Modeling Training Loads and Injuries: The Dangers of Discretization

DAVID L. CAREY[1,2], KAY M. CROSSLEY[1], ROD WHITELEY[3], ANDREA MOSLER[1,3], KOK-LEONG ONG[4], JUSTIN CROW[2], and MEG E. MORRIS[1,5]

[1]La Trobe Sport and Exercise Medicine Research Centre, College of Science, Health and Engineering, La Trobe University, Melbourne, AUSTRALIA; [2]Essendon Football Club, Melbourne, AUSTRALIA; [3]Rehabilitation Department, Aspetar Orthopedic and Sports Medicine Hospital, Doha, QATAR; [4]Research Centre for Data Analytics and Cognition, La Trobe University, Melbourne, AUSTRALIA; and [5]Healthscope, Northpark Private Hospital, Melbourne, AUSTRALIA

ABSTRACT

CAREY, D. L., K. M. CROSSLEY, R. WHITELEY, A. MOSLER, K.-L. ONG, J. CROW, and M. E. MORRIS. Modeling Training Loads and Injuries: The Dangers of Discretization. *Med. Sci. Sports Exerc.*, Vol. 50, No. 11, pp. 2267–2276, 2018. **Purpose**: To evaluate common modeling strategies in training load and injury risk research when modeling continuous variables and interpreting continuous risk estimates; and present improved modeling strategies. **Method**: Workload data were pooled from Australian football (n = 2550) and soccer (n = 23,742) populations to create a representative sample of acute:chronic workload ratio observations for team sports. Injuries were simulated in the data using three predefined risk profiles (U-shaped, flat and S-shaped). One-hundred data sets were simulated with sample sizes of 1000 and 5000 observations. Discrete modeling methods were compared with continuous methods (spline regression and fractional polynomials) for their ability to fit the defined risk profiles. Models were evaluated using measures of discrimination (area under receiver operator characteristic [ROC] curve) and calibration (Brier score, logarithmic scoring). **Results**: Discrete models were inferior to continuous methods for fitting the true injury risk profiles in the data. Discrete methods had higher false discovery rates (16%–21%) than continuous methods (3%–7%). Evaluating models using the area under the ROC curve incorrectly identified discrete models as superior in over 30% of simulations. Brier and logarithmic scoring was more suited to assessing model performance with less than 6% discrete model selection rate. **Conclusions**: Many studies on the relationship between training loads and injury that have used regression modeling have significant limitations due to improper discretization of continuous variables and risk estimates. Continuous methods are more suited to modeling the relationship between training load and injury. Comparing injury risk models using ROC curves can lead to inferior model selection. Measures of calibration are more informative judging the utility of injury risk models. **Key Words:** ACUTE:CHRONIC WORKLOAD RATIO, INJURY RISK, ROC CURVES, CALIBRATION

EPIDEMIOLOGY

One of the challenges for coaches, physical preparation practitioners, clinicians, and researchers in sports science and sports medicine is estimating the risk of injury during sporting competitions and training (1,2). Relationships between training loads and injuries have been studied extensively in recent publications (2–14). Training load has been reported as a key injury risk factor in recent consensus statements (1,15). Studies of training loads and injuries often model the relationships between continuous risk factors (e.g., cumulative load or acute:chronic workload ratio [ACWR]) and binary outcomes (injury or no-injury) (4–14).

Discretization is the practice of transforming continuous data into discrete categories and is a prevalent methodology in studies of training load and injury risk (4–10). Discretization methods in sports medicine research include median splits (5,7), percentiles (5,6,13), z-score categories (4,7), and arbitrary bins (8–10). These methodologies have not been critically examined in the context of modeling training loads and injuries. Discretization of continuous covariates in risk models has been criticized in other fields (16–19). Discretization of a continuous risk factor into categories assumes that each individual within that category has equal risk. For example, if cumulative training load is split into low, medium, and high categories using percentiles, then it is assumed that each athlete in the high category has identical risk, irrespective of how broad the category is (i.e., an athlete at the 67th percentile is considered to be at the same risk as one at the 99th percentile). This practice causes a loss of information because within-category variation is ignored (17). The loss of information lowers the statistical power of the study and may reduce the ability to detect relationships between variables, increasing the

Address for correspondence: David L. Carey, M.Sc., La Trobe Sport and Exercise Medicine Research Centre, College of Science, Health and Engineering, La Trobe University, Plenty Road & Kingsbury Drive, Melbourne 3086, Australia; E-mail: d.carey@latrobe.edu.au.
Submitted for publication February 2018.
Accepted for publication May 2018.
Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.acsm-msse.org).

0195-9131/18/5011-2267/0
MEDICINE & SCIENCE IN SPORTS & EXERCISE®

DOI: 10.1249/MSS.0000000000001685

likelihood of a false-negative result (17,18,20). Discretization can also lead to inflated false discovery rates (17,18). It is common for studies using discretization to analyze results by choosing a reference category and making multiple comparisons to each other category, increasing the chance of finding a significant result (17). It has also been shown by Wainer et al. (21) that categorization of continuous variables can make trends appear in otherwise unrelated data if there is freedom to choose the boundaries of the categories. Modeling methods that allow risk factors to vary continuously, such as cubic regression splines and fractional polynomials have therefore been advocated as appropriate alternatives to discretization for modeling nonlinear risk profiles in epidemiology (16,20).

The increase in studies investigating training load as a risk factor for injury has been accompanied by an increase in studies exploring injury prediction (5,11,12,14,22,23). Injury prediction models have been evaluated and compared using metrics, such as sensitivity, specificity or area under the receiver operator characteristic curve (AUC) (5,11,12,14,22,23). These scoring metrics are designed to evaluate binary predictions (i.e., injury or no-injury) and look at how often the model predictions match the actual outcomes (24). In a practical setting, where there is a clinician or coach to synthesize other sources of information to make a contextualized judgment, a model would not be expected to make a yes/no decision. In this scenario, it could be more informative to evaluate injury risk models using measures of calibration (19,25). Calibration refers to how well a model is able to estimate the probability of an event (24).

In this study, we critically evaluated the modeling and evaluation methodologies found in the existing literature on the relationships between training load and injury. Training load data collected from Australian football (6) and soccer (26) were used to generate a set of hypothetical data sets with known injury risk profiles (27). Discrete risk models using *z* score, percentile, and arbitrary binning methods (4–10) were compared with continuous methods, regression splines, and fractional polynomials (16,20). Models were evaluated using measures of discrimination (AUC) and calibration (Brier score) to assess which metrics were the most informative for assessing the utility of risk models (24).

EPIDEMIOLOGY

METHODS

Training Load Data

The ACWR is a relative training load variable calculated by dividing an athlete's acute workload (typically 1 wk) by their chronic workload (typically 4 wk) (2,27). It is a bounded continuous variable that has been studied extensively as an injury risk factor (2–14). The ACWR data were pooled from two studies on separate male populations; a two season study at a single Australian Football club (6) (n = 2550), and a two-season study of 17 soccer teams in the Qatar Stars League (26) (n = 23,742). One-week acute and 4-wk chronic periods (overlapping) were used for both data sets. Total distance was used as the load variable in the Australian football data set and training/match duration in the soccer data set (the only available load metric). Combined, these data had a mean and standard deviation of 1.05 and 0.42; similar to values reported in previous studies (5,7,8) (see Figure, Supplemental Digital Content 1, histogram of ACWR values, http://links.lww.com/MSS/B302). The pooling of data from independent sources was done to ensure the distribution of values used in the simulations was as representative as possible (i.e., it is a good approximation of what a researcher could expect to collect in a hypothetical future study). Alternate ACWR calculation methods that use exponentially weighted averages (28) or decouple the acute and chronic time windows (29) have been proposed. These modifications likely change the distribution of ACWR values (e.g., decoupling causes the ACWR to become unbounded). Despite this, each method still produces a continuous variable and the investigation into the effects of discretization in this study remains relevant irrespective of the ACWR calculation method. Ethical approval for this study was obtained from the Shafallah Medical Genetics Centre, Approval number: 2012-017 and the La Trobe University Faculty of Health Sciences Human Ethics Committee (FHEC14/233). Informed consent was obtained from the participating teams for the analysis of deidentified data.

Injury Risk Profiles

Traditional research designs collect data and build models in an attempt to estimate the true relationship between variables of interest (e.g., ACWR and injury). To evaluate different modeling approaches we have used a different strategy. Artificial injuries were inserted into existing training load data based on predefined risk profiles. This enabled us to compare different models based on how well they were able to recover the true relationship in the data. Three predefined theoretical risk profiles were considered (Fig. 1).

- U-shaped: To align with the hypothesized relationship between ACWR and injury (2,27), with minimum risk corresponding to ACWR = 1.

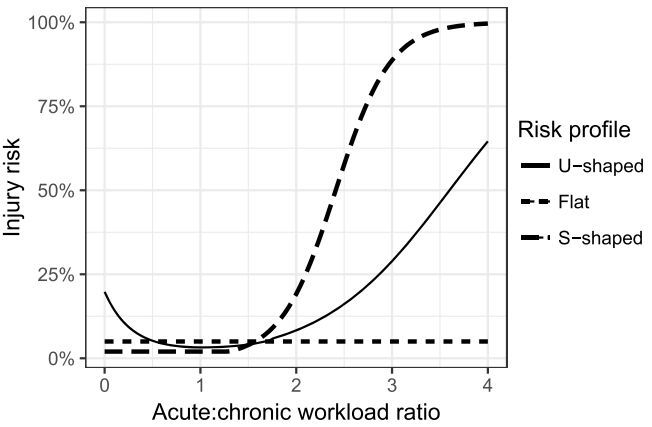

FIGURE 1—Theoretical risk profiles used to simulate injuries.

- Flat: To represent the null hypothesis that ACWR does not influence injury risk (every observation poses a uniform 5% injury risk).
- S-shaped: An alternative risk profile that has injury risk as constant (2%) for ACWR < 1 then rises sharply to very high injury risk.

Details on the mathematical form of risk curves can be found in the supplementary text (see text, Supplemental Digital Content 2, equations of risk profiles, http://links.lww.com/MSS/B303).

Simulating Study Data

To examine the outcomes of different modeling approaches, we simulated hypothetical new studies using the data collected from Australian football and soccer (6,26). The simulation procedure was:

- Choose a sample size (N_s) and randomly choose N_s observations of ACWR from the pooled data distribution (see Figure, Supplemental Digital Content 1, histogram of ACWR values, http://links.lww.com/MSS/B302).
- Assign an injury probability (p_i) to each observation using one of the predefined theoretical risk profiles (Fig. 1).
- Randomly generate injuries by treating each observation as a Bernoulli trial with probability of injury p_i. Simply, this means for an observation with injury risk of 20%, we randomly assigned it an injury or no-injury label with probability 0.2 and 0.8, respectively.

We considered study sizes of 1000 and 5000 observations (representing a single-season or multiseason study in team sport) and three different risk profiles (U-shaped, flat, and S-shaped). For each of these six combinations, we simulated 100 studies to estimate the variability in any results. Simulations were performed using the R statistical computing language (30). An implementation of the simulation procedure is included in the supplementary code (see text, Supplemental Digital Content 3, simulation code, http://links.lww.com/MSS/B304).

Training Load—Injury Models

Two types of modeling approach were considered, discrete and continuous. Discrete models were defined as those that applied a discretization strategy to the ACWR values *before* modeling them against injury incidence. We considered three different discretization methods to reflect those found in the existing literature (4,5,7–10,13).

- D1: Normalize ACWR values (*z*-score) then split into seven categories using cutpoints: $\{-\infty, -2, -1, 0, 1, 2, 3, \infty\}$ (4,7).
- D2: Split into five quantiles (5,6,13).
- D3: Split the ACWR into five categories using the cut-points: $\{0, 1, 1.35, 1.5, 2, \infty\}$ (10).

After discretization, ACWR was modeled against injury incidence using binary logistic regression, with the central group used as the reference level. This method of analysis replicates that commonly used in previous studies (8–10,13).

To contrast the discrete models, two continuous modeling methodologies (C1 and C2) were considered. The continuous methods apply a transformation to the independent variable (ACWR) within the logistic regression. This allows for nonlinear relationships that vary continuously to be modeled.

- C1: Restricted cubic splines model relationships by subdividing the range of values of the covariate (at locations called knots), and fitting a cubic polynomial between each pair of knots. The polynomials are constrained to join smoothly at each knot and to be linear in the two outermost regions (19). Restricted cubic splines are a common method of analysis in epidemiological studies of nonlinear dose–response relationships (16,19,20). Spline models were fitted in R using the *splines* package (30). Spline regression models were constructed with three internal knots placed at equally spaced percentiles (19,20). The number of knots was chosen *a priori* in this study but in general can be chosen by comparing multiple options using an objective criterion (e.g., Akaike information criterion (AIC)) (19).
- C2: Fractional polynomials are a flexible method of modeling nonlinear, continuous relationships. Fractional polynomials consider a combination of candidate functions and select a final model after a series of tests for nonlinearity and complexity (31). A potential benefit of fractional polynomials over cubic splines is that they are more interpretable. The final model can be described by a closed form equation and offers potential insight into the underlying relationship. Their drawback is that they are a global model (i.e., they fit the entire range of data with a single function) and, therefore, cannot fit local features as well as splines. Fractional polynomial models were fitted in R using the *mfp* package (30,32).

Presently, few studies of the relationship between ACWR and injury have used modeling methods that allow for nonlinear trends and avoid discretization of the ACWR. An implementation of each modeling method considered is included in the supplementary code (see text, Supplemental Digital Content 3, simulation code, http://links.lww.com/MSS/B304).

Each of the models (discrete and continuous) was used to produce estimates of injury risk for each ACWR observation in the simulated data sets. This replicates a study design from a team sport environment where workload risk factor and injury outcomes are recorded daily.

Evaluating Injury Models

Comparison between true and modeled risk curves. A direct comparison can be made between the modeled risk profile and the true risk profile in this study

because the function used to simulate the injuries was predefined (i.e., it is exactly known, as shown in Fig. 1). Root mean square error (RMSE) (24) was calculated for the difference between the true risk and predicted risk for each observation in each simulated study. This provides a measure of how well the modeling procedure was able to recover the true risk profile used to generate the data.

False discovery and false rejection rates. The flat injury risk profile was used to estimate the false discovery rate for each modeling approach (Fig. 1). Data simulated under the flat profile contained no association between ACWR and injury risk. Therefore, any simulated study finding a significant relationship in the data could be considered a false discovery (Type I error). Significance testing for discrete models (D1, D2, D3) was performed by comparing the reference ACWR level to all other levels in the discretized ACWR (4,5,7–10,13). A simulation was deemed to have a significant finding if any of the 95% confidence intervals for the odds ratios did not contain 1. Significance testing for spline regression (C1) and fractional polynomials (C2) was performed by comparing to a null model using the likelihood ratio test with $\alpha = 0.05$ (32).

False rejection rates (type II error) were estimated for each model by counting the number of times no significant result was found when the data were simulated with a U-shaped or S-shaped risk profile. Discretizing continuous variables causes a decrease in statistical power (17,18), potentially causing the false rejection rates of discrete models to increase.

Receiver operator characteristic. The AUC has been used to evaluate predictive models of training load and injury in previous studies (5,11,12,14). The AUC measures the ability of the model to discriminate between the two outcome classes (injury and no-injury). It has been used as a way to select the best performing injury prediction model in studies comparing multiple methods (11,12,14,23). Cross-validation (10-fold) was used to obtain estimates of AUC for each simulated study. Without some kind of resampling or out-of-sample testing, the results can be positively biased (i.e., they will be better than could be expected in practice) (24).

Calibration. Calibration is a measure of how well a model is able to estimate the probability of an event. It can be assessed visually by constructing calibration curves (19,33). Calibration curves show how closely the predicted probabilities match the observed event rates (i.e., for observations estimated to have injury risk of 20%—was the actual injury incidence rate on those days around 20%?). Calibration can also be assessed quantitatively by computing the Brier score or logarithmic scoring rule (19). In the case of a binary outcome variable, the Brier score is calculated as the square of the probability assigned to the incorrect class (e.g., if the model predicts injury with probability 0.2, and there was no injury, the Brier score would be $0.2^2 = 0.04$, but if there *was* an injury, then the score would be $0.8^2 = 0.16$). A lower Brier score indicates a better model. The logarithmic scoring rule is evaluated by taking the natural logarithm of the probability assigned to the correct class (e.g., a predicted injury probability of 0.2 and no injury would score $\log(0.8) = -0.22$, and if there was, an injury would score $\log(0.2) = -1.61$). A higher score indicates better probability estimates. Logarithmic scoring may be more appropriate than the Brier score in the case of rare event estimation (34). Brier and logarithmic scores were estimated for each model and simulated study using 10-fold cross-validation (24).

Longitudinal Models of Training Loads and Injury

For clarity of message in the previous sections, we have simulated ACWR and injury data with no correlation structure and used logistic regression assuming independence of observations to illustrate the effects of discretization. However, training load monitoring data collected from sporting teams often consists of repeated measurements taken from the same athletes. It is therefore possible that the observations from the same athletes will be correlated. To investigate the effects of this correlation on injury risk modeling, we simulated longitudinal training load data sets using the *SimCorMultRes* package (35) (see text, Supplemental Digital Content 3, simulation code, http://links.lww.com/MSS/B304). Injuries were simulated in the data by defining a marginal risk profile and specifying a within-subject correlation strength (35). Four longitudinal data sets were simulated (100 times each) to investigate the effects of different sample sizes, within-subject correlations and marginal risk profiles. The first simulated 50 observations from 20 participants with a U-shaped marginal risk (Fig. 1) and a within-subject correlation of 0.1. The second increased the correlation strength to 0.7. The third considered a larger sample size of 100 observations from 50 participants. The fourth considered the effect of reducing the strength of the marginal risk by reducing the injury risk by a factor of 0.5 for each ACWR value.

Each longitudinal data set was analyzed using naïve logistic regression (i.e., assuming independence of observations) and generalized estimating equations (GEE) (36). The GEE models have been used in previous studies of training load and injury (5,12,14). The GEE models were fitted using the R package *geepack* (37) using a binomial link and exchangeable working correlation structure. Both analysis methods allowed for the relationship between ACWR and injury risk to vary continuously using restricted cubic splines (as previously described). Modeling approaches were compared for their ability to recover the predefined marginal effect of ACWR on injury risk using RMSE. Additionally, significance testing was performed for each simulated study result by comparing to a null model using a likelihood ratio test (see text, Supplemental Digital Content 3, simulation code, http://links.lww.com/MSS/B304).

RESULTS

Simulated Studies

Details of the simulated studies (injury summary statistics) are found in supplementary Table 1 (see Table, Supplemental Digital Content 4, simulated injury statistics, http://links.lww.com/MSS/B305).

Comparison between True and Modeled Risk Profiles

A visual inspection of how well each modeling procedure was able to recover the true risk profile used to generate the data is shown in Figure 2. It is clear that models that discretized the ACWR (Fig. 2(D1-3)) are unable to fully capture the U-shaped relationship. Continuous modeling methods C1 (spline regression) and C2 (fractional polynomials) fared much better at fitting the true risk profile (see Figures, Supplemental Digital Content 5, S-shaped risk, http://links.lww.com/MSS/B306 and Supplemental Digital Content 6, flat risk, http://links.lww.com/MSS/B307).

The RMSE performance of each modeling strategy under the different simulation parameters is shown in Figure 3. Continuous modeling methods (C1, C2) had noticeably lower RMSE for data generated using a U or S-shaped injury risk profile (particularly in larger simulated studies with $N_s = 5000$). The difference between discrete and continuous methods was less pronounced for the flat injury risk profile. In general, the total error and variance in error for each model tended to decrease when the simulated sample size increased from 1000 to 5000 observations.

False Discovery Rates

Discrete modeling methods had higher false discovery rates than continuous methods (Fig. 4). For 100 simulated studies with flat injury risk profile (i.e., no relationship in the data) and 5000 observations, discrete models (D1, D2, D3) had false discoveries 21, 16, and 16 times, respectively. The continuous methods had false discovery rates of 7/100 and 3/100 for C1 and C2, respectively. Alarmingly, in the 100 simulated studies, we found that at least one of the three discrete methods had a false discovery 42 times.

Choice of reference level. Discretizing the ACWR then running a logistic regression introduces another choice into the modeling procedure when the reference level is chosen by the researcher. There has been little consistency in existing studies, with the lowest (10,13), highest (8), and central ACWR interval (5) being used. This freedom of choice is an issue because it can change the reported findings. For example, using discrete model D1 and a flat risk profile; 11 of 100 simulations had a false discovery if the highest interval was used as the reference, but if the central interval was used, this increased to 21 of 100 false discoveries. Avoiding discretization and modeling a continuous relationship removes this choice.

False Rejection Rates

Discrete methods (D2, D3) had higher false rejection rates when data were simulated with U-shaped or S-shaped risk profiles (see Table, Supplemental Digital Content 7, false rejection rates, http://links.lww.com/MSS/B308). For data sets with 1000 observations and a U-shaped risk relationship, 59 and 57 of 100 simulated studies did not find a significant result when analyzed using discrete methods D2 and D3 respectively. The false rejection rate was much lower when using methods D1 (5/100), C1 (12/100) or C2 (19/100). As expected, increasing the sample size from 1000 to 5000 observations reduced the false rejection rates for each modeling approach

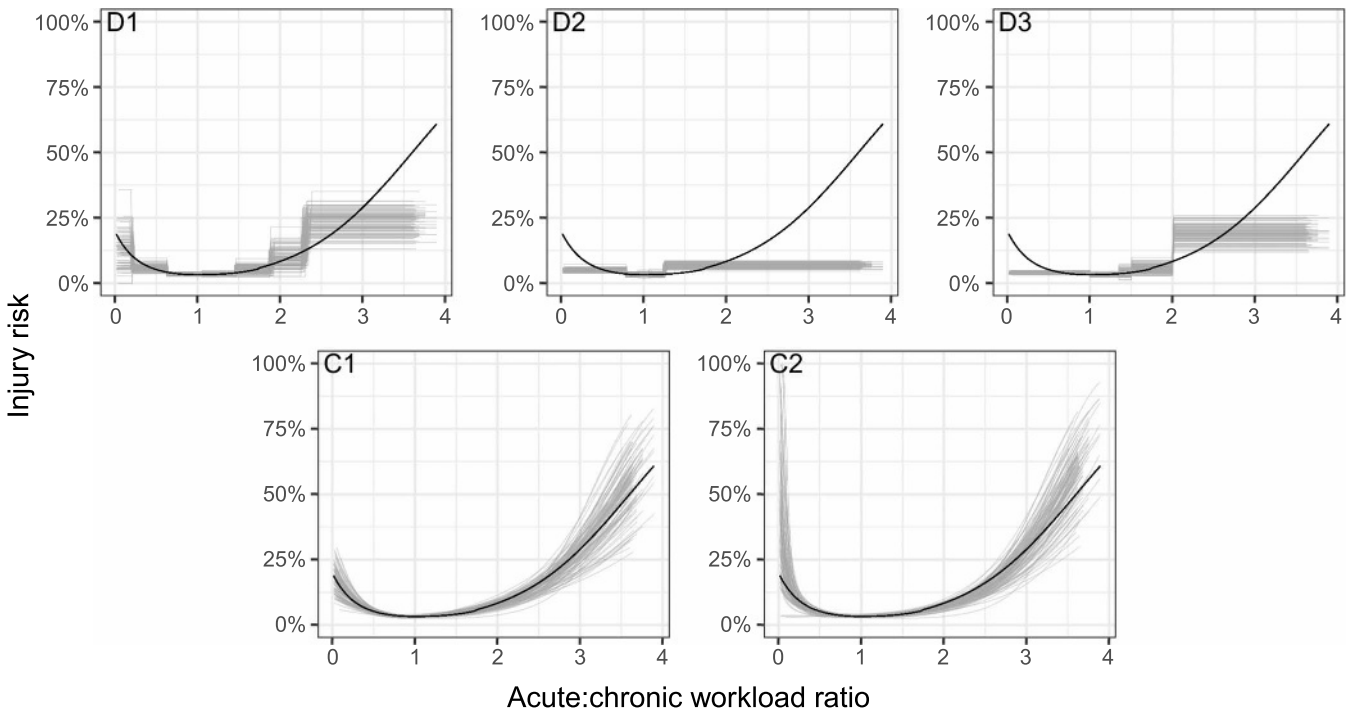

FIGURE 2—Comparison of 100 simulated study results ($N_s = 5000$ and U-shaped risk) analyzed using discrete models (D1, D2, D3) and continuous models (C1, C2). *Solid line* represents the true risk profile used to generate the data and each grey line represents one simulated study result.

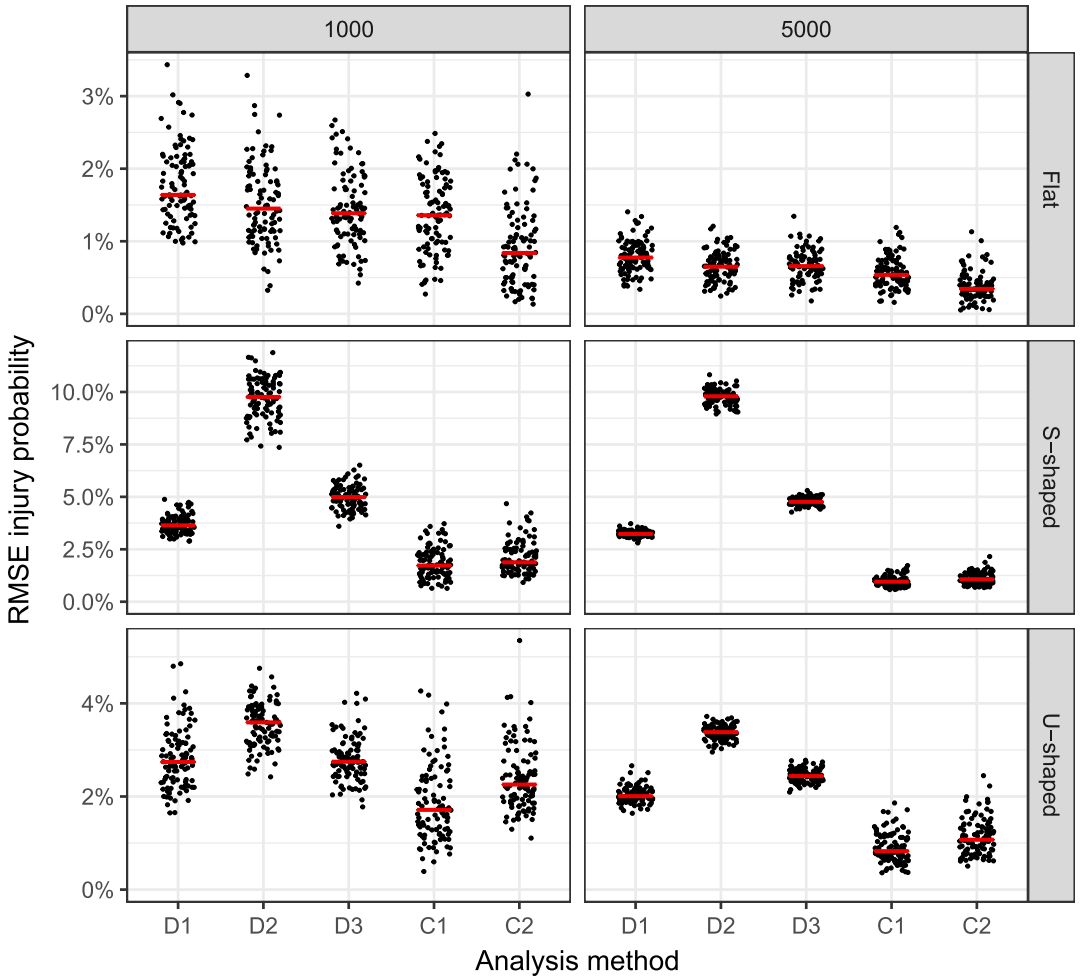

FIGURE 3—RMSE of model probability estimates for 100 trials of each theoretical risk profile and sample size (*red bar*, median).

(see Table, Supplemental Digital Content 7, false rejection rates, http://links.lww.com/MSS/B308).

Receiver Operator Characteristics

Area under the receiver operator characteristic (ROC) curve was estimated for each of the 100 simulated studies using 10-fold cross-validation (24). The continuous analysis methods had higher median AUC values but did not clearly outperform discrete methods under this evaluation metric. If AUC was used to select the best performing model in each simulation, we found that one of the discrete models was chosen on 38 of 100 occasions and 31 of 100 occasions for U-shaped and S-shaped risks, respectively (Table 1).

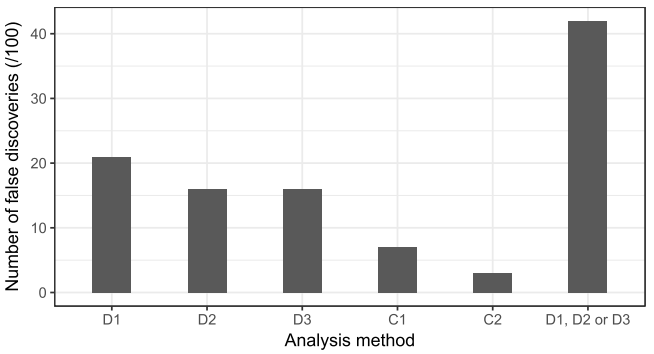

FIGURE 4—False discovery rates (of 100 simulated studies with N_s = 5000 and flat risk profile).

Calibration

To compare with ROC curves, Brier and logarithmic scores were estimated for each model using 10-fold cross-validation (19,24). When the Brier score was used to select the best performing model in each simulation, discrete models were chosen on only 6 of 100 occasions and 0 of 100 occasions for U-shaped and S-shaped risks, respectively (Table 1). When logarithmic scoring was used the rates were 3 of 100 and 1 of 100. Brier and logarithmic scoring favored the continuous methods far more than evaluation with ROC curves.

Calibration curves offer a way to visually evaluate injury risk models (Fig. 5). A calibration curve shows the relationship between the predicted probabilities and actual event occurrence rate (perfect calibration is represented by the diagonal line). An exemplar set of calibration curves is shown in Figure 5 (one simulated study with U-shaped risk and N_s = 5000). Ideally, a

TABLE 1. Comparison of model selection rates using AUC, Brier score, and logarithmic scoring as the evaluation metric (N_s = 5000).

	No. Times Selected as Best Model (/100 Simulated Studies)					
	U-Shaped Risk			S-Shaped Risk		
Method ID	AUC	Brier	Logarithmic	AUC	Brier	Logarithmic
D1	28	6	3	15	0	1
D2	2	0	0	0	0	0
D3	8	0	0	16	0	0
C1	35	80	70	26	73	75
C2	27	14	27	43	27	24

well-calibrated risk model will have a curve that is close to the diagonal line and covers a large range of probabilities (i.e., has confidence in identifying both high- and low-risk scenarios). Discrete model D2 provided little information other than the baseline injury rate. Model D3 did not appear to be well calibrated. Models D1, C1, and C2 were well calibrated (close to diagonal line); however, the continuous methods covered a much larger range of probabilities.

Longitudinal Data Models

The GEE and naive logistic regression models had similar ability to recover the marginal effect in each simulated longitudinal data set (Table 2). Median RMSE values were near identical for each approach. Increasing the sample size (100 observations from 50 participants) lowered the median RMSE values whilst increasing the within-participant correlation strength increased the median RMSE (Table 2).

The naïve logistic regression approach (assuming independence of observations) had higher false rejection rates (i.e., lower statistical power) than the GEE approach (Table 2). The difference in false rejection rates became more pronounced when the strength of relationship between ACWR and injury risk was decreased (47/100 for logistic vs 18/100 for GEE). Using a larger sample size caused the false rejection rate to drop to zero for both methods. Increasing the strength of within-participant correlation did not have a strong effect on false rejection rates.

DISCUSSION

Discrete versus Continuous Modeling Strategies

Discrete models showed limited ability to capture the risk profiles used to generate the simulation data (Figs. 2–3). Discretization forced the models to fit an unrealistic and discontinuous step profile to the data (Fig. 2 and Supplemental Digital Content 5, S-shaped risk, http://links.lww.com/MSS/B306). This illustrates how discretization of continuous risk factors can lead to inaccurate estimation of effects (17,20). Figure 2 shows how using percentile splits (method D2) groups a large range of ACWR values together and provides an inaccurate estimated effect that is far lower than the true risk for ACWR values greater than 2. Similarly, the ACWR categories used in method D3 assume homogeneity

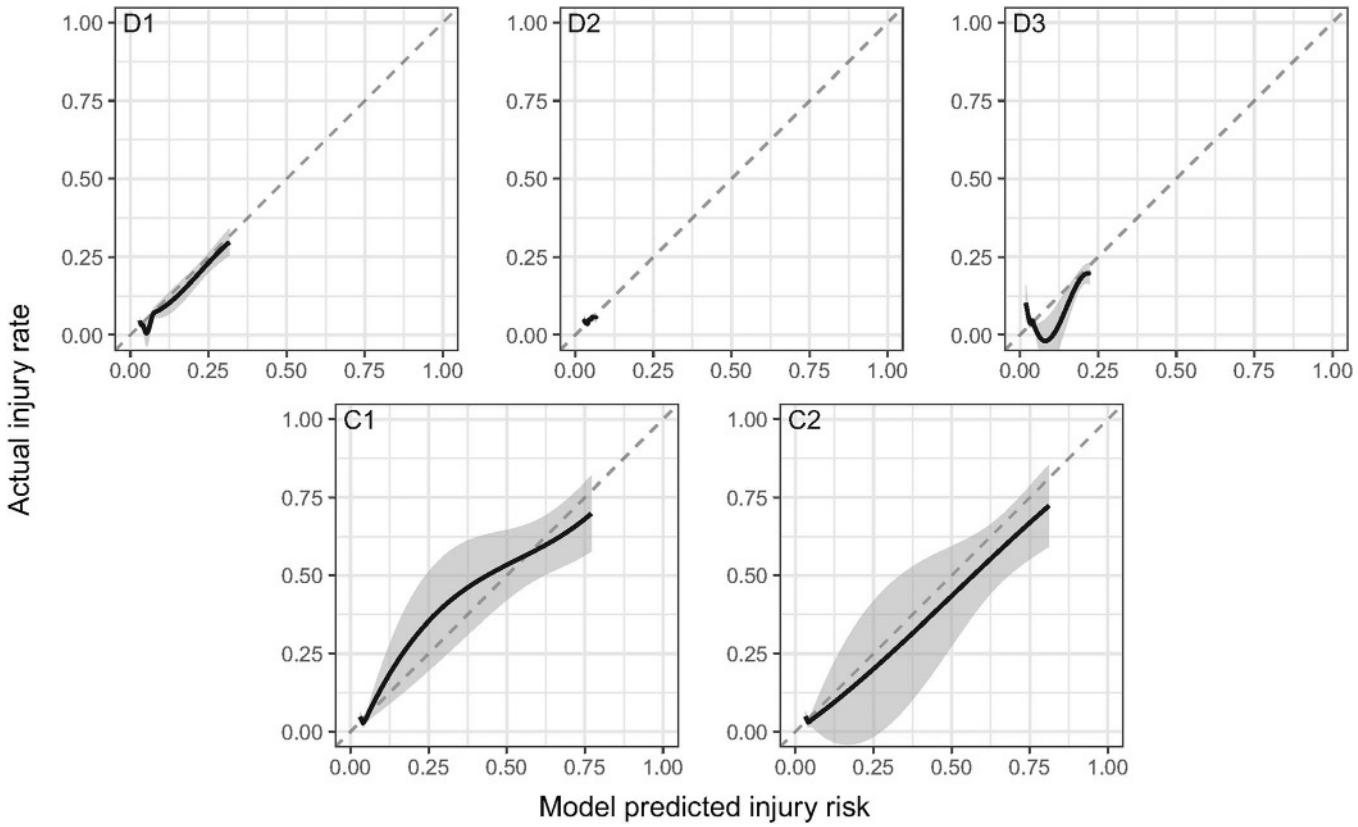

FIGURE 5—Comparison of cross-validated calibration curves from a single simulated study (N_s = 5000 and U-shaped risk) analyzed using discrete (D1, D2, D3) and continuous models (C1, C2). *Diagonal line* represents perfect calibration and shaded area represents 95% CI. 95% CI, 95% confidence interval.

TABLE 2. Comparison of logistic regression and GEE modeling for longitudinal data.

Marginal Risk Profile	Study Size (Observations × Participants)	Correlation Strength	Significant Results (/100)		Median RMSE (IQR)	
			Logistic	GEE	Logistic	GEE
U-shaped	50 × 20	0.1	84	95	0.020 (0.013–0.025)	0.021 (0.014–0.025)
U-shaped	50 × 20	0.7	86	97	0.030 (0.024–0.037)	0.028 (0.021–0.035)
U-shaped	100 × 50	0.1	100	100	0.009 (0.007–0.012)	0.009 (0.007–0.011)
U-shaped (dilated ½)	50 × 20	0.1	53	82	0.013 (0.011–0.017)	0.014 (0.011–0.017)

of risk over the range 0 to 1, leading to an estimated effect that cannot capture the rise in risk seen for small ACWR values. Our simulations suggest that the discrete methods found in the current literature (4–8,10,13) are unsuited to modeling the continuous U-shaped risk profile between ACWR and injury proposed in the literature (2,27).

Continuous modeling methods (spline regression and fractional polynomials) were better suited to fitting the nonlinear risk profiles (U-shaped and S-shaped) and provided more accurate estimated effects. This was demonstrated by lower RMSE scores (Fig. 3) and also confirmed visually by the 100 simulations shown in Figure 2. These findings align with recommendations from other fields that continuous modeling methods are preferable to discretization (17,20). Future studies may benefit from using continuous modeling methods instead of discretizing continuous training load variables when analyzing their relationship to injury.

False Discovery Rates

Data generated under the assumption that ACWR had no relationship to injury risk (Fig. 1, flat risk profile) was used to estimate the false discovery rate for each modeling approach. False discovery rates were inflated by using discrete models (16%–21%) (Fig. 4). Splitting the ACWR into multiple categories before modeling leads to multiple comparisons between groups and may explain the higher false discovery rates (17,18). Discrete method D1 used the most categories (7 groups) and had the highest false discovery rate (21%). A secondary issue was the choice of reference level when categorical predictors are used in generalized linear models (e.g., logistic or Poisson regression). Discrete model D1 had 21/100 false discoveries when the central ACWR category was used as the reference but only 11/110 if the highest was used.

There is currently no consensus in the literature regarding the discretization strategy or choice of reference level when modeling ACWR and injury risk (5,8,10,13). The apparent freedom of choice of discretization and reference level may have caused highly inflated false discovery rates in previous studies (38). When a choice of only three methods was considered in our simulations false discovery rates were as high as 42% (Fig. 4). Continuous modeling methods do not require choosing a reference level and do not suffer from multiple comparisons between predictor categories. Spline regression and fractional polynomials had substantially lower false discovery rates (7% and 3%).

False Rejection Rates

Models that transformed the continuous ACWR into discrete categories showed higher false rejection rates in the simulated studies (see Table, Supplemental Digital Content 7, false rejection rates, http://links.lww.com/MSS/B308). This aligns with findings from other studies that discretization lowers statistical power (17,18,20). Simulations using a larger sample size (N_s = 5000) were not as prone to false rejections, highlighting the benefits of larger sample sizes. The negative consequences of discretization on statistical power are particularly relevant for research in elite sport cohorts where sample sizes are often constrained.

Evaluating Injury Risk Models

ROC curves. Comparing models using the area under the ROC curve did not always identify that continuous methods were better fits to the risk profiles (Table 1). RMSE scores showed that continuous methods were clearly superior when modeling U-shaped or S-shaped risk profiles when a sample size of 5000 observations was used (Fig. 3). Despite this, AUC incorrectly identified discrete methods as superior in 38 and 31/100 simulations for U and S-shaped risk (Table 1). This suggests that using AUC as the sole evaluation metric when selecting injury prediction models (11,12,14) runs the risk of selecting an inferior model.

A ROC curve is constructed by sampling through the possible decision thresholds that could be applied (i.e., cut points where the models makes an injury or no-injury prediction). This may not realistically represent the purpose of the model if it to be used for risk estimation. If the output of the model is used along with context and clinical judgment, and not required to make a binary decision, then AUC may not be an appropriate evaluation metric (25). The ROC curves also assume that false positive errors and false negative errors are of equal consequence (39). This is likely not the case when a false negative means an injured athlete and a false positive may be a modified or missed training session. We suggest that ROC curves in isolation are insufficient to evaluate the performance of injury prediction models.

Probabilistic scoring and calibration curves. Evaluating the modeling strategies with Brier scores and logarithmic scoring strongly favored the continuous models (Table 1). Discrete models were selected in only 6/100 and 0/100 simulations using the Brier score and 3/100 and 1/100 when using logarithmic scoring. These provide a much closer reflection of the RMSE scores (ground truth) than evaluating

models with AUC. The Brier and logarithmic scores are probabilistic scoring rules designed to evaluate probability estimates (19) and are therefore better suited to assessing injury risk models. We suggest that Brier scores, logarithmic scoring, or another comparable probabilistic scoring rule (34) be included in future studies to compare injury risk models.

Calibration curves (Fig. 5) provided an informative visualization of the performance of injury risk models (33). They showed how closely the risk estimates of each model matched the observed injury rates and how well each model discriminated between high and low risk instances. Figure 5 clearly shows that continuous models gave more informative probability estimates (closer to the observed event rates and over a larger range of values) than the discrete models. Calibration curves show absolute risks and thus may be a more important result for clinicians and decision makers (40).

Longitudinal Models

Extending the simulation study to include correlated within-individual observations showed the negative effects of incorrectly assuming independence between repeated measurements. The naïve logistic regression approach had higher false rejection rates than a GEE approach (Table 2). Assuming independence can cause the standard errors for time varying covariates to be overestimated (41) and may have been the cause of the inflated false rejection rates. When the strength of the "signal" in the data was decreased the difference between logistic and GEE approaches became more pronounced, and the naïve logistic approach had very high false rejection rate (47/100). This highlights the importance of accounting for correlated observations when modeling longitudinal training load data, particularly if the expected strength of signal in the data is small.

Both longitudinal modeling approaches showed similar ability to recover the true marginal risk profile. This is likely because the parameter estimates from logistic regression and GEE models are generally very similar (41). In all simulations, larger sample sizes improved the accuracy of model estimated effects, suggesting the potential benefits of collaborative studies with large sample sizes.

Limitations and Extensions

Restricted cubic spline regression and fractional polynomials were considered as the alternative modeling methods in this study. Although they are common approaches for modeling nonlinear relationships (16,20) they are not the only possible approaches. A number of other nonparametric and semi-parametric methods may have been suitable (e.g., locally weighted regression, generalized additive models, and smoothing splines) (19).

We did not consider multivariable modeling and used only a single covariate (ACWR) in our simulations. This was done for clarity of the message. The issues caused by discretization are equally problematic in multivariable modeling. Spline and fractional polynomial techniques can still be used when there is more than one covariate to allow for proper modeling of continuous variables (19,32). For example, recent studies have investigated the effect of ACWR on injury risk moderated by absolute chronic workload dichotomized using a median split (4,5,7). This dichotomization removes a significant amount of variation in the data, leading to decreased statistical power and inaccurate estimation of effects (17,20). It is possible, and we would suggest more appropriate, to avoid discretization and model both risk factors continuously using a technique such as restricted cubic surfaces (19). This study did not consider time-to-event approaches for modeling training loads and injury (e.g., survival analysis and Cox regression (42)). Discretization of baseline or time-varying covariates can have similar consequences on statistical power and estimated effects in these contexts and continuous approaches are advised (19).

CONCLUSIONS

Modeling methods that discretize continuous risk factors are inappropriate for studying the relationship between training loads and injuries. Discrete models have inflated false discovery and false rejection rates and are unsuited to fitting nonlinear risk profiles. Strong justification is required for research that chooses a discrete approach and we suggest avoiding discretization and modeling relationships with continuous methods, such as spline regression or fractional polynomials. Accounting for correlated observations in longitudinal training data decreases the risk of false rejection. Evaluating injury risk models using ROC curves may not reflect their practical use and may lead to inferior model selection. Probabilistic scoring methods, such as Brier scores, logarithmic scoring, and calibration curves, may be more informative when assessing injury models.

The authors report no conflicts of interest. This research is supported by an Australian Government Research Training Program (RTP) Scholarship. The results of the study are presented clearly, honestly, and without fabrication, falsification, or inappropriate data manipulation. The results of the present study do not constitute endorsement by ACSM.

REFERENCES

1. Soligard T, Schwellnus M, Alonso JM, et al. How much is too much? (Part 1) International Olympic Committee consensus statement on load in sport and risk of injury. *Br J Sports Med*. 2016; 50(17):1030–41.
2. Gabbett TJ. The training-injury prevention paradox: should athletes be training smarter and harder? *Br J Sports Med*. 2016;50(5):273–80.
3. Drew MK, Finch CF. The relationship between training load and injury, illness and soreness: a systematic and literature review. *Sports Med*. 2016;46(6):861–83.
4. Bowen L, Gross AS, Gimpel M, Li FX. Accumulated workloads and the acute:chronic workload ratio relate to injury risk in elite youth football players. *Br J Sports Med*. 2017;51(5):452–9.

5. Colby MJ, Dawson B, Peeling P, et al. Multivariate modelling of subjective and objective monitoring data improve the detection of non-contact injury risk in elite Australian footballers. *J Sci Med Sport*. 2017;20(12):1068–74.
6. Carey DL, Blanch P, Ong KL, Crossley KM, Crow J, Morris ME. Training loads and injury risk in Australian football-differing acute: chronic workload ratios influence match injury risk. *Br J Sports Med*. 2017;51(16):1215–20.
7. Hulin BT, Gabbett TJ, Lawson DW, Caputi P, Sampson JA. The acute:chronic workload ratio predicts injury: high chronic workload may decrease injury risk in elite rugby league players. *Br J Sports Med*. 2016;50(4):231–6.
8. Murray NB, Gabbett TJ, Townshend AD, Hulin BT, McLellan CP. Individual and combined effects of acute and chronic running loads on injury risk in elite Australian footballers. *Scand J Med Sci Sports*. 2017;27(9):990–8.
9. Malone S, Owen A, Newton M, Mendes B, Collins KD, Gabbett TJ. The acute:chonic workload ratio in relation to injury risk in professional soccer. *J Sci Med Sport*. 2017;20(6):561–5.
10. Malone S, Roe M, Doran DA, Gabbett TJ, Collins KD. Protection against spikes in workload with aerobic fitness and playing experience: the role of the acute:chronic workload ratio on injury risk in elite Gaelic football. *Int J Sports Physiol Perform*. 2017;12(3): 393–401.
11. Ruddy JD, Shield AJ, Maniar N, et al. Predictive modeling of hamstring strain injuries in elite Australian footballers. *Med Sci Sports Exerc*. 2018;50(5):906–14.
12. Thornton HR, Delaney JA, Duthie GM, Dascombe BJ. Importance of various training load measures on injury incidence of professional rugby league athletes. *Int J Sports Physiol Perform*. 2016:1–17.
13. Malone S, Owen A, Mendes B, Hughes B, Collins K, Gabbett TJ. High-speed running and sprinting as an injury risk factor in soccer: can well-developed physical qualities reduce the risk? *J Sci Med Sport*. 2018;21(3):257–62.
14. Carey DL, Ong K-L, Whiteley R, Crossley KM, Crow J, Morris ME. Predictive modelling of training loads and injury in Australian football. *Int J Comp Sci Sport*. Forthcoming 2018:17.
15. Bourdon PC, Cardinale M, Murray A, et al. Monitoring athlete training loads: consensus statement. *Int J Sports Physiol Perform*. 2017;12(2 Suppl):S2161–70.
16. Greenland S. Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. *Epidemiology*. 1995;6(4): 356–65.
17. Bennette C, Vickers A. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. *BMC Med Res Methodol*. 2012;12:21.
18. Altman DG, Royston P. The cost of dichotomising continuous variables. *BMJ*. 2006;332(7549):1080.
19. Harrell FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer; 2001. xxii, p. 568.
20. Kahan BC, Rushton H, Morris TP, Daniel RM. A comparison of methods to adjust for continuous covariates in the analysis of randomised trials. *BMC Med Res Methodol*. 2016;16:42.
21. Wainer H, Gessaroli M, Verdi M. Visual revelations. *Chance*. 2013;19(1):49–52.
22. Gabbett TJ. The development and application of an injury prediction model for noncontact, soft-tissue injuries in elite collision sport athletes. *J Strength Cond Res*. 2010;24(10):2593–603.
23. Rossi A, Pappalardo L, Cintia P, Fernandez J, Iaia FM, Medina D. Who is going to get hurt? Predicting injuries in professional soccer. In: *Proc the Machine Learning and Data Mining for Sports Analytics Workshop (MLSA'17), ECML/PKDD; 2017 Sep 18*. Skopje (Macedonia); CEUR Workshop Proceedings; 2017. pp. 21–30.
24. Kuhn M, Johnson K. *Applied Predictive Modeling*. New York, NY: Springer New York; 2013. p. 600.
25. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007;115(7):928–35.
26. Mosler AB, Weir A, Eirale C, et al. Epidemiology of time loss groin injuries in a men's professional football league: a 2-year prospective study of 17 clubs and 606 players. *Br J Sports Med*. 2018;52(5): 292–7.
27. Blanch P, Gabbett TJ. Has the athlete trained enough to return to play safely? The acute: chronic workload ratio permits clinicians to quantify a player's risk of subsequent injury. *Br J Sports Med*. 2016;50(8):471–5.
28. Williams S, West S, Cross MJ, Stokes KA. Better way to determine the acute:chronic workload ratio? *Br J Sports Med*. 2017;51(3): 209–10.
29. Lolli L, Batterham AM, Hawkins R, et al. Mathematical coupling causes spurious correlation within the conventional acute-to-chronic workload ratio calculations. *Br J Sports Med*. 2017; pii: bjsports-2017-098110. 10.1136/bjsports-2017-098110.
30. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria; 2014.
31. Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. *Int J Epidemiol*. 1999;28(5):964–74.
32. Ambler G, Benner A. mfp: Multivariable Fractional Polynomials. *R package version 14 9*. 2010.
33. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Stat Med*. 2014;33(3):517–35.
34. Benedetti R. Scoring rules for forecast verification. *Mon Weather Rev*. 2010;138(1):203–11.
35. Touloumis A. Simulating correlated binary and multinomial responses under marginal model specification: the SimCorMultRes package. *The R Journal*. 2016;8(2):79–91.
36. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73(1):13–22.
37. Halekoh U, Højsgaard S, Yan J. The R package geepack for generalized estimating equations. *Journal of Statistical Software*. 2006;15(2):1–11.
38. Simpson D, Whiteley R, Carey D. Workload and Gaelic Football injury risk–have we been fooled by spurious correlations or is Gaelic Football really different gravy? *Phys Ther Sport*. 2017;28:e25.
39. Holmberg L, Vickers A. Evaluation of prediction models for decision-making: beyond calibration and discrimination. *PLoS Med*. 2013;10(7):e1001491.
40. Nielsen RO, Bertelsen ML, Verhagen E, et al. When is a study result important for athletes, clinicians and team coaches/staff? *Br J Sports Med*. 2017;51(20):1454–5.
41. Twisk JWR. *Applied Longitudinal Data Analysis for Epidemiology: a Practical Guide*. Cambridge, UK; New York: Cambridge University Press; 2003. xvi, p. 301.
42. Møller M, Nielsen RO, Attermann J, et al. Handball load and shoulder injury rate: a 31-week cohort study of 679 elite youth handball players. *Br J Sports Med*. 2017;51(4):231–7.