0%

Note. ICC = between-player variance / (between-player variance + within-player variance). A one-way random-intercept ANOVA used unequal-group-size correction and non-negative between-player variance. Sleep duration reached that boundary. Variances retain the squared units of each predictor; larger ICCs describe clustering, not causal effects.

Table 5
Descriptive statistics for MD-1 predictors (N = 124 records from 25 players).

Variable	Mean	SD	Minimum	Maximum
Mood	3.27	0.51	2.00	5.00
Stress	3.15	0.52	2.00	5.00
Fatigue	2.99	0.56	2.00	4.00
Readiness to train	7.00	0.93	5.00	9.00
Sleep duration (h)	8.27	0.94	6.00	11.00
Sleep quality	3.24	0.68	1.00	5.00
Muscle soreness	2.82	0.64	1.00	4.00
Acute load (AU)	324.37	103.45	0.00	507.14
Chronic load (AU)	315.98	86.72	0.00	481.43

Outcome distribution

Before historical-baseline transformation, raw performance ratings are summarised in Table 7. Below-baseline prevalence was 31.5% for team, 26.6% for offensive and 32.3% for defensive ratings. Table 7 also reports typical and above-baseline counts so that the information combined in the primary binary outcome remains visible.

Predictive performance across outcomes

At the primary −0.50 below-baseline threshold, event prevalence was 31.5% for team performance, 26.6% for offensive performance, and 32.3% for defensive performance. These prevalence values were used as AP reference baselines because precision-recall performance is base-rate dependent under class imbalance (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015).

Discrimination was limited and inconsistent across outcomes. For the joint logistic reference model, AP was 0.450 (95% CI 0.346–0.570) for team, 0.308 (0.245–0.409) for offensive and 0.280 (0.241–0.363) for defensive ratings. Corresponding ROC-AUCs were 0.590, 0.496 and 0.401 (Table 8). The team model contained some ranking information, but offensive and defensive discrimination was weak. At the inner-selected thresholds, false-discovery proportions were 67.8%, 76.1% and 68.4%, and false-positive rates were 82.1%, 79.1% and 96.2%, respectively. F1 values of 0.467, 0.354 and 0.472 did not exceed the corresponding all-positive benchmarks of 0.479, 0.420 and 0.488. False-discovery proportion refers to false alerts among all positive predictions, whereas false-positive rate uses all actual non-events as its denominator. Fig. 1b and Supplementary Table S1 report all 18 primary specifications to avoid presenting selected best models as confirmed winners.

Sensitivity to the below-baseline threshold

Sensitivity analyses examined dependence on the operational below-baseline threshold (Supplementary Figure S1). Event prevalence increased at −0.25 and decreased at −0.75, as expected. The same joint logistic reference model was used at each cut-point rather than reselecting a best model. Results remained outcome-specific and did not establish a generally reliable classifier. On 118 paired records with sufficient seven-day reporting, AP for MD-1 versus seven-day wellness means was 0.427 versus 0.422 for team, 0.294 versus 0.292 for offensive and 0.267 versus 0.301 for defensive ratings. There was no consistent window advantage. Above-baseline and three-category results, alternative history lengths and other sensitivity checks are reported in Supplementary Tables S3–S4. Three-category macro-F1 values were 0.362,

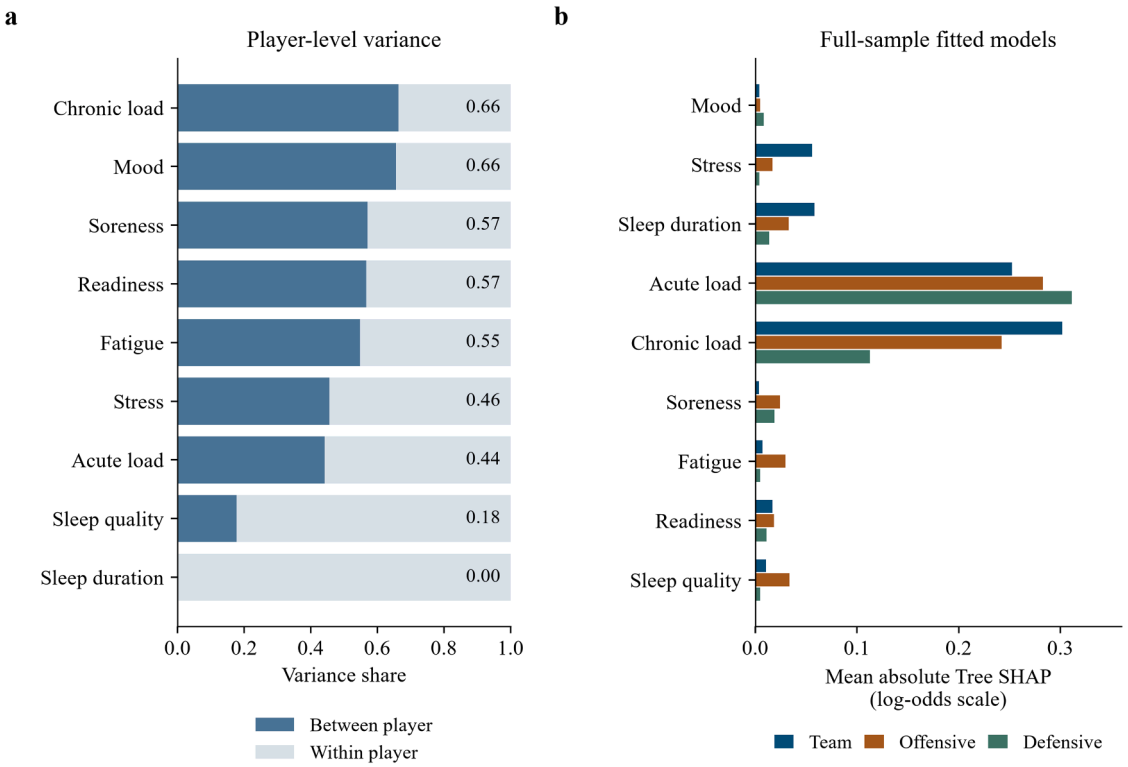


Fig. 2. Player-level clustering and joint XGBoost feature attribution. Panel a shows within- and between-player variance shares, with labels giving ICCs from the unequal-group-size ANOVA decomposition. Panel b shows mean absolute Tree SHAP values on the log-odds scale for joint XGBoost models fitted to the full analytic sample. Neither panel establishes causal predictor importance.

Table 6
Descriptive between-club comparisons using independent player means.

Variable	Team A Mean (SD)	Team B Mean (SD)	Difference (95% CI)	t (df)	p
Mood	3.43 (0.50)	3.14 (0.35)	0.29 (-0.07, 0.65)	1.68 (19.24)	0.110
Stress	3.31 (0.51)	3.12 (0.35)	0.18 (-0.18, 0.55)	1.05 (19.33)	0.308
Raw team performance	6.37 (1.11)	6.95 (1.23)	-0.58 (-1.69, 0.53)	-1.13 (12.58)	0.279

Note. Each player contributes one mean. Team A n = 17 players; Team B n = 8. Difference = A − B. Welch comparisons are descriptive and unadjusted for multiplicity. Shared team and match conditions limit independence and generalisation beyond these squads.

Table 7
Historical-baseline outcome distributions (124 records from 25 players).

Outcome	Below baseline n (%)	Typical baseline n (%)	Above baseline n (%)	Raw mean (SD)
Team performance	39 (31.5%)	36 (29.0%)	49 (39.5%)	6.34 (1.73)
Offensive performance	33 (26.6%)	44 (35.5%)	47 (37.9%)	5.66 (1.42)
Defensive performance	40 (32.3%)	38 (30.6%)	46 (37.1%)	5.70 (1.39)

0.261 and 0.272.

Supplementary leave-one-player-out validation

Leave-one-player-out validation was used as a supplementary check of player-level generalisation. For the joint logistic reference model, pooled AP values were 0.452, 0.309 and 0.280 for team, offensive and defensive ratings, respectively. Performance was summarised using pooled held-out predictions because some players had no below-baseline events; each prediction retained its own training-derived threshold. The 2020-to-2021 temporal check trained on only 20 eligible records and tested on 104, yielding AP values of 0.402, 0.305 and 0.326. This small training sample limits conclusions about temporal transportability. Club-adjusted and leave-one-club-out checks are also reported in Supplementary Table S3.

Logistic regression standardised coefficient patterns

Standardised logistic coefficients were examined as descriptive fitted-model associations. Positive coefficients indicated higher predicted odds of below-baseline performance, whereas negative coefficients indicated lower predicted odds. Because estimates came from fitted prediction models in a small retrospective repeated-measures dataset, they were not interpreted as causal effects, stable risk factors, or monitoring thresholds (Collins et al., 2024; Steyerberg et al., 2010).

Coefficient patterns remained outcome-specific. In the physical-status model for team performance, chronic load had a coefficient of +0.729, whereas acute load and soreness had coefficients of −0.564 and −0.140. In the joint team model, chronic load (+0.710), acute load (−0.513) and stress (−0.445) had the largest absolute coefficients. In the joint offensive model, the corresponding load coefficients were +0.539 and −0.501. In the joint defensive model, mood (+0.286) and readiness (−0.247) had the largest absolute coefficients. Supplementary Table S5 provides all joint coefficients and their ranges across outer training fits. These ranges are descriptive, not confidence intervals; correlated

Table 8
Joint logistic reference-model performance under repeated player-grouped cross-validation.

Outcome	AP (95% CI)	ROC-AUC (95% CI)	Precision (95% CI)	Recall (95% CI)	F1 (95% CI)	Event prevalence
Team performance	0.450 (0.346, 0.570)	0.590 (0.500, 0.684)	0.322 (0.252, 0.396)	0.851 (0.787, 0.924)	0.467 (0.389, 0.537)	0.315
Offensive performance	0.308 (0.245, 0.409)	0.496 (0.390, 0.609)	0.239 (0.177, 0.295)	0.685 (0.496, 0.865)	0.354 (0.271, 0.418)	0.266
Defensive performance	0.280 (0.241, 0.363)	0.401 (0.340, 0.480)	0.316 (0.245, 0.389)	0.935 (0.840, 0.994)	0.472 (0.389, 0.550)	0.323

Note. AP = non-interpolated average precision; ROC-AUC = receiver operating characteristic area under the curve. Values are means across five repeat-level pooled held-out estimates; parentheses give 95% player-cluster bootstrap intervals conditional on fitted predictions. Classification thresholds were selected only within outer training data. All 18 primary model specifications are reported in Supplementary Table S1.

predictors and changes in the training sample limit substantive interpretation of signs and magnitudes.

SHAP characterisation of XGBoost joint model behaviour

Tree SHAP values were computed for the joint XGBoost models predicting the outcome defined by the primary −0.50 below-baseline cut-off. This secondary transparency analysis characterised fitted-model attribution in the full nine-variable non-linear models; it was not evidence that XGBoost was the preferred predictive model (Collins et al., 2024; Lundberg & Lee, 2017). Mean absolute Tree SHAP values are reported in Table 9 and visualised in Fig. 2b.

Larger mean absolute SHAP values indicate greater attribution within the fitted model. The largest attributions were load-related. Chronic load and acute load ranked first and second for team ratings; acute load and chronic load ranked first and second for offensive and defensive ratings (Table 9; Fig. 2b). Outer held-out attribution summaries are provided in Supplementary Table S6. These rankings describe fitted-model behaviour and should not be interpreted as causal mechanisms, stable athlete characteristics or evidence that modifying load would improve performance.

Across the analyses, several MD-1 variables showed substantial between-player variance, whereas discrimination of below-baseline ratings was limited and inconsistent. Load variables contributed prominently to fitted-model explanations, but attribution prominence did not establish useful prediction or causal effects.

Discussion

We examined whether MD-1 monitoring indicators could identify below-baseline self-ratings of overall team performance and of each player's own offensive and defensive performance among elite female footballers competing in Norway. This exploratory prediction analysis defined performance relative to each athlete's strictly historical reference, and validation was grouped by player to respect repeated observations within athletes.

Table 9
Mean absolute Tree SHAP values for all nine MD-1 predictors in joint XGBoost models for the outcome defined by the primary −0.50 below-baseline cut-off.

Predictor	Team performance	Offensive performance	Defensive performance
Mood	0.004	0.005	0.008
Stress	0.056	0.017	0.004
Fatigue	0.007	0.030	0.005
Readiness to train	0.017	0.018	0.011
Sleep duration	0.058	0.033	0.014
Sleep quality	0.011	0.034	0.005
Muscle soreness	0.004	0.024	0.019
Acute load	0.252	0.283	0.311
Chronic load	0.302	0.242	0.113

Note. Values are mean absolute native Tree SHAP contributions on the raw log-odds scale from joint XGBoost models fitted to all 124 eligible records. These are in-sample model descriptions, not held-out performance estimates. Bias contributions are excluded. Outer held-out attribution summaries are provided in Supplementary Table S6.

Principal findings

The principal finding was limited and inconsistent discrimination, with some ranking information for team appraisals but weak offensive and defensive results. The joint logistic model's AP values were 0.450, 0.308 and 0.280, against prevalence values of 0.315, 0.266 and 0.323. The high false-discovery proportions and failure to outperform simple all-positive F1 benchmarks preclude a claim of useful classification. These findings are consistent with the difficulty of developing stable prediction models from small repeated-measures datasets (Collins et al., 2024; Steyerberg et al., 2010; Vabalas et al., 2019). Their value is to establish the limits of the available indicators under an explicitly historical outcome definition, rather than to confirm a deployable monitoring system. The contribution remains exploratory and hypothesis-generating, providing a basis for prospective validation rather than clinical screening, athlete selection or stand-alone decision support.

Individual-baseline monitoring as within-athlete measurement

Person-specific centring addresses the measurement problem that raw self-reports combine within-athlete fluctuation with between-athlete differences in reporting level (Curran & Bauer, 2011; Wang & Maxwell, 2015). Defining performance relative to strictly earlier ratings reduces reliance on a common raw-score threshold and prevents current or later outcomes from entering the reference. This logic is compatible with idiographic approaches, while recognising that group-level associations need not describe individual processes (Fisher et al., 2018; Hamaker et al., 2015; Molenaar, 2004). The ICCs in Table 4 support attention to repeated observations and between-player structure. Grouped validation assesses transfer to held-out players who already have reporting histories, not performance for a new player with no history. The three-report minimum trades data retention against an imprecisely estimated baseline. The five-report sensitivity analysis and limited seven-day coverage make those constraints visible.

Physical-status predictors and performance-related functioning

Acute load, chronic load and soreness are embedded in established athlete-monitoring frameworks and may reflect recent exposure, residual fatigue or musculoskeletal status relevant to performance expression (Bourdon et al., 2017; Halson, 2014; Impellizzeri et al., 2019; Kellmann et al., 2018). In the present analyses, load variables were prominent in the fitted-model coefficients and SHAP summaries. This does not demonstrate that physical-status predictors are generally superior or closer to a causal performance mechanism. Attribution magnitude describes how a model uses a predictor, and the available models still showed limited predictive utility. Elite female footballers face multidimensional physical demands (Datson et al., 2014), and the available SoccerMon variables provide only a partial representation of external load, internal load, neuromuscular readiness, sleep and longitudinal recovery. Prospective studies integrating richer load and recovery measures are needed before physical-status patterns can be linked to

practice.

Psychological-behavioural indicators and restricted variance

The limited discrimination from psychological-behavioural indicators warrants careful interpretation. Restricted dispersion may reflect genuinely stable reported states, restricted use of scale anchors or other reporting processes (Sackett & Yang, 2000). Research on mental-health help-seeking among young elite athletes identifies stigma, confidentiality, limited mental-health literacy, prior help-seeking experiences and relationships with coaches or providers as relevant barriers or facilitators (Gulliver et al., 2012). Whether those processes influenced routine monitoring responses in this cohort is unknown because they were not measured.

Shared response tendencies and unmeasured match experiences could affect both the pre-match reports and post-match appraisals; temporal ordering alone does not eliminate common-method bias (Podsakoff et al., 2003). Subjective measures can be useful for monitoring perceived training responses (Saw et al., 2016), but that evidence does not validate these single-item match ratings against objective performance. Position, playing time, opponent strength and independent event or coach ratings would be needed to test whether the signals extend beyond subjective appraisal.

Reporting context and transferability

Contextual sport psychology emphasises locating research within the relationships and organisations in which practice occurs (Ryba & Wright, 2005; Ryba et al., 2024; Schinke & Stambulova, 2017). For the present study, however, reporting context was not measured. The data therefore cannot determine how players understood the monitoring system or whether staff relationships influenced their responses.

Recent work on indigenised science likewise emphasises explicit reporting of geographic, linguistic and sociopolitical context (Schinke et al., 2026). Here, Norway, Toppserien, the two clubs and the 2020–2021 seasons delimit the source setting and the scope of transferability. They do not establish athletes' nationality, cultural identities, national values or disclosure expectations. Historical baseline adjustment concerns rating level and is not a cultural measure.

Restricted variance may reflect stable reported states, scale use or other reporting practices, but the present data cannot distinguish these accounts. Neither the predictive results nor the two-club contrasts provide evidence for a cultural explanation. Any future study of reporting context should measure the proposed mechanisms directly rather than infer them from country or club membership.

Algorithmic caution and prediction-model boundaries

The modelling results also support algorithmic caution. Only two algorithms were evaluated, and XGBoost did not consistently outperform L2-regularised logistic regression. This is consistent with evidence that flexible machine-learning methods do not necessarily outperform logistic regression in small clinical prediction settings (Christodoulou et al., 2019). Small repeated-measures datasets increase the risk of overfitting, unstable model attribution, and optimistic performance estimates (Collins et al., 2024; Vabalas et al., 2019; Varoquaux, 2018). Post-hoc SHAP values were therefore treated as descriptive model attributions, not as stable explanations. The added outer held-out summaries and coefficient ranges document sensitivity to fitting, but do not validate a mechanism.

Implications for monitoring practice

The immediate practical implication is that the current models should not be used to flag or select athletes. Their discrimination and false-alert burden do not establish a net benefit for training decisions or welfare support. Ethical analysis of athlete biometric monitoring identifies validity, privacy, confidentiality, data security, conflicts of interest and coercion as relevant concerns (Karkazis & Fishman, 2017). Accordingly, any future implementation should specify who can access the data, the purposes for which they may be used, how long they are retained and how athletes can raise concerns. These are governance safeguards, not benefits demonstrated by the present predictions.

Monitoring should be interpreted with attention to the athlete's own account rather than used as stand-alone evidence of mental-health status, motivation, professionalism or readiness to compete. Permission for secondary analysis does not establish consent for consequential deployment (Karkazis & Fishman, 2017), and historical-baseline feedback was not tested as an intervention here. Future studies must evaluate whether such feedback improves dialogue or support and what harms may arise from false alerts.

Limitations

Several limitations constrain interpretation. The study was retrospective and observational, so causal inference is not possible. The sample was small, with 124 player-match records nested within 25 players. Predictive discrimination was limited and inconsistent under both player-grouped cross-validation and leave-one-player-out validation. This limits immediate application, but also underscores the difficulty of predicting within-player match-performance deviations from routine MD-1 monitoring data alone. Performance outcomes were subjective 1–10 post-match ratings rather than objective match-event measures or validated performance constructs. The dataset did not include clinical mental-health outcomes, direct measures of psychological safety, trust, perceived surveillance, disclosure climate, athlete agency, or data-use norms.

Additional limitations concern generalisability and measurement. The analysis was restricted to two Norwegian Toppserien clubs and does not constitute a cross-cultural comparison. Club membership and country of competition cannot establish players' nationality, cultural identity or shared beliefs. The −0.50 threshold was operational and dataset-specific. Complete-case analysis and the historical-report requirement may select players with more consistent reporting. Physical-status indicators may not fully capture external load, internal load, neuromuscular status or longitudinal recovery. Recorded dates did not verify actual submission times; grouped folds were not strictly forward-time evaluations; and bootstrap intervals excluded model-refitting and between-club population uncertainty. Players may also share match or team experiences, a dependence structure not fully represented by player-only resampling. Residual confounding, common-method bias and measurement error remain plausible.

Future directions

Future research would benefit from prospective multi-site designs with external validation across teams, seasons and leagues. Studies that propose reporting-context mechanisms should measure psychological safety, trust, disclosure climate, athlete agency, perceived surveillance and data-use norms directly. Mixed-methods work could clarify how athletes and staff understand monitoring outputs and whether those interpretations affect response variance, validity or predictive information.

Future modelling should also integrate objective match-event data, technical-tactical indicators, GNSS external load, heart-rate variability, actigraphy-derived sleep, and longitudinal fatigue and recovery trajectories. Combined with direct contextual measures, these additions would allow researchers to test rather than assume whether reporting conditions are associated with self-report variance and predictive information. Benefit-harm evaluation should precede any applied deployment, particularly where monitoring outputs could influence selection, training decisions, or access to support.

Conclusions

MD-1 monitoring indicators showed limited and inconsistent discrimination for historically defined below-baseline self-rated performance among elite female footballers competing in Norway. Some ranking information was present for team appraisals, but the false-alert burden did not support applied classification. Historical person-specific references defined deviations without using current or future ratings.

These athlete-monitoring data came from a defined two-club setting, but the study did not measure cultural mechanisms or clinical mental health. The models are not diagnostic, causal, selection-ready or deployable decision tools. Prospective, externally validated research with independent performance criteria and direct measurement of any proposed reporting-context mechanisms is needed.

Data and code availability

SoccerMon is openly available through the repository identified by Midoglu et al. (2024). The derived analytic records, revision analysis scripts, model outputs and figure source data accompany the revision reproducibility materials.

Declaration of generative AI use

During preparation of this revision, the authors used OpenAI Codex to assist with literature searches and reference checks, analysis programming, document consistency checks and English editing. The authors are responsible for reviewing the resulting analyses and text and for the scientific content of the manuscript.

Declaration of competing interest

The authors declare that they have no known financial or non-financial competing interests, personal relationships, professional affiliations, or other circumstances that could have appeared to influence the work reported in this paper.

All authors have reviewed and approved this declaration and confirm that there are no competing interests to disclose.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ijchp.2026.100727.

References

Baumgartner, H., & Steenkamp, J. B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*(2), 143–156. https://doi.org/10.1509/jmkr.38.2.143.18840

Beato, M., Madsen, E. E., Clubb, J., Emmonds, S., & Krustrup, P. (2024). Monitoring readiness to train and perform in female football: Current evidence and recommendations for practitioners. *International Journal of Sports Physiology and Performance, 19*(3), 223–231. https://doi.org/10.1123/ijspp.2023-0405

Bourdon, P. C., Cardinale, M., Murray, A., Gastin, P., Kellmann, M., Varley, M. C., Gabbett, T. J., Coutts, A. J., Burgess, D. J., Gregson, W., & Cable, N. T. (2017). Monitoring athlete training loads: Consensus statement. *International Journal of Sports Physiology and Performance, 12*(Suppl 2). https://doi.org/10.1123/ijspp.2017-0208. S2-161–S2-170.

Bouwmeester, W., Zuithoff, N. P. A., Mallett, S., Geerlings, M. I., Vergouwe, Y., Steyerberg, E. W., Altman, D. G., & Moons, K. G. M. (2012). Reporting and methods in clinical prediction research: A systematic review. *PLOS Medicine, 9*(5), Article e1001221. https://doi.org/10.1371/journal.pmed.1001221

Carpenter, J. R., & Kenward, M. G. (2013). *Multiple imputation and its application*. John Wiley & Sons. https://doi.org/10.1002/9781119942283

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology, 110*, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A. L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., … Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ, 385*, Article e078378. https://doi.org/10.1136/bmj-2023-078378

Costello, J. T., Bieuzen, F., & Bleakley, C. M. (2014). Where are all the female participants in Sports and Exercise Medicine research? *European Journal of Sport Science, 14*(8), 847–851. https://doi.org/10.1080/17461391.2014.911354

Cowley, E. S., Olenick, A. A., McNulty, K. L., & Ross, E. Z. (2021). Invisible sportswomen": The sex data gap in sport and exercise science research. *Women in Sport and Physical Activity Journal, 29*(2), 146–151. https://doi.org/10.1123/wspaj.2021-0028

Curran, P. J., & Bauer, D. J. (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. *Annual Review of Psychology, 62*, 583–619. https://doi.org/10.1146/annurev.psych.093008.100356

Datson, N., Hulton, A., Andersson, H., Lewis, T., Weston, M., Drust, B., & Gregson, W. (2014). Applied physiology of female soccer: An update. *Sports Medicine, 44*(9), 1225–1240. https://doi.org/10.1007/s40279-014-0199-1

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 233–240). https://doi.org/10.1145/1143844.1143874

Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly, 44*(2), 350–383. https://doi.org/10.2307/2666999

Edmondson, A. C., & Lei, Z. (2014). Psychological safety: The history, renaissance, and future of an interpersonal construct. *Annual Review of Organizational Psychology and Organizational Behavior, 1*, 23–43. https://doi.org/10.1146/annurev-orgpsych-031413-091305

Emmonds, S., Heyward, O., & Jones, B. (2019). The challenge of applying and undertaking research in female sport. *Sports Medicine - Open, 5*, 51. https://doi.org/10.1186/s40798-019-0224-x

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences, 115*(27), 6106–6115. https://doi.org/10.1073/pnas.1711978115

Fullagar, H. H. K., Skorski, S., Duffield, R., Hammes, D., Coutts, A. J., & Meyer, T. (2015). Sleep and athletic performance: The effects of sleep loss on exercise performance, and physiological and cognitive responses to exercise. *Sports Medicine, 45*(2), 161–186. https://doi.org/10.1007/s40279-014-0260-0

Gulliver, A., Griffiths, K. M., & Christensen, H. (2012). Barriers and facilitators to mental health help-seeking for young elite athletes: A qualitative study. *BMC Psychiatry, 12*, 157. https://doi.org/10.1186/1471-244x-12-157

Halson, S. L. (2014). Monitoring training load to understand fatigue in athletes. *Sports Medicine, 44*(Suppl 2), 139–147. https://doi.org/10.1007/s40279-014-0253-z

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods, 20*(1), 102–116. https://doi.org/10.1037/a0038889

Hecksteden, A., Pitsch, W., Julian, R., Pfeiffer, M., Kellmann, M., Ferrauti, A., & Meyer, T. (2017). A new method to individualize monitoring of muscle recovery in athletes. *International Journal of Sports Physiology and Performance, 12*(9), 1137–1142. https://doi.org/10.1123/ijspp.2016-0120

Hopkins, W. G. (2000). Measures of reliability in sports medicine and science. *Sports Medicine, 30*(1), 1–15. https://doi.org/10.2165/00007256-200030010-00001

Impellizzeri, F. M., Marcora, S. M., & Coutts, A. J. (2019). Internal and external training load: 15 years on. *International Journal of Sports Physiology and Performance, 14*(2), 270–273. https://doi.org/10.1123/ijspp.2018-0935

Karkazis, K., & Fishman, J. R. (2017). Tracking U.S. professional athletes: The ethics of biometric technologies. *The American Journal of Bioethics, 17*(1), 45–60. https://doi.org/10.1080/15265161.2016.1251633

Kellmann, M., Bertollo, M., Bosquet, L., Brink, M., Coutts, A. J., Duffield, R., Erlacher, D., Halson, S. L., Hecksteden, A., Heidari, J., Kallus, K. W., Meeusen, R., Mujika, I., Robazza, C., Skorski, S., Venter, R., & Beckmann, J. (2018). Recovery and performance in sport: Consensus statement. *International Journal of Sports Physiology and Performance, 13*(2), 240–245. https://doi.org/10.1123/ijspp.2017-0759

Long, H., Li, L., & Zhang, L. (2026). Prioritizing next-day readiness decline in elite female soccer using readiness, workload, and wellness data: A temporally validated and calibrated monitoring framework. *PLOS One, 21*(9), Article e0358036. https://doi.org/10.1371/journal.pone.0358036

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774. https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Luteberget, L. S., Houtmeyers, K. C., Vanrenterghem, J., Jaspers, A., Brink, M. S., & Helsen, W. F. (2021). Load monitoring practice in elite women association football. *Frontiers in Sports and Active Living, 3*, Article 715122. https://doi.org/10.3389/fspor.2021.715122

Midoglu, C., Kjæreng Winther, A., Boeker, M., Dahl Pettersen, S., Pedersen, S., Ragab, N., Kupka, T., Hicks, S. A., Bredsgaard Randers, M., Jain, R., Dagenborg, H. J., Pettersen, S. A., Johansen, D., Riegler, M. A., & Halvorsen, P. (2024). A large-scale multivariate soccer athlete health, performance, and position monitoring dataset. *Scientific Data, 11*, 553. https://doi.org/10.1038/s41597-024-03386-x

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research and Perspectives, 2*(4), 201–218. https://doi.org/10.1207/s15366359mea0204_1

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Reardon, C. L., Hainline, B., Aron, C. M., Baron, D., Baum, A. L., Bindra, A., Budgett, R., Campriani, N., Castaldelli-Maia, J. M., Currie, A., Derevensky, J. L., Glick, I. D., Gorczynski, P., Gouttebarge, V., Grandner, M. A., Han, D. H., McDuff, D., Mountjoy, M., Polat, A., … Engebretsen, L. (2019). Mental health in elite athletes: International Olympic Committee consensus statement. *British Journal of Sports Medicine, 53*(11), 667–699. https://doi.org/10.1136/bjsports-2019-100715

Ryba, T. V., Schinke, R. J., Quartiroli, A., Wong, R., & Gill, D. L. (2024). ISSP position stand on cultural praxis in sport psychology: Reaffirming our commitments to the ethics of difference, cultural inclusion, and social justice. *International Journal of Sport and Exercise Psychology, 22*(3), 533–552. https://doi.org/10.1080/1612197x.2024.2310988

Ryba, T. V., & Wright, H. K. (2005). From mental game to cultural praxis: A cultural studies model's implications for the future of sport psychology. *Quest, 57*(2), 192–212. https://doi.org/10.1080/00336297.2005.10491853

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology, 85*(1), 112–118. https://doi.org/10.1037/0021-9010.85.1.112

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE, 10*(3), Article e0118432. https://doi.org/10.1371/journal.pone.0118432

Saw, A. E., Main, L. C., & Gastin, P. B. (2016). Monitoring the athlete training response: Subjective self-reported measures trump commonly used objective measures: A systematic review. *British Journal of Sports Medicine, 50*(5), 281–291. https://doi.org/10.1136/bjsports-2015-094758

Schinke, R. J., Gao, Q., Li, Y., Raab, M., & Ge, Y. (2026). Indigenized science. *Journal of Sport and Exercise Psychology, 48*(4), 189–197. https://doi.org/10.1123/jsep.2025-0305

Schinke, R. J., & Stambulova, N. (2017). Context-driven sport and exercise psychology practice: Widening our lens beyond the athlete. *Journal of Sport Psychology in Action, 8*(2), 71–75. https://doi.org/10.1080/21520704.2017.1299470

Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ, 338*, b2393. https://doi.org/10.1136/bmj.b2393

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology, 21*(1), 128–138. https://doi.org/10.1097/ede.0b013e3181c30fb2

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLOS ONE, 14*(11), Article e0224365. https://doi.org/10.1371/journal.pone.0224365

Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage, 180*, 68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061

Wang, L. P., & Maxwell, S. E. (2015). On disaggregating between-person and within-person effects with longitudinal data using multilevel models. *Psychological Methods, 20*(1), 63–83. https://doi.org/10.1037/met0000030

White, I. R., & Carlin, J. B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine, 29*(28), 2920–2931. https://doi.org/10.1002/sim.3944

Zentgraf, K., Musculus, L., Reichert, L., Will, L., Roffler, A., Hacker, S., Hilpisch, C., Wiedenbrüg, K., Cermak, N., Lenz, C., de Haan, H., Mutz, M., Wiese, L., Al-Ghezi, A., Raab, M., & Krüger, K. (2024). Advocating individual-based profiles of elite athletes to capture the multifactorial nature of elite sports performance. *Scientific Reports, 14*, Article 26351. https://doi.org/10.1038/s41598-024-76977-8