Diminishing Returns from Added Sensing in Athlete Injury Forecasting: A Two-Cohort Evaluation of Monitoring Tiers with Data-Integrity Audits

Zhuopeng Li[a]

The Experimental High School Attached to Beijing Normal University, Beijing, China

[a]E-mail: stephenlee011111@gmail.com

Abstract

Background. Multimodal "digital athlete" systems assume that adding sensor streams improves injury forecasting, but the marginal value and data integrity of each layer are rarely tested. Objective. To quantify the predictive gain from successive monitoring tiers and audit failure modes that inflate performance. Methods. This retrospective prognostic-modelling study analysed SoccerMon (50 elite women footballers; 25,529 monitored player-days; 63 injury episodes) and a seven-year runner cohort (74 athletes; 42,766 athlete-days; 583 injury days). Five nested tiers (T0–T4) were ordered by instrumentation and reporting burden and compared within cohort. Penalised logistic regression and gradient boosting were assessed with athlete-grouped cross-validation, athlete-level bootstrap intervals, natural-rate calibration, and explicit alert budgets. Pre-modelling audits tested injury-report ascertainment and historical coverage. Results. SoccerMon injury incidence fell 11.9-fold at a September 2020 changepoint despite stable or rising monitoring. In runners, injury rows contained 10.8 recoverable prior days versus 25.8 for non-injury rows; coverage alone achieved AUROC 0.961, and reconstructed-history features yielded AUROC 0.988 versus 0.660 within the released window. Cohort A showed no generalisable forecasting (AUROC 0.32–0.48). In Cohort B, AUROC increased from 0.641 at T0 to 0.663 at T3; the full stack added ΔAUROC 0.043 for logistic regression and 0.019 for gradient boosting over exposure alone. Self-report carried 38.2% of attribution. At a 10% alert budget, the best model detected 19.6% of injuries with 38 alerts per detection. Conclusions. Added sensing showed diminishing returns, hardware-free self-report carried the most signal, and operational utility remained low. Ascertainment-stability and coverage-balance audits should precede performance claims.

Keywords

Injury prediction, athlete monitoring, wearable sensors, data leakage, ascertainment bias, clinical prediction models, value of information.

1. Introduction

Athlete health management in elite sport is still predominantly reactive. In medicine, we intervene after injury rather than preventively; streams that need to support prevention-wearable inertial and positional sensors, physiological & biochemical sensing modalities, periodic imaging protocols, training-load logs, self-reported wellness, etc.-are generally captured in isolation [1,2]. Recently, a substantial engineering literature has attempted to bridge this gap by constructing integrated "digital athlete" or digital-twin ecosystems: multimodal sensor fusion feeding into deep learning and increasingly large multimodal models, yielding patient-specific risk stratification/anticipation of injury along with recommendations for intervention [3–5]. Individual studies have also provided positive results - for example,

predicting injuries from GPS training load in professional football [6], modelling hamstring strain in elite Australian football [7], and utilising public game statistics to predict next-season injury in professional ice hockey [8] - while reviews compile dozens of others [9,10].

That question comes in three parts-and none of them is answered cleanly by the existing literature. The marginal value of each modality is seldom isolated; studies usually report results on one full feature set, and the contribution from expensive layers cannot be separated from cheap ones. Second, reported performance is often overestimated by validation choices - random splits that place the same athlete, or near-duplicate overlapping windows on both sides of a partition - so that the baseline against which any increment is compared is itself unreliable [9,11]. Third, discrimination is often reported without calibration or a decision-analytic context, even though injury is a rare event for which AUROC is a poor marker of clinical usefulness [12–14].

Scepticism about injury prediction is not new. At the individual level, Bahr has argued that screening tests to predict injury are not worth it and likely never will, because the signal-to-base-rate ratio is ominously adverse [15]. At the same time, complex-systems accounts argue that injury results from interacting, time-varying factors rather than from a relatively stable set of measurable predictors (for discussion, see [16,17]). For a decade, the acute:chronic workload ratio-arguably the dominant risk construct in the field [18]-has been shown to rest on fragile conceptual and statistical foundations that do not support its prescriptive use [19,20]. Worse, proposing to spend engineering effort on extra sensing requires evidence that the additional sensing has a positive return.

Answering this question empirically requires public data, and this field is constrained in ways that are often underappreciated. We thus started by surveying the literature (Section 2) and found that truly multimodal athlete monitoring paired with injury labels is rare; that no such dataset seems available for basketball - the sport most often omitted in digital-athlete proposals; and that the two largest public releases each suffer from a data-integrity defect in separate sub-populations, thereby generating large, yet uncorrectable, spectacular results.

We report a deflationary result, and we report it precisely. This is not an argument against building athlete-monitoring systems; it is an argument about where the engineering effort should go, which the closing sections develop.

2. The Public Data Landscape

We catalogued openly available datasets that combine athlete monitoring with injury labels before modelling. This population is small, and its members differ widely in modality coverage, labelling quality, and access.

SoccerMon [21] is the largest publicly available multimodal athlete-monitoring release: 50 elite women footballers at two Norwegian Toppserien clubs across the 2020 and 2021 seasons. The objective component consists of high-frequency positional, accelerometric and heart-rate streams from GPS units (approx. 99 GB); the subjective component, distributed under a CC BY 4.0 license, includes daily self-report wellness (fatigue, mood, readiness, sleep duration, sleep quality, soreness, stress), session-level training load as well as derived load analytics; game performance ratings and athlete-reported injury and illness logs. The present study applied the subjective component; analyses of the objective component were omitted, for reasons stated in Section 5.5.

Competitive-runner training logs [22] comprise a cohort of 74 middle- and long-distance runners from one elite Dutch squad monitored over seven years. Each released row comprises seven consecutive days of 10 training variables (session counts, total and zone-specific distance, sprint distance, strength and alternative training, and self-rated exertion, recovery, and training

success), matched to a binary injury label for the next day. Injuries were defined from staff-recorded time-loss data.

3. Methods

3.1. Design

We conducted a retrospective prognostic-modelling study in two independent public cohorts and followed the TRIPOD+AI recommendations for machine-learning-based prediction models [25]. The target condition was the first occurrence of a time-loss injury; the intended use case is a daily-volume risk readout for training-planning decisions. We performed all analyses in Python 3.12 using scikit-learn, XGBoost [26], and SHAP; all code and derived tables are available as described under Data and Code Availability.

3.2. Cohorts and Eligibility

Cohort A (footballers). We created a player-day panel for all 50 players over our period of interest (2020-01-01 to 2021-12-31), resulting in 731 days and ~36,550 cells. Squad memberships were restricted to the period between the first and last day with recorded activity, resulting in 25,529 monitored player-days. Day-level injury reports were collapsed into episodes, where reports separated by three days or less were considered to be in the same episode; this created 63 episodes across 15 players, with a total of 306 body-site entries (230 graded minor, 76 major), most commonly at right knee (66 entries), left knee (50) and right foot (49). Athlete-days registered during an episode, or in a return-to-play washout of 7 days following one, were excluded from both training and testing since the forecast is meaningless for an athlete who is already injured.

Cohort B (runners). All data were available for the released day-approach file: 42,766 athlete-days from a sample of 74 athletes (1.363% injury days, distributed across 63 athletes; 583/42,766). No rows were excluded.

Cohort characteristics are given in Table 1.

Table 1. Cohort characteristics. Provenance, size, event counts and licensing for both cohorts.

Item	Cohort A (SoccerMon)	Cohort B (runners)
Sport	Women's elite football (Toppserien)	Middle- and long-distance running
Athletes	50	74
Observation span	2020-01-01 to 2021-12-31 (2 seasons)	7 years (day index 0–2673)
Monitored athlete-days	25,529	42,766
Days with athlete-reported wellness	17,008	All rows carry self-report items
Injury events	63 episodes (306 body-site entries)	583 injury days
Athletes with ≥1 event	15	63
Event rate in analysis set	1.82%	1.36%
Analysis rows	4,563 (ascertainment-stable window)	42,766
Label source	Athlete self-report via monitoring app	Coaching/medical staff training log
Licence	CC BY 4.0 (Zenodo 10033832)	Public release accompanying Lövdal et al. (2021)

3.3. Monitoring-Tier Taxonomy

One asymmetry requires explicit statement. In Cohort A, T4 uses the load analytics produced and made available by the data providers from the continuous club record. T4 is bounded to within-week quantities in Cohort B. Foster's monotony and strain are defined on a single training week, and therefore they are computable from the published 7-day window [27]; the acute:chronic workload ratio is not, because it requires a 28-day chronic load which is not made available in the release. What happens if this restriction is violated is depicted in Section 4.2.

Table 2. Monitoring-tier taxonomy. Variables assigned to each tier in each cohort, with the instrumentation or reporting burden each tier implies (Section 3.3).

Tier	Cohort A (footballers)	Cohort B (runners)	Instrumentation / reporting burden
T0 exposure only	athlete tenure, day of week	athlete tenure, day of week	None
T1 passive wearable	sleep duration, sleep quality	number of sessions, total km, hours of alternative training	Commodity wrist-worn device; no athlete action required
T2 (see burden column)	fatigue, soreness, mood, stress, readiness	km in zones 3–4, km in zones 5/T1/T2, km sprinting, strength training	Cohort A: 30 s daily questionnaire. Cohort B: calibrated heart-rate or pace zones
T3 (see burden column)	number of sessions, session minutes, mean RPE, daily load, weekly load	perceived exertion, perceived recovery, perceived training success	Cohort A: per-session logging by athlete or staff. Cohort B: 30-second daily questionnaire
T4 load dynamics	ACWR, ATL, CTL (28 d), CTL (42 d), monotony, strain	within-week load dynamics block	Sustained monitoring history plus an analytics stack

3.4. Data-Integrity Screens

Two screens were specified and run before any tier comparison.

Ascertainment stability (Cohort A). Because athletes report injuries through the same application that collects wellness information, incidence (reported incidence hereafter) is confounded with reporting behaviour. We calculated monthly injury-onset counts, using monitored player-days as the exposure offset. We fitted a single-changepoint Poisson model, with the changepoint selected by maximum likelihood, and report the likelihood-ratio statistic against an equivalent no-changepoint model. So, modelling was limited to the period before the changepoint. We computed wellness-reporting compliance and player-days monitored in the same month as controls; a true reduction in injury incidence should not occur with stable monitoring effort.

Coverage balance (Cohort B). Each released row overlaps six out of every seven days, thus providing a perfect inversion of the underlying daily series in these overlapped windows. This enables chronic-load features. So, we measured how much of the past 28 days each released row can recover and calculated this for both injury and non-injury rows. However, where coverage is imbalanced, any feature spanning more than the release window reflects data availability rather than training behaviour.

3.5. Feature Construction and Labels

Cohort A. Descriptors of past-only data per athlete-day for each monitoring stream: last observed value (carried forward for at most three days), 3- and 7-day means, 7-day standard

deviation, difference between the means over the previous 7 days and 28 days, a within-athlete z-score against the baseline of their mean over the previous 28 days, comparison to baseline on the previous day, day since last observation. All rolling statistics were calculated within each athlete and did not use any information after the index date. This label marked the start of an injury episode occurring within a week (horizons of 3 and 14 days were also studied in sensitivity analysis), and we kept rows only when a wellness report was performed within the last seven days so features weren't completely stale.

Cohort B. Computed per variable within each of the seven-day windows released: last day, mean, SD, sum, maximum value over the week (max), linear trend (slope/unit), ratio of 7th day to window mean (ratio), and count of days equal to zero during the week– plus a block for within-week load-dynamic. The label is the release - injury on the following day of the window - which makes results consistent with previous work using this release. We used the reconstructed-history features solely to demonstrate leakage in Section 4.2, not for any other tier analysis.

3.6. Models and Validation

We built two model families: penalised logistic regression (L2, with median imputation, missingness indicators and standardisation), and gradient-boosted trees [26]. Both studies used class weighting inversely proportional to class frequency for discrimination analyses. No tuning information leaked into evaluation because the regularisation strength of an inner athlete-grouped loop was nested within each training fold for the logistic model.

And all validation was stratified by athlete using five-fold GroupKFold: no athlete shared rows between the training and test side of a given split. This matters in both cohorts—in Cohort B as well, since six of seven days are shared between any two consecutive rows; an arbitrary row split places near-duplicates on either side. Section 4.6 quantifies that.

3.7. Performance, Uncertainty and Decision Utility

We report areas under the receiver operating characteristic curve (AUROC), area under the precision–recall curve (AUPRC), ratio over prevalence, the Brier score, and calibration intercept and slope from logistic recalibration of predicted log-odds [13,28]. Reporting AUPRC alongside AUROC is important because event rates here are 1–2%, so AUROC is optimistic [12].

Confidence intervals are athlete-level cluster bootstraps (percentile method) resampling athletes with replacement, since athlete-days within an athlete are not independent. We use the paired bootstrap (n=7500) on the same resampled athletes for comparisons between adjacent tiers, reporting the associated bootstrap probability that the increment exceeds zero.

For deployment analyses, we refitted unweighted models at the natural event rate and used only those predictions for calibration, net benefit, detection rates and alert burden. Operational utility was expressed as an alert budget: the fraction of athlete-days staff could act on, the proportion of injuries detected, and alerts generated per injury detected.

4. Results

4.1. Injury Ascertainment Collapses While Monitoring Continues (Cohort A)

In SoccerMon, the monthly incidence of injury onset is not stationary (Figure 1). The single changepoint with maximum likelihood occurs in September 2020: incidence drops from 8.00 to 0.67 onsets per 1000 monitored player-days-an 11.9-fold reduction-and has a likelihood-ratio statistic of 83.9 against the no-changepoint model.

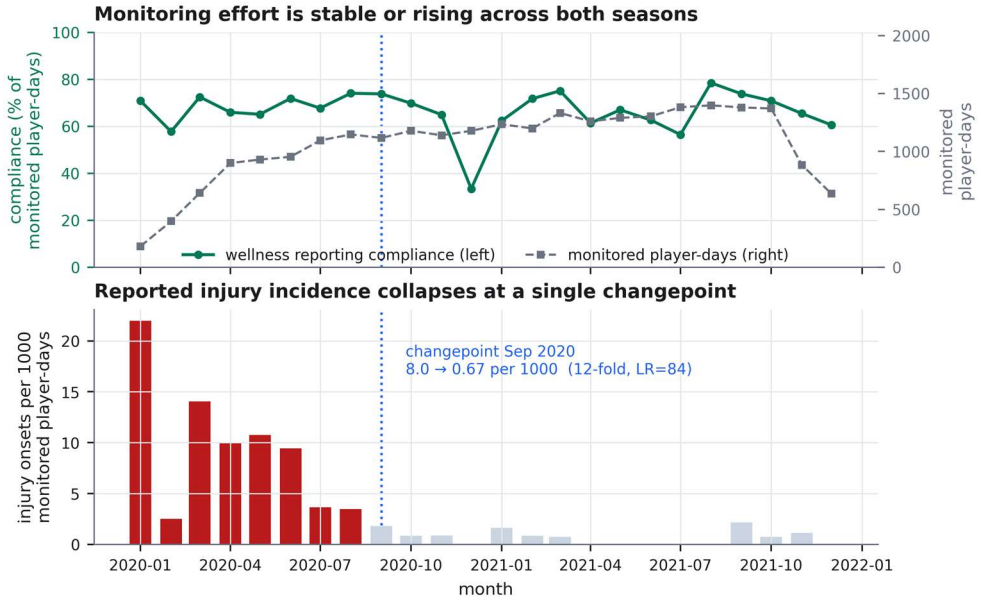


Figure 1. Injury ascertainment collapses while monitoring continues (Cohort A). Upper panel: monthly wellness-reporting compliance (left axis) and monitored player-days (right axis) are stable or rising across both seasons. Lower panel: injury onsets per 1000 monitored player-days, with months before the maximum-likelihood Poisson changepoint (September 2020) highlighted. Reported incidence falls 11.9-fold across the changepoint (likelihood-ratio statistic 83.9) with no corresponding fall in monitoring effort.

The control series excludes a real epidemiological decrease. Wellness-reporting compliance throughout the window fluctuates between 0.56 and 0.79 of monitored player-days, with no discernible downward trend (green bars), while routinely monitored player-days rise from January 2020 (182) to plateau near 1,300 per month across 2021. Of 63 injury incidents, 54 (86%) occurred before the changepoint, while 57 (90%) came from just one team. The 2021 season provides more monitoring data than the 2020 season and nine episodes in total; across some ~1300 monitored player-days per month from April 2021, the release logs no instances of injury onset for eight consecutive months. Therefore, a true injury incidence rate of zero across an entire elite football squad for two-thirds of a competitive season seems implausible, given prospective surveillance in professional football reports injury incidences several orders of magnitude greater than zero across full seasons [29]; athlete-reported injury ascertainment ceased, but wellness reporting continued.

We therefore restricted Cohort A modelling to 2020-01-01 to 2020-08-31. The analysis set includes 4,563 athlete-days across 36 athletes after applying the in-episode and washout exclusions, combined with a recency requirement; this includes 83 event-days (1.82%) contributed by 11 athletes.

4.2. Recoverable History Is Imbalanced by Outcome, and Reconstructing It Inflates Performance by a Factor of 25 (Cohort B)

Inverting the overlapping windows fully recovers the underlying daily series: 299,362 windowed observations collapse into 49,162 unique athlete-days, with a maximum within-day disagreement of 0.0 across each of the ten variables, so this inversion is lossless.

However, as shown in Figure 2a, the recovered history remains extremely outcome-biased. For non-injury rows, we recover on average 25.8 days (median 28) of the prior 28 days, while for injury rows only an average of 10.8 days (median 13). The injury bands are centred on roughly 7 and 13 recoverable days, the signature of windows released in isolation rather than during a

continuous training cycle. All that separates injury rows from non-injury rows - carrying no physiological information of any kind other than coverage, the count of available days - with an AUROC value of 0.961.

This modelling result is dramatic (Figure 2b). With gradient boosting (0.978 and 0.333 with logistic regression), a pipeline equipped with features computed through the reconstructed 28-day history, under athlete-grouped cross-validation, yields AUROC 0.988 and AUPRC 0.594. The same pipeline, using models and splits limited to the available 7-day window, yields AUROC 0.660 and AUPRC 0.024. The observed AUPRC is inflated by about a factor of 25. The most strongly associated feature is the last observed acute:chronic workload ratio, with a univariate AUROC of 0.910; a number one should be read as an analytic fire alarm rather than a finding, since no true load ratio segregates a 1.4%-prevalence outcome so well.

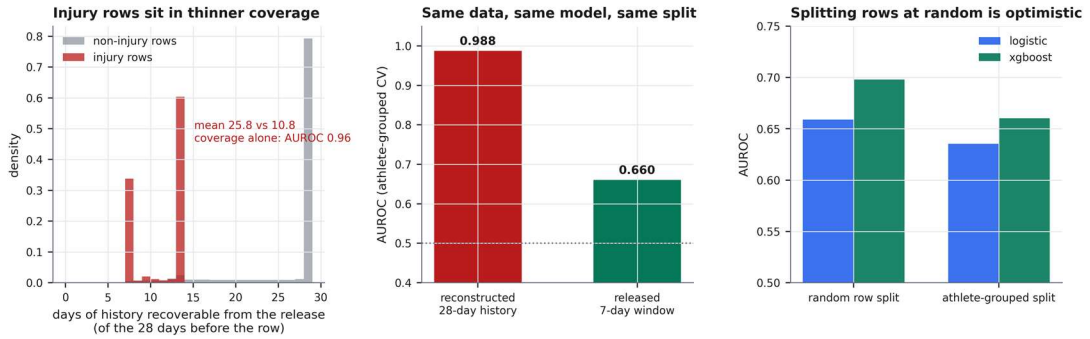


Figure 2. Two leakage mechanisms, quantified. (a) Days of athlete history recoverable from the runners release for the 28 days preceding each row, by outcome: injury rows sit in systematically thinner coverage (mean 10.8 vs 25.8 days), and coverage alone separates the classes with AUROC 0.961. (b) The same pipeline, models and athlete-grouped splits, with features computed across the reconstructed history versus confined to the released seven-day window. (c) Splitting rows at random rather than by athlete, with identical features and models.

All tier results below therefore use window-confined features only.

4.3. The Full Monitoring Stack Adds Little over Exposure Alone

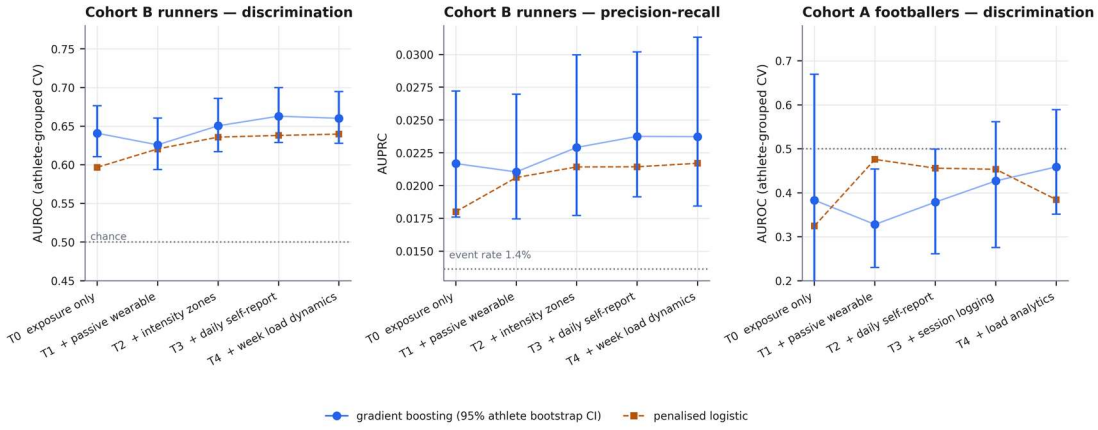


Figure 3. Value of each monitoring tier. Out-of-fold performance under athlete-grouped five-fold cross-validation for nested monitoring tiers. Points are gradient boosting with 95% athlete-level cluster-bootstrap intervals; dashed lines are penalised logistic regression. Left and centre, Cohort B (runners), AUROC and AUPRC against the 1.36% event rate. Right, Cohort A (footballers) within the ascertainment-stable window, where every interval spans chance.

Table 3. Tier performance. AUROC, AUPRC, AUPRC lift over prevalence, Brier score and sensitivity (Sens.) at a 10% alert budget for every tier, model and cohort, with the number of features (Feat.) and 95% athlete-level cluster-bootstrap intervals.

Cohort	Tier	Model	Feat.	AUROC (95% CI)	AUPRC (95% CI)	Lift	Brier	Sens.
B (runners)	T0 exposure only	Logistic regression	2	0.597 (0.562–0.636)	0.018 (0.014–0.023)	1.32	0.244	0.144
B (runners)	T0 exposure only	Gradient boosting	2	0.641 (0.610–0.676)	0.022 (0.018–0.027)	1.59	0.226	0.190
B (runners)	T1 passive wearable	Logistic regression	26	0.621 (0.586–0.656)	0.021 (0.016–0.028)	1.51	0.238	0.172
B (runners)	T1 passive wearable	Gradient boosting	26	0.626 (0.594–0.660)	0.021 (0.017–0.027)	1.54	0.168	0.173
B (runners)	T2 intensity/zone sensing	Logistic regression	58	0.636 (0.604–0.671)	0.021 (0.017–0.029)	1.57	0.232	0.187
B (runners)	T2 intensity/zone sensing	Gradient boosting	58	0.650 (0.617–0.686)	0.023 (0.018–0.030)	1.68	0.149	0.194
B (runners)	T3 self-report	Logistic regression	82	0.638 (0.606–0.675)	0.021 (0.016–0.030)	1.57	0.231	0.185
B (runners)	T3 self-report	Gradient boosting	82	0.663 (0.629–0.700)	0.024 (0.019–0.030)	1.74	0.134	0.220
B (runners)	T4 within-week dynamics	Logistic regression	87	0.640 (0.606–0.676)	0.022 (0.017–0.030)	1.59	0.230	0.196
B (runners)	T4 within-week dynamics	Gradient boosting	87	0.660 (0.628–0.695)	0.024 (0.018–0.031)	1.74	0.134	0.220
A (footballers)	T0 exposure only	Logistic regression	2	0.324 (0.150–0.606)	0.039 (0.007–0.116)	2.13	0.247	0.108
A (footballers)	T0 exposure only	Gradient boosting	2	0.383 (0.198–0.670)	0.041 (0.007–0.180)	2.27	0.158	0.205
A (footballers)	T1 passive wearable	Logistic regression	18	0.476 (0.351–0.621)	0.017 (0.007–0.035)	0.93	0.241	0.084
A (footballers)	T1 passive wearable	Gradient boosting	18	0.328 (0.230–0.454)	0.012 (0.005–0.023)	0.68	0.042	0.012
A (footballers)	T2 self-report wellness	Logistic regression	58	0.456 (0.331–0.603)	0.018 (0.007–0.041)	1.00	0.167	0.096
A (footballers)	T2 self-report wellness	Gradient boosting	58	0.379 (0.261–0.500)	0.013 (0.005–0.026)	0.73	0.028	0.000
A (footballers)	T3 session logging	Logistic regression	98	0.453 (0.279–0.621)	0.018 (0.008–0.038)	1.01	0.144	0.133
A (footballers)	T3 session logging	Gradient boosting	98	0.427 (0.276–0.561)	0.017 (0.006–0.045)	0.96	0.022	0.060
A (footballers)	T4 load analytics	Logistic regression	146	0.384 (0.247–0.609)	0.014 (0.007–0.025)	0.78	0.115	0.036
A (footballers)	T4 load analytics	Gradient boosting	146	0.459 (0.351–0.589)	0.016 (0.007–0.029)	0.88	0.022	0.084

Results for Cohort B are shown in Figure 3 and Table 3. Gradient-boosting AUROC was 0.641 (95% CI 0.610–0.676) at T0, 0.626 at T1, 0.650 at T2, 0.663 at T3 and 0.660 at T4; corresponding logistic-regression values were 0.597, 0.621, 0.636, 0.638 and 0.640. AUPRC remained 0.018–0.024 against a prevalence of 0.0136 (lift 1.3–1.7).

Incremental tier increases are marginal, indeed nearly indistinguishable from zero. The largest individual AUROC gain is T1→T2 using gradient boosting (+0.025, 95% CI 0.010–0.039, bootstrap P(Δ>0) = 0.998). From the self-report step T2→T3, we provide +0.012 (−0.002 to 0.027, P = 0.950). From the load-analytics step, T3→T4, we get −0.003 (−0.010 to 0.005, P = 0.253) - that is, once we have our underlying streams detected, the analytics tier adds no additional detectable gain.

Taken end to end, the entire monitoring stack is worth ΔAUROC 0.043 (95% CI 0.014–0.068, P(Δ>0) = 1.000) over exposure alone for logistic regression and 0.019 (−0.006 to 0.044, P = 0.945) for gradient boosting, with ΔAUPRC of 0.0037 (0.0004–0.0079) and 0.0021 (−0.0005 to 0.0055) respectively on a 1.4% base rate.

4.4. Cohort A Supports No Generalisable Forecasting at Any Tier

No tier in the Cohort A group was discriminated above chance within the ascertainment-stable window. AUROC ranged from 0.324 to 0.476 (logistic) and 0.328 to 0.459 (gradient boosting); AUPRC lift ranged from 0.68 to 2.27; each cluster-bootstrap interval spanned 0.5, and most point estimates fell below it. No adjacent-tier increment was significantly greater than zero.

This null is robust. AUROC ranged from 0.282 to 0.578 across the full cohort, with a single squad contributing the most events for prediction horizons of 3, 7, and 14 days (Supplementary Table S1). The configuration with the highest performance for this target - a 14-day horizon only to the squad and involving 107 event-days - attained an AUROC of 0.578, well within an interval crossing random inference.

We take this to be a power and label-quality result, not evidence that football injuries are unpredictable per se. All events are attributable to 11 individual athletes in the exploration window; under athlete-grouped validation, a model must generalise across individuals trained on several injured exemplars, and with only 11 outcome-contributing clusters this is far below what is required for stable prediction-model development [30,31]. This null is important to report because the same dataset, with a random or temporal split across both seasons, will yield much nicer-looking numbers, as described in Section 4.1.

4.5. The Cheapest Layer Carries the Most Signal

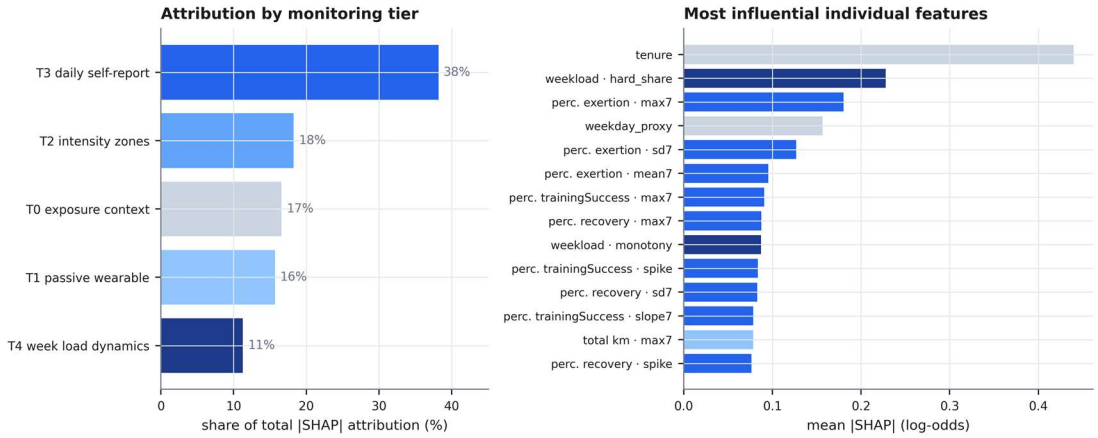


Figure 4. Where the signal sits. Mean absolute SHAP attribution for the full-tier Cohort B gradient-boosting model, aggregated by monitoring tier (left) and for the most influential individual features (right). The daily self-report tier, which requires no sensor hardware, carries 38.2% of total attribution.

Attribution analysis on the full-tier Cohort B model (Figure 4) places 38.2% of total mean absolute SHAP attribution on the daily self-report tier - the layer requiring no hardware at all. Intensity and zone sensing accounts for 18.3%, exposure context for 16.6%, passive wearable

variables for 15.7%, and derived within-week load dynamics for 11.3%. The most influential individual features are summaries of perceived exertion and perceived recovery, consistent with univariate discrimination (perceived exertion, AUROC 0.630; perceived recovery, 0.604) and with the long-standing finding that subjective athlete-reported measures respond to training stress at least as sensitively as objective ones [32].

4.6. Calibration Is Acceptable; Operational Utility Is Not

Unweighted models fitted at the natural event rate were approximately calibrated in the large but produced overly variable predictions (mean predicted risk 0.0153; observed prevalence 0.0136; calibration intercept −0.117 and slope 0.573 for gradient boosting (for logistic regression, 0.0149, −0.095 and 0.477); Figure 5a).

The decision-analytic readout is adverse (Figure 5b,c). The best model relays 10.6% of the injuries at an alert budget of 5% athlete-days, producing 34 alerts per injury detected. At alert budgets of 10% and 20%, the model detected 19.6% and 36.7% of injuries, requiring 38 and 40 alerts per detected injury, respectively. While detection outperforms random flaggers at all budgets-the model is far from useless-a system that requires practitioners to act on about forty alerts to predict one injury, while missing four out of five injuries per budget, cannot logically serve as a reference for automatic intervention.

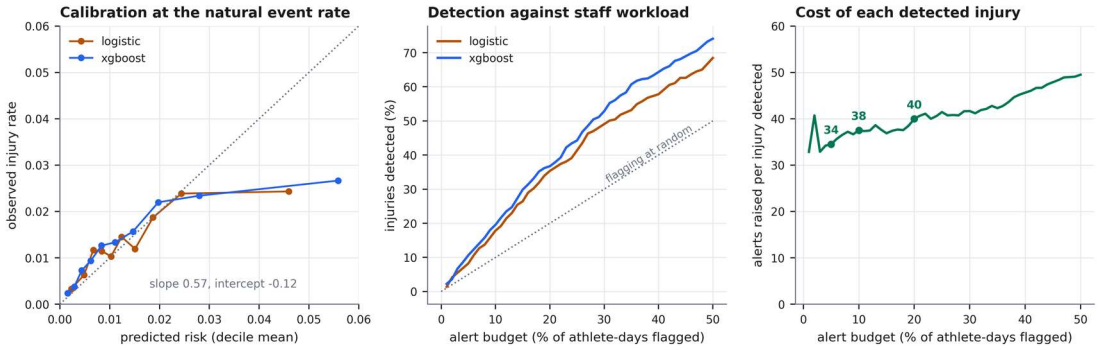


Figure 5. Calibration and operational utility (Cohort B). (a) Calibration of unweighted models fitted at the natural event rate, by decile of predicted risk. (b) Injuries detected as a function of the alert budget, against random flagging. (c) Alerts raised per injury detected, annotated at 5%, 10% and 20% budgets.

Lastly, how you design your validation matters measurably. Splitting rows at random rather than by athlete resulted in an AUROC that was 0.038 higher (P=0.047) and an AUPRC that was 0.0075 higher (P=0.046), a relative inflation of AUPRC of 29% (Figure 2c), given identical features and models. This is accordingly a smaller effect than that of the coverage artefact of Section 4.2. Still, it is systematic and arises here in part because consecutive rows share six days out of seven.

5. Discussion

5.1. Principal Findings

In this study of 68,295 athlete-days from two independent public cohorts, we find that (1) the marginal predictive value of successive layers of athlete-monitoring is small and diminishing; (2) the layer carrying most signal requires no sensor hardware; (3) the derived load-analytics tier - a construct upon which much applied sports science rests - adds nothing detectable once its input streams become available; and (4) while our risk scores are better than chance and approximately calibrated in the large, they make for weak operational tokens. In addition, our

research reveals that the two largest public datasets in this field have a different integrity defect each that can lead to major over-inflation of results, and we quantify both.

5.2. Implications for Digital-Athlete and Digital-Twin Programmes

These results directly shape how an integrated sensor–AI athlete ecosystem should be designed.

Injury prediction alone does not justify sensor investment. This gap is worth ~2.5 AUROC points here, going from a passive consumer wearable with calibrated intensity sensing and then nothing more, to building a complete derived-analytics stack. If the rationale for an expensive instrumentation tier is injury forecasting, public evidence does not currently support that case. Other functions - performance profiling, rehabilitation monitoring, load prescription, biomechanical assessment - may still justify instrumentation on their own terms.

The digital-twin simulation inherits the accuracy of your risk model. Simulations of stress scenarios and testing counterfactual rehabilitation or load-modification strategies are enticing proposals, yet a simulator's counterfactual outputs cannot be more trustworthy than the risk model beneath; indeed, a model at AUROC 0.66 with a calibration slope of 0.57 does not warrant confident individual-level counterfactuals. In addition, counterfactual load modification is a causal statement, and even the observational designs described here-nor those in available datasets-identify causal effects [20,33]. For this kind of work-i.e., the digital twin in this domain-scenario exploration with explicit assumptions should be described and validated, not predictions of intervention effects.

5.3. Basketball Specifically

The literature on proposals often focuses on professional basketball, where the clinical problem is especially acute because of dense schedules, travel, and high-intensity loading [15]; public data sources also document schedule–injury associations [23,24]. According to our survey, no publicly available dataset matches multimodal monitoring with injury labels for basketball. This means a basketball digital-athlete programme has two honest paths: either importing evidence from other sports and declaring the transfer assumption, or performing and-ideally-publishing the missing data. The second appears to add more value, with its worth lying less in additional sensor channels than in label quality: dense, staff-adjudicated, consistently verified injury records with documented definitions [34].

5.4. Relation to Prior Work

The reported discrimination we report for Cohort B (AUROC 0.66) is within the range of previously reported values for this release and similar to prior machine-learning injury-prediction studies more broadly [6,9,10,22]. At the same time, our finding that self-report disproportionately carries weight is consistent with the findings of Saw and colleagues [32]. The original finding that the derived load-analytics tier added no detectable information improves on systematic reviews that have reported largely inconsistent direction and effect for the training-load–injury relationship across studies [35], and it addresses key methodological critiques of ratio-based workload constructs [19,20]. Our null for Cohort A opposes previously reported increased performance on SoccerMon, and Section 4.1 provides a simple mechanism by which such performance can emerge without signalling injury; we encourage re-examining results around the ascertainment changepoint. More broadly, the study gives precision to long-standing conceptual critiques [15,16,19] for open data.

To the best of our knowledge, neither of these two integrity screens is standard practice. Both are cheap. The ascertainment test requires only event counts, an exposure offset and a changepoint fit, while the coverage audit requires only available history count per row, stratified by outcome. Indeed, both showed positive results on the first datasets we tested them against, suggesting that defects are common and implementing these techniques would be worthwhile.

5.5. Limitations

First, the objective SoccerMon GPS stream (approximately 99 GB) could provide an external-load tier above T4, but archive download rates made acquisition impracticable for this study. The evaluated ceiling is therefore analytics-derived rather than elite external-load sensing.

Second, with ascertainment restriction, Cohort A is massively underpowered (83 event-days from 11 athletes), making its null weak evidence of absence rather than absence.

Third, both cohorts apply heterogeneous, unvalidated time-loss injury labels (self-reported in Cohort A; organisational logs in Cohort B) without diagnosis, increasing exposure misclassification risk in unpredictable ways.

Fourth, tier taxonomy annotates variables on the cheapest device class that could realistically produce them. Cohort A scores related to things like sleep quality, for example, are self-rated and treated as if they were what a consumer sleep tracker outputs automatically; that mapping is a modelling assumption, not measurement equivalence.

Fifth, while runners are partially normalised and single-squad, Cohort A is sampled from the disrupted 2020 season schedule due to COVID-19 disruption; neither cohort can be viewed as a random sample of any wider population.

Sixth, we compared two families of models and did not test deeper architectures; we applied the standard framework for model evaluation (discrimination, calibration, and clinical usefulness) [28,36]. We expect the architecture choice not to change the conclusions, as we found strong evidence of leakage in more expressive modelling of artefactual features. Thus, the signal ceiling here is low-but we have not demonstrated this.

5.6. Recommendations

For researchers using public athlete-monitoring data, we suggest three checks before any performance claim: (i) plot event incidence against monitored exposure over time and test for a changepoint; (ii) tabulate, by outcome class, how much history each row's features actually span; and (iii) validate with athlete-grouped splits and report AUPRC, calibration and an alert-budget readout alongside AUROC. For dataset publishers, releasing an explicit per-row coverage indicator and documenting periods of known reporting change would prevent both defects found here.

6. Conclusion

Using data from the two largest publicly available multimodal athlete-tracking datasets, we show that each successive sensing layer yields one to two AUROC points, with diminishing returns as the stack grows: a full monitoring stack adds about two to four AUROC points over knowing only an athlete's exposure. Hardware-free, a daily questionnaire delivers the layer with greatest attribution. Under realistic alert budgets, the best model identifies one in five injuries for 40 alerts per detection. We found an 11.9-fold ascertainment collapse in one dataset and a coverage imbalance (exacerbated by our outcome-driven modelling approach) that artificially inflated apparent precision-recall performance by ∼25-fold versus two reproducible integrity screens applied pre-modelling to assess reporting quality.

The key takeaway for the many digital-athlete programmes out there is that it is not about building less, but building differently; the binding constraint in this space today is label quality, density, and ascertainment-not model input.

Data and Code Availability

Both datasets are public. SoccerMon is available at https://doi.org/10.5281/zenodo.10033832 under CC BY 4.0 [21]; the competitive-runner day-approach release accompanies Lövdal et

al. [22]. No data are redistributed with this work. All analysis code, derived tables and figure scripts are provided in the accompanying repository, including the two integrity screens as standalone modules (labels.poisson_changepoint and leakage_case.coverage_diagnostic).

Author Contributions

Zhuopeng Li: Conceptualization, Methodology, Software, Formal analysis, Data curation, Visualization, Writing – original draft, Writing – review & editing.

Conflicts of Interest

The author declares no conflicts of interest.

Acknowledgments

Funding: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

References

[1] Bourdon, P. C., Cardinale, M., Murray, A., Gastin, P., Kellmann, M., Varley, M. C., et al. (2017). Monitoring athlete training loads: Consensus statement. International Journal of Sports Physiology and Performance, 12, S2-161–S2-170. https://doi.org/10.1123/ijspp.2017-0208

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

[3] Seshadri, D. R., Li, R. T., Voos, J. E., Rowbottom, J. R., Alfes, C. M., Zorman, C. A., et al. (2019). Wearable sensors for monitoring the physiological and biochemical profile of the athlete. npj Digital Medicine, 2. https://doi.org/10.1038/s41746-019-0150-9

[4] Corral-Acero, J., Margara, F., Marciniak, M., Rodero, C., Loncaric, F., Feng, Y., et al. (2020). The 'digital twin' to enable the vision of precision cardiology. European Heart Journal, 41, 4556–4564. https://doi.org/10.1093/eurheartj/ehaa159

[5] Zadeh, A., Taylor, D., Bertsos, M., Tillman, T., Nosoudi, N., & Bruce, S. (2021). Predicting sports injuries with wearable technology and data analysis. Information Systems Frontiers, 23, 1023–1037. https://doi.org/10.1007/s10796-020-10018-3

[6] Rossi, A., Pappalardo, L., Cintia, P., Iaia, F. M., Fernàndez, J., & Medina, D. (2018). Effective injury forecasting in soccer with GPS training data and machine learning. PLOS ONE, 13, e0201264. https://doi.org/10.1371/journal.pone.0201264

[7] Ruddy, J. D., Shield, A. J., Maniar, N., Williams, M. D., Duhig, S., Timmins, R. G., et al. (2018). Predictive modeling of hamstring strain injuries in elite Australian footballers. Medicine & Science in Sports & Exercise, 50, 906–914. https://doi.org/10.1249/mss.0000000000001527

[8] Luu, B. C., Wright, A. L., Haeberle, H. S., Karnuta, J. M., Schickendantz, M. S., Makhni, E. C., et al. (2020). Machine learning outperforms logistic regression analysis to predict next-season NHL player injury: An analysis of 2322 players from 2007 to 2017. Orthopaedic Journal of Sports Medicine, 8. https://doi.org/10.1177/2325967120953404

[9] Van Eetvelde, H., Mendonça, L. D., Ley, C., Seil, R., & Tischer, T. (2021). Machine learning methods in sport injury prediction and prevention: A systematic review. Journal of Experimental Orthopaedics, 8. https://doi.org/10.1186/s40634-021-00346-x

[10] Claudino, J. G., Capanema, D. d. O., de Souza, T. V., Serrão, J. C., Machado Pereira, A. C., & Nassis, G. P. (2019). Current approaches to the use of artificial intelligence for injury risk assessment and performance prediction in team sports: A systematic review. Sports Medicine – Open, 5. https://doi.org/10.1186/s40798-019-0202-3

[11] Wolff, R. F., Moons, K. G., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., et al. (2019). PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. Annals of Internal Medicine, 170, 51–58. https://doi.org/10.7326/m18-1376

[12] 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, e0118432. https://doi.org/10.1371/journal.pone.0118432

[13] Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. BMC Medicine, 17. https://doi.org/10.1186/s12916-019-1466-7

[14] Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. Medical Decision Making, 26, 565–574. https://doi.org/10.1177/0272989x06295361

[15] Bahr, R. (2016). Why screening tests to predict injury do not work—and probably never will…: A critical review. British Journal of Sports Medicine, 50, 776–780. https://doi.org/10.1136/bjsports-2016-096256

[16] Bittencourt, N. F. N., Meeuwisse, W. H., Mendonça, L. D., Nettel-Aguirre, A., Ocarino, J. M., & Fonseca, S. T. (2016). Complex systems approach for sports injuries: Moving from risk factor identification to injury pattern recognition—Narrative review and new concept. British Journal of Sports Medicine, 50, 1309–1314. https://doi.org/10.1136/bjsports-2015-095850

[17] Hulme, A., & Finch, C. F. (2015). From monocausality to systems thinking: A complementary and alternative conceptual approach for better understanding the development and prevention of sports injury. Injury Epidemiology, 2. https://doi.org/10.1186/s40621-015-0064-1

[18] Gabbett, T. J. (2016). The training—injury prevention paradox: Should athletes be training smarter and harder? British Journal of Sports Medicine, 50, 273–280. https://doi.org/10.1136/bjsports-2015-095788

[19] Impellizzeri, F. M., Tenan, M. S., Kempton, T., Novak, A., & Coutts, A. J. (2020). Acute:Chronic workload ratio: Conceptual issues and fundamental pitfalls. International Journal of Sports Physiology and Performance, 15, 907–913. https://doi.org/10.1123/ijspp.2019-0864

[20] Kalkhoven, J. T., Watsford, M. L., Coutts, A. J., Edwards, W. B., & Impellizzeri, F. M. (2021). Training load and injury: Causal pathways and future directions. Sports Medicine, 51, 1137–1150. https://doi.org/10.1007/s40279-020-01413-6

[21] Midoglu, C., Kjæreng Winther, A., Boeker, M., Dahl Pettersen, S., Pedersen, S., Ragab, N., et al. (2024). A large-scale multivariate soccer athlete health, performance, and position monitoring dataset. Scientific Data, 11. https://doi.org/10.1038/s41597-024-03386-x

[22] Lövdal, S. S., Den Hartigh, R. J., & Azzopardi, G. (2021). Injury prediction in competitive runners with machine learning. International Journal of Sports Physiology and Performance, 16, 1522–1531. https://doi.org/10.1123/ijspp.2020-0518

[23] Teramoto, M., Cross, C. L., Cushman, D. M., Maak, T. G., Petron, D. J., & Willick, S. E. (2017). Game injuries in relation to game schedules in the National Basketball Association. Journal of Science and Medicine in Sport, 20, 230–235. https://doi.org/10.1016/j.jsams.2016.08.020

[24] Lewis, M. (2018). It's a hard-knock life: Game load, fatigue, and injury risk in the National Basketball Association. Journal of Athletic Training, 53, 503–509. https://doi.org/10.4085/1062-6050-243-17

[25] Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., et al. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385, e078378. https://doi.org/10.1136/bmj-2023-078378

[26] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). https://doi.org/10.1145/2939672.2939785

[27] Foster, C. (1998). Monitoring training in athletes with reference to overtraining syndrome. Medicine & Science in Sports & Exercise, 30, 1164–1168. https://doi.org/10.1097/00005768-199807000-00023

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

[29] Ekstrand, J., Hägglund, M., & Waldén, M. (2011). Injury incidence and injury patterns in professional football: The UEFA injury study. British Journal of Sports Medicine, 45, 553–558. https://doi.org/10.1136/bjsm.2009.060582

[30] Riley, R. D., Ensor, J., Snell, K. I. E., Harrell, F. E., Martin, G. P., Reitsma, J. B., et al. (2020). Calculating the sample size required for developing a clinical prediction model. BMJ, 368, m441. https://doi.org/10.1136/bmj.m441

[31] van Smeden, M., Moons, K. G., de Groot, J. A., Collins, G. S., Altman, D. G., Eijkemans, M. J., et al. (2019). Sample size for binary logistic prediction models: Beyond events per variable criteria. Statistical Methods in Medical Research, 28, 2455–2474. https://doi.org/10.1177/0962280218784726

[32] 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, 281–291. https://doi.org/10.1136/bjsports-2015-094758

[33] Windt, J., & Gabbett, T. J. (2017). How do training and competition workloads relate to injury? The workload-injury aetiology model. British Journal of Sports Medicine, 51, 428–435. https://doi.org/10.1136/bjsports-2016-096040

[34] Fuller, C. W., Ekstrand, J., Junge, A., Andersen, T. E., Bahr, R., Dvorak, J., et al. (2006). Consensus statement on injury definitions and data collection procedures in studies of football (soccer) injuries. British Journal of Sports Medicine, 40, 193–201. https://doi.org/10.1136/bjsm.2005.025270

[35] Drew, M. K., & Finch, C. F. (2016). The relationship between training load and injury, illness and soreness: A systematic and literature review. Sports Medicine, 46, 861–883. https://doi.org/10.1007/s40279-015-0459-8

[36] Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine, 15, 361–387. https://doi.org/10.1002/(sici)1097-0258(19960229)15:4<361::aid-sim168>3.0.co;2-4