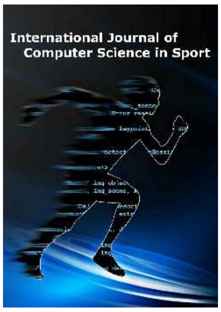
International Journal of Computer Science in Sport

Volume 17, Issue 1, 2018

Journal homepage: http://iacss.org/index.php?id=30

sciendo

DOI: 10.2478/ijcss-2018-0002



Predictive Modelling of Training Loads and Injury in Australian Football

Carey, D. L.[1,4], Ong, K.[2], Whiteley, R.[3], Crossley, K. M.[1], Crow, J.[3, 1], Morris, M. E.[1]

[1]La Trobe Sport and Exercise Medicine Research Centre, College of Science, Health and Engineering, La Trobe University, Melbourne, Australia

[2]SAS Analytics Innovation Lab, La Trobe Business School, La Trobe University, Melbourne, Australia

[3]Aspetar Orthopedic and Sports Medicine Hopsital, Doha, Qatar

[4]Essendon Football Club, Melbourne, Australia

Abstract

To investigate whether training load monitoring data could be used to predict injuries in elite Australian football players, data were collected from athletes over 3 seasons at an Australian football club. Loads were quantified using GPS devices, accelerometers and player perceived exertion ratings. Absolute and relative training load metrics were calculated for each player each day. Injury prediction models (regularised logistic regression, generalised estimating equations, random forests and support vector machines) were built for non-contact, non-contact time-loss and hamstring specific injuries using the first two seasons of data. Injury predictions were then generated for the third season and evaluated using the area under the receiver operator characteristic (AUC). Predictive performance was only marginally better than chance for models of non-contact and non-contact time-loss injuries (AUC<0.65). The best performing model was a multivariate logistic regression for hamstring injuries (best AUC=0.76). Injury prediction models built using training load data from a single club showed poor ability to predict injuries when tested on previously unseen data, suggesting limited application as a daily decision tool for practitioners. Focusing the modelling approach on specific injury types and increasing the amount of training observations may improve predictive models for injury prevention.

KEYWORDS: INJURY, MACHINE LEARNING, TRAINING LOAD

Introduction

Training loads are known to be associated with injuries in team sport athletes (Carey et al., 2016; Colby, Dawson, Heasman, Rogalski, & Gabbett, 2014; Hulin, Gabbett, Lawson, Caputi, & Sampson, 2015; Malone et al., 2016; Murray, Gabbett, Townshend, & Blanch, 2016; Rogalski, Dawson, Heasman, & Gabbett, 2013; Thornton, Delaney, Duthie, & Dascombe, 2016). Monitoring and adjusting loads to reduce injury risk is considered to be an important aspect of athlete management (Soligard et al., 2016), especially since injuries can have a detrimental effect on team sport performance (Hägglund et al., 2013). Existing studies have found associations between injuries and both absolute and relative training loads (Drew & Finch, 2016). Therefore monitoring these metrics has been recommended in recent reviews (Bourdon et al., 2017; Drew & Finch, 2016; Soligard et al., 2016). Relative training loads are typically quantified using the acute:chronic workload ratio (Blanch & Gabbett, 2016; Gabbett, 2016; Hulin et al., 2015; Murray et al., 2016; Williams, West, Cross, & Stokes, 2016b). This is calculated by dividing the short term (acute) training load for an athlete, typically around 7 days, by their load over a longer (chronic) period, typically 3 weeks or longer. Absolute training loads are usually reported using cumulative totals or absolute weekly loads (Colby et al., 2014; Rogalski et al., 2013). Other methods of training load calculation that take into account daily variations, such as monotony and strain, are also useful for athlete monitoring (Foster et al., 2001).

Multiple metrics exist for quantifying load in team sport athletes (Bourdon et al., 2017; Soligard et al., 2016). External loads refer to objective measures of training activity such as training duration, running distance, or accelerations (Bourdon et al., 2017; Soligard et al., 2016). They can be measured using sensor technologies such as global positioning systems (GPS) and accelerometers (Bourdon et al., 2017; Soligard et al., 2016). Internal training load is defined as the physiological and psychological response to external loads. Internal load is typically assessed using metrics derived from heart rate, blood lactate and athlete ratings of perceived exertion (RPE) (Bourdon et al., 2017; Soligard et al., 2016).

It has been suggested that machine learning approaches may provide a way to progress the field of training load monitoring and injury prediction by allowing for non-linear pattern recognition and interactions between variables (Bittencourt et al., 2016; Bourdon et al., 2017). Despite existing studies reporting associations between training load and injury (Carey et al., 2016; Colby et al., 2014; Rogalski et al., 2013), the ability of load monitoring to predict future injury in Australian football is not well established. Studies in rugby league populations have examined the performance of training load models to predict injuries in new data (Gabbett, 2010; Thornton et al., 2016). Currently no specific modelling methodology has established superiority for accurate injury predictions. The ability of models to predict particular injury types is unknown. Yet to be explored are techniques to deal with the large imbalance between injured and non-injured observations, and how the volume of data used to build predictive models affects the quality of predictions.

This study investigates the ability of training loads to predict future injuries in elite Australian football players using statistical learning approaches. The proposed aetiology of sports injury is complex and multifactorial (Windt & Gabbett, 2017). Recent evidence suggests that training loads are a major risk factor (Soligard et al., 2016). Insights into the ability of training load models to predict injury may assist decision making by coaches and athletes, and narrow the focus of future research.

Relative and absolute training loads were included as predictor variables in this investigation to align with previous studies (Drew & Finch, 2016). Age was included as a proxy for playing experience, which has been identified as a potential moderator of the relationship between training loads and injuries (Rogalski et al., 2013). A binary indicator variable for whether a player was scheduled to play a match was also included because injury rates in Australian football are higher in matches than training sessions (Carey et al., 2016). The modelling approaches considered were informed by comparable studies in rugby league that used random forests, generalized estimating equations and logistic regression (Gabbett, 2010; Thornton et al., 2016). In particular, generalised estimating equations have been proposed as an appropriate way to model repeated observations in injury risk factor studies (Williamson, Bangdiwala, Marshall, & Waller, 1996). Support vector machines were also used to fit potentially non-linear relationships (Vapnik & Vapnik, 1998). Different data processing protocols to deal with collinear predictors and unbalanced classes were compared, and learning curves were constructed to explore how the amount of available data influenced the quality of future predictions.

Sampling strategies to combat class imbalance have not been considered in previous modelling studies. Problems with mulitcollinearity were examined by using PCA for dimensionality reduction, as has been suggested for multivariate training load analysis (Weaving, Jones, Till, Abt, & Beggs, 2017; Williams, Trewartha, Cross, Kemp, & Stokes, 2016a). Univariate and multivariate predictive models were trained on two years of player monitoring data and evaluated on one year of unseen future data in order to get a realistic estimation of predictive ability. Models were compared using the area under the receiver operator characteristic (AUC). Particular emphasis was given to the method of evaluating predictive performance given recent commentary on the misuse of the term prediction in sports science studies measuring association between risk factors and injury (McCall, Fanchini, & Coutts, 2017).

Methods

Participants

The participants involved in the study were from one professional Australian football club. The club fielded 45, 45 and 43 players in the 2014, 2015 and 2016 seasons respectively, giving 133 player seasons from 75 unique athletes. Informed consent was received from the club for collection and analysis of de-identified training and injury data. The project was approved by the La Trobe University Faculty of Health Sciences Human Ethics Committee (FHEC14/233).

In the 2016 season the club participating in this study fielded a comparatively high number of new players due to multiple season long suspensions. The impact of this on the predictive models was explored by comparing the results for new versus returning players. This enabled evaluation of the impact of introducing new players on the performance of injury models.

Data Collection

Player tracking data were collected using commercially available 10 Hz global positioning system (GPS) devices and 100 Hz triaxial accelerometers (Catapult Optimeye S5). All players were monitored during all outdoor training sessions and matches. The devices used in the study are valid for use in this athletic population (Boyd, Ball, & Aughey, 2011; Rampinini et al., 2015). Additionally, players gave a rating of perceived exertion (RPE) after each session (Foster et al., 2001). Missing data were imputed using predictive mean matching (Buuren & Groothuis-Oudshoorn, 2011).

Club medical staff recorded all injuries. Injuries were classified using the Orchard Sports Injury Classification System (OSICS) (Rae & Orchard, 2007) and were categorised as contact or non-contact. Injury severity was classified as either transient (not causing unavailability for training or matches) or time-loss (causing the player to be unavailable for regular training or match activity). Hamstring injuries were defined to include all injures when the OSICS 'specific' category was 'Hamstring strain' or 'Hamstring tendon injury'.

Training load quantification

Training loads were quantified using 5 different training load variables (Table 1). For each workload variable (w) a number of metrics were derived on each day and used as the predictor variables for injury models. The training load metrics were chosen from the existing literature associating them with injury risk.

Table 1. Workload variables and descriptions.

Training load variable	Definition
(i) Distance (m)	Distance above 3 km/h
(ii) Moderate speed running (MSR) (m)	Distance between 18-24 km/h
(iiii) High speed running (HSR) (m)	Distance above 24 km/h
(iv) Session-RPE (arbitrary units) (Foster et al., 2001)	Rating of perceived exertion multiplied by session duration
(v) Player load (arbitrary units) (Boyd et al., 2011)	Proprietary metric measuring the magnitude of rate of change of acceleration

(i) Total distance covered has been used a load metric in previous injury risk studies in Australian football (Carey et al., 2016; Colby et al., 2014; Hulin et al., 2015; Murray et al., 2016) and provides a measure of global running load.

(ii) Moderate speed running (MSR; 18-24 km/h) was included because it had previously been identified as a risk factor in Australian football (Carey et al., 2016).

(iii) High speed running (HSR; 24+ km/h) has been identified as a specific risk factor for hamstring injury as well as general non-contact injury (Colby et al., 2014; Duhig et al., 2016) and is potentially a better measure of training intensity than total distance.

(iv) Session-RPE was included as internal load metric. RPE is a subjective metric that can take into account influences such as heat stress, residual fatigue and other contextual factors that the other objective measures cannot. Additionally it has been identified as an injury risk factor in previous studies (Gabbett, 2010; Rogalski et al., 2013).

(v) Player load is a workload metric based on accelerometer data, that has the potential to capture information about activities such as tacking, jumping and collisions that the running based metrics cannot (Gabbett, 2015).

Rolling averages (C) were calculated on each day of the season (i) using a 3, 6 and 21 day accumulation window. These time periods have been identified as appropriate for quantifying cumulative load in Australian football (Carey et al., 2016; Colby et al., 2014).

$$C_i = \sum_{j=i-c}^{i-1} \frac{w_j}{c} \text{ for } c \in \{3, 6, 21\}$$

Exponentially weighted moving averages ($EWMA$) were calculated daily with 3, 6 and 21 day decay parameters (N). An exponentially weighted moving average weights the influence of training loads less the longer ago they happened. The method used was adapted from Williams et al. (Williams et al., 2016b) so that the value calculated each on day 'i' had no dependence on the training load that day (w_i). This is necessary to avoid information recorded on the day of an injury being used to try and predict that injury.

$$EWMA_i = \lambda \cdot w_{i-1} + (1-\lambda) \cdot EWMA_{i-1}$$
$$\lambda = \frac{2}{N+1} \text{ for } N \in \{3, 6, 21\}$$

Training monotony was calculated each day as the average training load in the previous 7 days divided by the standard deviation of daily loads over the same time (Foster et al., 2001). Monotony represents the variation in training done by an athlete, with higher values indicating more monotonous training. Training strain was calculated as the sum of load in the previous 7 days multiplied by the training monotony. Strain is an extension of cumulative training volume that incorporates a weighting factor based on the amount of daily variation (Foster et al., 2001). Monotony and strain were not calculated for HSR due to players frequently accumulating zero HSR load in the previous 7 days.

Daily acute:chronic workload ratios (*r*) were derived for each workload variable. The acute:chronic workload ratio represents the relative amount of short term (acute) training load compared to the long term (chronic) load. 3 and 6 day acute time windows were used with a 21 day chronic window. These choices have been identified as appropriate for discriminating between high and low risk athletes in the study cohort (Carey et al., 2016). When players had no chronic workload (and by definition zero acute load) they were assigned an acute:chronic workload ratio of zero.

$$r_i = \sum_{j=i-a}^{i-1} \frac{w_j}{a} \bigg/ \sum_{j=i-c}^{i-1} \frac{w_j}{c}$$

Exponentially weighted acute:chronic workload ratios were included as a modification of the acute:chronic workload ratio where the rolling averages were replaced by exponentially weighted moving averages (Murray et al., 2016; Williams et al., 2016b). Murray et al. suggested that the exponentially weighted acute:chronic workload ratio gave a better indicator of injury risk than the rolling average method (Murray et al., 2016).

In total, for each of the 5 training load metrics (Table 1), 12 features were derived; 3, 6 and 21-day rolling and exponentially weighted average load. ACWR and exponentially weighted ACWR (using 3:21 and 6:21 day windows), as well as monotony and strain (excluding HSR), resulting in 58 workload features. The inclusion of a large number of features reflected the multiple injury risk factor findings from previous studies (Carey et al., 2016; Colby et al., 2014; Duhig et al., 2016; Hulin et al., 2015; Malone et al., 2016; Murray et al., 2016; Rogalski et al., 2013). There was no a-priori justification to exclude any. The size of the feature space and the possible correlations between predictor variables was given particular consideration when choosing the modelling and data processing approaches.

Injury classification

Three different types of injury were considered; any non-contact (NC), non-contact causing time-loss (NCTL) and any hamstring (HS). Separate models were built for each injury outcome to investigate whether predictive models performed better for specific injury types. Hamstring injuries were chosen as they were the most frequently occurring specific injury in this cohort and are a common injury in Australian football (Orchard, Seward, & Orchard, 2013). A possible lag period between spikes in training load and increased injury risk has been reported in previous studies (Hulin et al., 2015). Models were also built for the likelihood of a player sustaining an injury in the next five days to investigate whether including a lag period could improve predictive performance.

Modelling approach

Predictive models were built on two years of load and injury data (model training data) and tested on one season of unseen future data (testing data) (Figure 1). Evaluating models on a season of unseen data is required to get an estimate of the ability to predict injuries. Multivariate statistical models can have many degrees of freedom and can be tuned to fit a particular data set very well. To test whether a model will generalise and be useful in practice, the predictions must be evaluated on a new data set (Kuhn & Johnson, 2013).

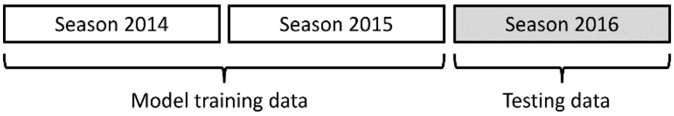


Figure 1. Data split for model training and testing.

Models were built to try and predict whether an athlete would be injured, given knowledge of their training loads. Each day that a player completed training or a match, or reported an injury, was included as an observation in the model training data. Predictions were then generated for each training or match day in the testing data set and evaluated against actual injury incidence.

Prediction algorithms

Multiple algorithms were tested to compare their ability to predict injury in team sport athletes. Multivariate models were constructed using all 58 training load variables as well as two additional features; player age (years) and a binary indicator variable for whether or not the player was scheduled to play a in a competitive match that day. The approaches considered were:

- Logistic regression (LR) is commonly used to model injury outcomes (Colby et al., 2014; Gabbett, 2010; Murray et al., 2016). Elastic net regularisation was introduced due to the large number of predictor variables used (Zou & Hastie, 2005).
- Random forest (RF) models were chosen for their ability to fit non-linear patterns in data and deal with collinear predictors (Breiman, 2001). Random forests have been used in injury prediction studies in rugby league (Thornton et al., 2016).
- Generalised estimating equations (GEE) are an extension of generalised linear models that account for correlations between repeated observations taken from the same subjects (Liang & Zeger, 1986). A binomial link function and auto-regressive correlation structure was used (Williamson et al., 1996).
- Support vector machines (SVM) with a radial basis function were used to model the potentially non-linear pattern between load and injury in high dimensional data (Vapnik & Vapnik, 1998).

In addition, univariate LR models were constructed for each training load variable (Colby et al., 2014; Gabbett, 2010; Murray et al., 2016) to provide a comparison for the more complex multivariate and non-linear models used.

Models hyperparameters (elastic net mixing parameter and regularisation strength for LR, number of trees and variables sampled at each node for RF and regularisation penalty for SVM) were tuned using 10-fold cross validation on the model training data and a grid search method (see supplementary table 1 for details). The choice of hyperparameter that gave the best performance (area under ROC curve) during cross validation was then used to construct a model on the full training data set (2014-15). This model was then tested on the hold-out season (2016). All analyses were performed using the R statistical programming language.

Data pre-processing

To allow for players to accumulate sufficient training loads to calculate workload ratios and exponentially weighted moving averages, the first 14 days of each season were removed from the model training and testing data. This loss of data could potentially be avoided in future studies if monitoring of chronic workloads extended into the off-season period. Players in rehabilitation training were excluded from modelling until they returned to full training.

Many of the training load variables collected were likely to be correlated (Weaving et al., 2017). Thus our prediction problem may suffer from multi-collinearity, potentially leading to instability and errors in the model building process (Kuhn & Johnson, 2013). Principal component analysis (PCA) is a dimensionality reduction process that reduces a large number of predictor variables to a smaller number of uncorrelated variables (called principal components) to combat the problems associated with multi-collinearity (Kuhn & Johnson, 2013). PCA has been advocated as a way of dealing with collinearity in multivariate training load modelling (Weaving et al., 2017). It has been employed in previous studies of training load monitoring (Weaving, Marshall, Earle, Nevill, & Abt, 2014; Williams et al., 2016a). To explore the effects of PCA pre-processing, each multivariate model was trained with unprocessed data and data pre-processed with PCA and the results were compared. Principal components were calculated using the singular value decomposition method (Jolliffe, 1986). A 95% cumulative variance threshold was used to extract m principal components, where m was the smallest number of components that explained at least 95% of the total variation in the data (Jolliffe, 1986; Kuhn & Johnson, 2013).

Class imbalance refers to prediction problems when one class is far more common than the other (Kuhn & Johnson, 2013). Injury prediction suffers from large class imbalance since injuries are far less common than days when a player doesn't get injured. Severe class imbalance can cause prediction algorithms to have trouble correctly predicting the rare class (Kuhn & Johnson, 2013). Two sampling techniques were implemented to combat class imbalance. Under-sampling randomly removes non-injury days from the model building data until there is an equal number of injury and non-injury days. Synthetic minority over-sampling (SMOTE) synthetically creates new injury samples in the model training data and under-samples a fraction of the non-injury days to even up the classes (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). Each model was built using unprocessed data, under-sampled data and SMOTE sampled data.

Model evaluation

To evaluate the performance of each modelling approach, predicted injury probabilities were generated for each training or match day in the full testing data set. A receiver operator characteristic curve was constructed and the AUC calculated. Importantly, each model was evaluated on exactly the same data set (i.e. no under-sampling or SMOTE applied to testing data) to allow for fair comparisons. A perfect model would have AUC=1.0 and random guessing would be expected to produce AUC=0.5. This gives a performance metric preferable to error rate for problems with unbalanced data; where low error can be achieved by simply predicting the more common class every time (Kuhn & Johnson, 2013; Thornton et al., 2016) (e.g. if the injury rate is 1%, a model always predicting no-injury is 99% accurate).

To estimate the variability introduced into the modelling procedure by randomly sampling the data during the pre-processing and model tuning stages, 50 repetitions of the entire process were run, generating a set of performance estimates for each model. Under-sampling and SMOTE sampling introduce variability into the modelling pipeline because they select a random subset of the training data before the model building stage. A different selection may result in a different model. Variability is also introduced during the model tuning stage. During cross validation the training data set is randomly partitioned into 10 folds, model hyperparameters are then chosen based on the performance across these folds. It is possible that when the process is repeated the composition of the folds will change. In this circumstance a different hyperparameter will be chosen and a different model created.

Results

The number of reported injuries (and frequency relative to the number of sessions) is shown in Table 2. Hamstring injury rates were similar across seasons (0.004 vs. 0.003 injuries per session), however non-contact (0.035 vs. 0.014 injuries per session) and non-contact time-loss (0.017 vs. 0.009 injuries per session) injury rates were lower.

Table 2: Injury counts and rates (relative to total number of sessions) in the model training and testing data. (NC = non-contact, NCTL = non-contact time-loss, HS = hamstring, 'lag' suffix indicates outcome is injury within 5 days).

Injury outcome	Model training data (2014 & 2015)	Testing data (2016)
NC	321 (0.035)	67 (0.014)
NCTL	156 (0.017)	42 (0.009)
HS	36 (0.004)	13 (0.003)
NC-lag	1601 (0.174)	479 (0.103)
NCTL-lag	784 (0.085)	295 (0.063)
HS-lag	183 (0.020)	88 (0.019)
Total records (match & training)	9203	4664

Predictive ability of different modelling approaches

Predictive performance was limited for multivariate models when using un-processed data (Figure 2). Using regularised LR to model hamstring injuries performed best (mean AUC=0.72), all other multivariate models had a mean AUC of less than 0.65 on the testing data. Univariate models performed worse than multivariate models for each injury outcome (best AUC<0.6 for NC and NCTL, and best AUC<0.7 for HS).

Attempting to account for a possible lag time between training load spikes and injury occurrence (NC lag, NCTL lag and HS lag in Figure 2) did not improve predictive performance (mean AUC=0.50-0.57). In general, models provided predictions only marginally better than chance.

The performance of HS injury models (particularly LR and RF) was more variable than non-specific injury models (Figure 2). The increased variability is likely due to the smaller number of HS injuries in the model training data (n=36). The random inclusion or exclusion of one or more of these injuries during the model building stage has a greater impact on the final model.

No particular prediction algorithm showed a strong tendency to outperform others across different injury outcomes. The more complex models (RF and SVM) did not tend to outperform generalised linear models (LR). Accounting for individual clustering effects (GEE) did not lead to better results.

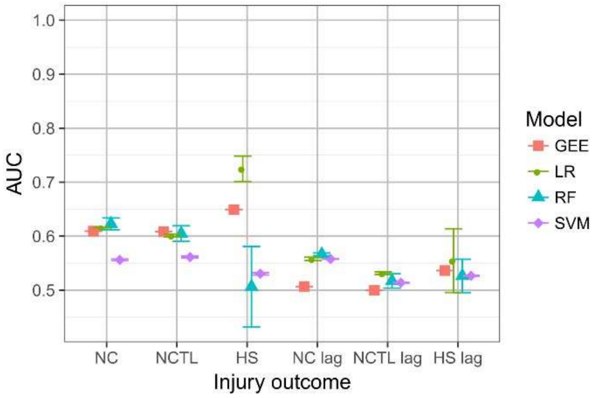


Figure 2: Area under ROC curve evaluated on the testing data set (mean and standard deviation of 50 repetitions) for different prediction algorithms and injury outcomes (no data pre-processing). (NC = non-contact, NCTL = non-contact time-loss, HS = hamstring, 'lag' suffix indicates outcome is injury within 5 days. GEE = Generalised estimating equation, LR = regularized logistic regression, RF = random forest, SVM = support vector machine).

Hamstring injury prediction

Regularised LR for hamstring injuries showed better performance (mean AUC=0.72) than other models (Figure 2). The model gives a predicted injury probability each day, how this translates into practice (i.e. modifying player training and preventing injuries) is dependent on the preferences of the decision maker. Specifically, how they weight the consequences (cost) of a false negative (missed injury) relative to a false positive prediction (unnecessarily cancelling or modifying a session by decreasing volume or intensity). To illustrate this we considered three arbitrary relative costs: 1 missed injury (i.e. one we didn't predict) costs as much as 50 unnecessarily modified sessions (aggressive risk), as much as 100 sessions (moderate risk), or as much as 1000 sessions (conservative risk). The estimated optimal performance metrics for the three choices are shown in Table 3. There is a trade-off between correctly predicting more injuries and incorrectly flagging non-injury sessions. In general, the AUC values for other models in the study were similar or below the HS model, suggesting they would not perform significantly better than the results in Table 3.

Table 3: Estimated optimal performance of HS injury models for different relative cost ratios (values reported as median of 50 repetitions).

Cost of false negative relative to false positive	**True positive rate**	**False positive rate**	**Positive likelihood ratio**	**Negative likelihood ratio**	**Probability injury given positive prediction**	**Probability injury given negative prediction**
50 (aggressive risk)	0.08	0.004	17.9	0.93	0.05	0.003
100 (moderate risk)	0.54	0.11	5.0	0.52	0.01	0.001
1000 (conservative risk)	0.92	0.53	1.7	0.16	0.005	0.0004

Effect of data pre-processing

The effects of different data pre-processing protocols are shown in Figure 3. Model performance varied under different protocols, yet the differences in predictive ability were generally small. Reducing the number of predictors to a smaller, uncorrelated set by applying PCA pre-processing caused minor performance improvements in the models considered (Figure 3). This suggested that multicollinearity was a potential cause of poor performance when using un-processed data. Additionally, the variability in performance tended to decrease (especially for RF models).

Under-sampling non-injury days led to performance decreases for all models except the SVM (Figure 3). This is possibly due to the information lost from the model training data when a large number of the non-injury days are removed. Under-sampling may not be appropriate for the injury prediction problem. SMOTE sampling did not lead to any major performance improvements (Figure 3). In the SMOTE procedure new injury observations were synthetically created using the common characteristics observed in actual injuries. This may not help the generalisability of models if new injuries show little resemblance to past ones. SVM models were the exception, and benefited from both sampling methods used, suggesting their performance was more negatively affected by imbalance between injury and non-injury observations. Combining SMOTE sampling and PCA pre-processing was similarly unsuccessful in improving predictive performance.

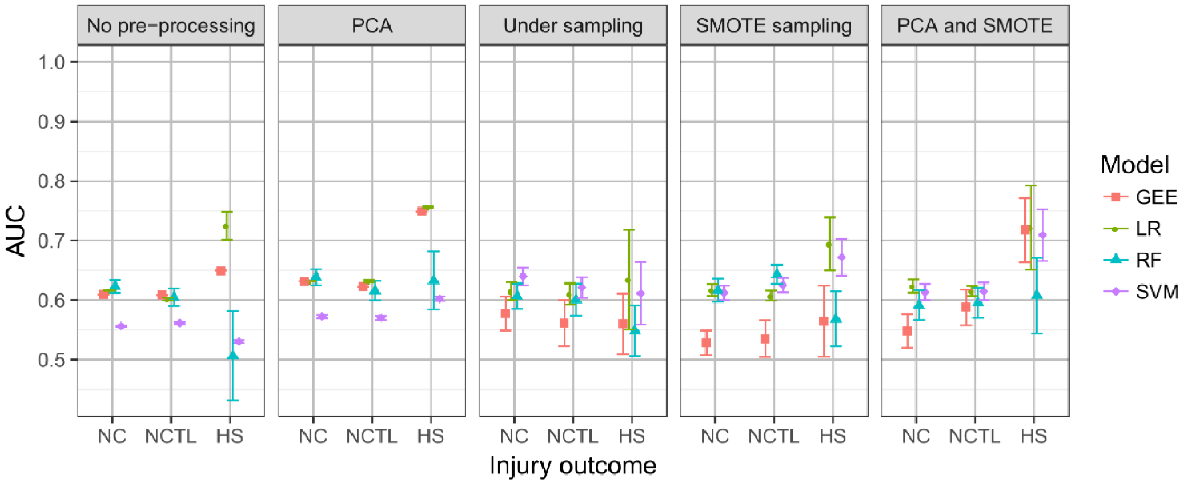


Figure 3: Effects of data pre-processing protocols and sampling methods on model performance (mean and standard deviation of 50 repetitions) for each injury outcome (NC = non-contact, NCTL = non-contact time-loss, HS = hamstring. GEE = Generalised estimating equation, LR = regularized logistic regression, RF = random forest, SVM = support vector machine).

Applying sampling methods to the data to try and reduce the amount of imbalance between the number of injured and non-injured observations led to increased variability in the results (Figure 3). This is a consequence of randomly removing different subsets of the data before building models in each simulation. ROC curves for the best performing models (highest mean AUC) for each injury type are shown in Figure 4. ROC curves for the modelling procedures that included under-sampling and SMOTE sampling of the training data showed much more variability between the 50 repetitions (Figure 4(a-b)). The best performing hamstring model (Figure 4(c)), which didn’t use any sampling methods, showed only very small variability between repetitions (nearly perfect overlap for each repetition).

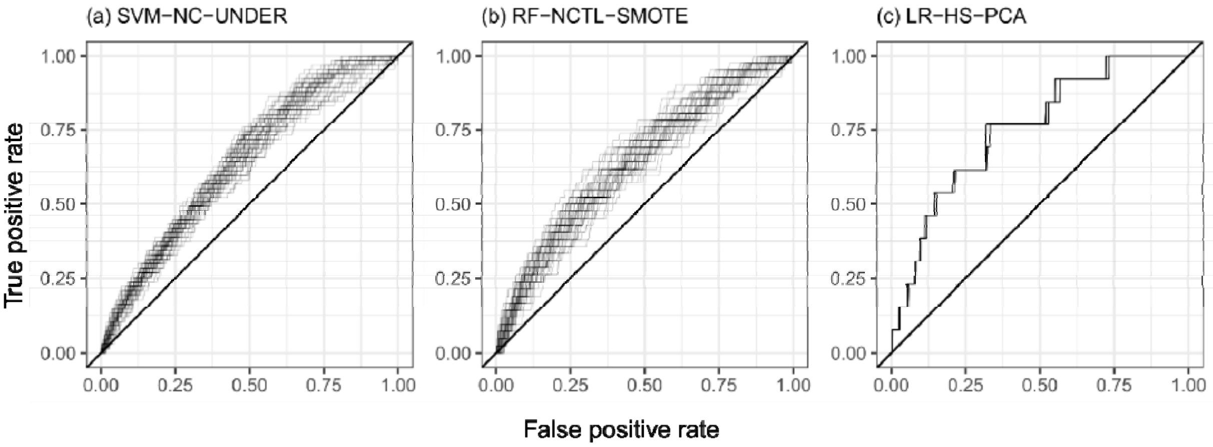


Figure 4: ROC curves for 50 repetitions of the best performing (highest mean AUC) modelling process for each injury type: (a) SVM with under-sampling for NC, (b) RF with SMOTE sampling for NCTL and (c) LR with PCA for HS injuries (NC = non-contact, NCTL = non-contact time-loss, HS = hamstring. LR = regularized logistic regression, RF = random forest, SVM = support vector machine).

Principal component analysis

PCA pre-processing applied to the full training data set (i.e. before any sampling strategies were applied) required 14 principal components to explain 95% of the variance in the data (Figure 5(a)). The first two components accounted for close to 60% of the variance. This implied that much of the information contained in the multiple training load metrics could be captured in a lower dimension. The correlation between the predictor variables and the first two principal components is shown in Figure 5(b). A clear grouping effect between variables of the same types was observed. The relative training load variables (ACWR and exponentially weighted ACWR) were related to the principal components in similar ways. The grouping could also be observed for 21-day cumulative loads and 3-day cumulative loads. Separation between monotony, strain (calculated over 7 days) and 6-day cumulative load variables was less clear. This was potentially due to their similar timescales. The two non-training related variables, age and match indicator, were not strongly correlated with either of the first two principal components.

Model performance for new versus returning players

Predictive accuracy of NC and NCTL injury models did not significantly change when new players and returning players were considered separately (Figure 6). Hamstring models appeared to perform better on returning players, suggesting they may have started to identify patterns leading to hamstring injury in the existing playing group. However, of the 13 total hamstring injuries in the testing data; 10 were to new players and only 3 to returning players. The inflated results for the returning players may be a consequence of the small sample size and not truly representative of expected future performance.

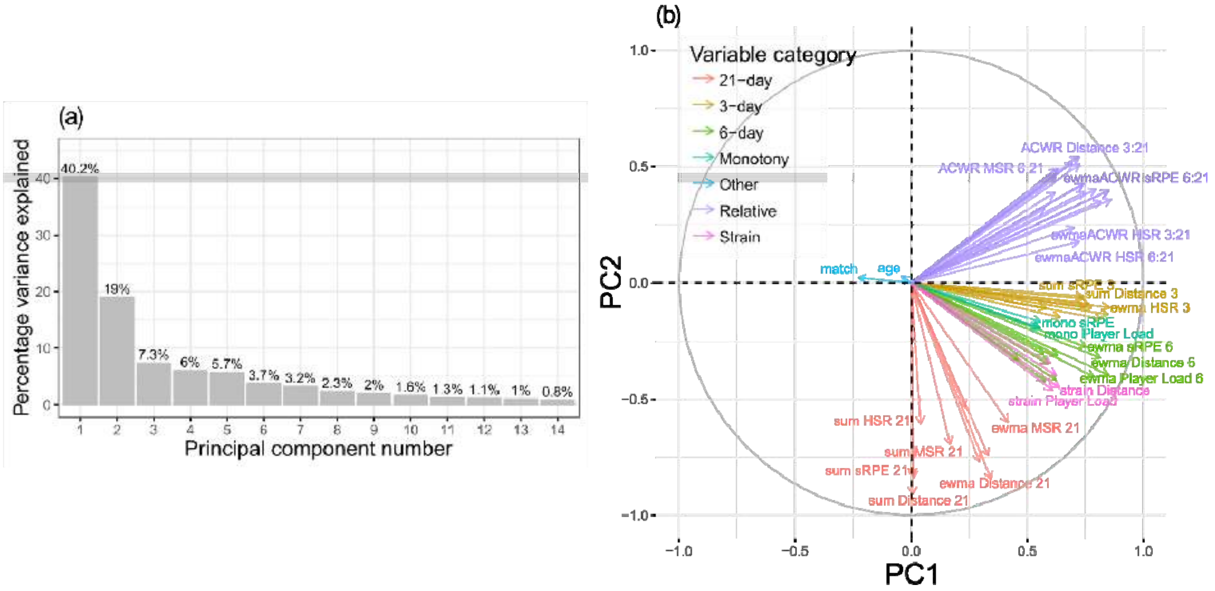


Figure 5: Results of principal component analysis applied to full training data set. (a) Scree plot showing the percentage variance explained by the first 14 principal components (the number required to pass 95% cumulative variance explained). (b) Variable factor map showing how correlated each predictor variable is with the first two principal components (a representative subset of variable labels were chosen for figure clarity).

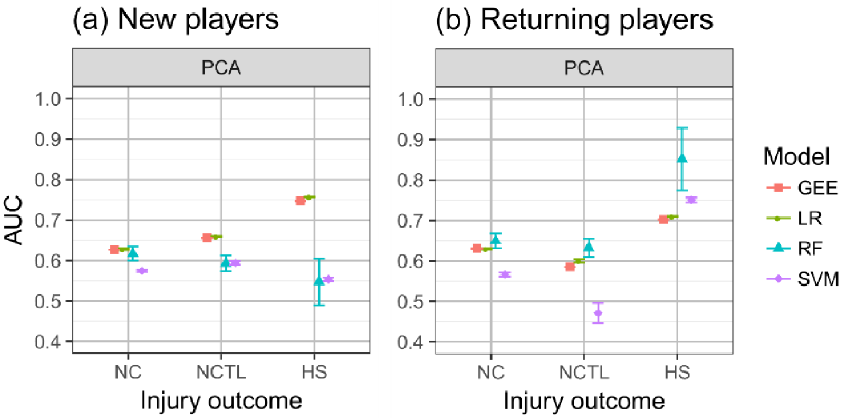


Figure 6: Model performance (mean and standard deviation of 50 repetitions) for (a) new players and (b) returning players (PCA pre-processing).

Effect of increasing the number of model training samples

The performance of predictive models can be influenced by the amount of data available to build models (Kuhn & Johnson, 2013). A learning curve shows the performance of a model on the training and testing data sets as the size of the training data set is increased. This indicates whether performance is improving or plateauing as more data is used. Learning curves were constructed for two injury models (Figure 7) to investigate whether the poor performance could be attributed to insufficient amounts of data.

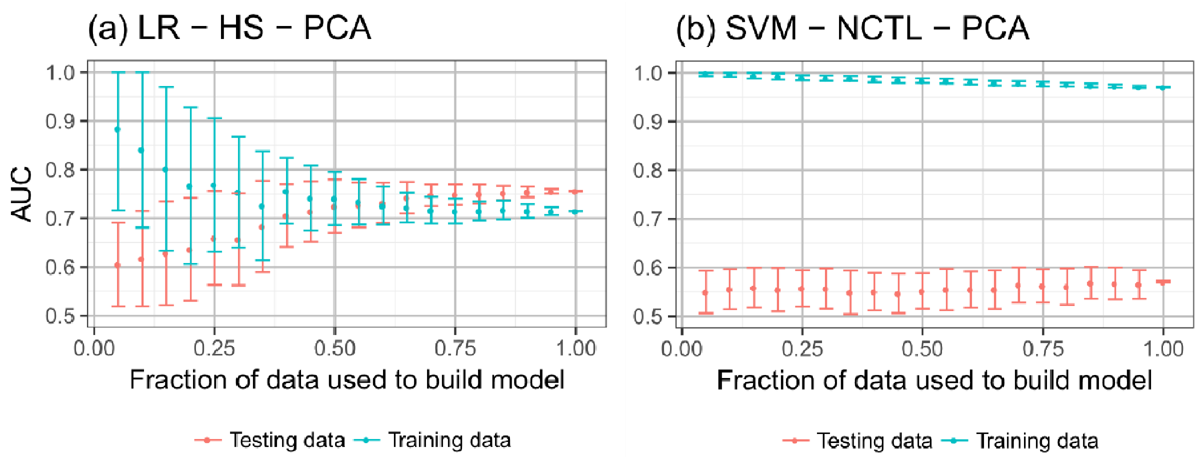


Figure 7: Learning curves showing mean and standard deviation of performance for; (a) LR model of HS injuries (b) SVM model of NCTL injuries (both with PCA pre-processing). (NCTL = non-contact time-loss, HS = hamstring. LR = regularized logistic regression, SVM = support vector machine)

The performance of LR to predict hamstring injury (with PCA) is shown in Figure 7(a). Test set performance increased as the amount of data used the build the model increased, suggesting the model was learning to recognise hamstrings injuries better. The performance on the training and testing sets appeared to converge to a similar level (mean AUC=0.7-0.8) once the full data set was used. However this represented a limited ability to predict injuries (Table 3); meaning the model was unable to fully explain the relationship between the predictor and outcome variables, or had underfit the problem. Underfitting suggested the model was unable to capture enough complexity in the data or that the predictor variables did not contain enough information to predict injuries (Kuhn & Johnson, 2013).

Figure 7(b) shows the learning curve for an SVM model for NCTL injuries (with PCA). The performance on the model training data was near perfect but test set performance was well below desired levels. The learning curve suggests the model may be suffering from overfitting; it has perfectly fit the injuries in the model training data but does not generalise well to new data. Potential strategies for addressing this are collecting more data (especially more positive injury samples) or penalising the model for increasing complexity (regularisation) (Kuhn & Johnson, 2013). The increase in performance observed when the size of the training data set was increased from 460 to 9203 samples was approximately 5%. Estimating further improvements is difficult given the potentially non-linear relationship, however it appears at least an order of magnitude (tenfold) more data may be needed.

Discussion

Models of the relationships between training loads and injuries in elite Australian footballers showed limited ability to predict future injuries. Multivariate models of non-contact and non-contact time-loss injuries performed better than univariate models, yet only marginally better than would be expected by random chance (mean AUC<0.7). The levels of performance were comparable to similar modelling in rugby league (AUC=0.64-0.74) (Thornton et al., 2016).

Considering hamstring injuries on their own led to improvements in model performance (best AUC=0.76) (Figure 2-4). Implementing such a model in practice would require practitioners to consider how much modification of player training they are willing to accept in an attempt to prevent injuries. Results suggested that predicting half of the hamstring injuries would incur a false positive rate above 10% (or more than 1 in 10 player sessions unnecessarily modified) (Table 3). The multivariate models used in this study provided improved predictive ability compared to findings by Ruddy et. al. (Ruddy et al., 2016) in a larger cohort of hamstring injuries in Australian football (AUC=0.5-0.63).

Pre-processing data to account for multicollinearity of predictors (with PCA) and sampling to reduce the level of class imbalance resulted in minor improvements to predictive performance (Figure 3). Particularly in the more complex models (SVM) that tended to over-fit the model training data. Consideration of these issues may improve predictive performance in future modelling studies.

Results from the PCA analysis showed the different training load variables to strongly correlated (Figure 5(b)). A large proportion of the total variance in the 60 variable data set could be captured by only a few principal components (Figure 5(a)). The metrics used in this study were originally chosen in an attempt to represent different physiological demands imposed on players. The results suggested that in Australian football the different metrics contain largely overlapping information. This finding highlights issues with performing multivariate training load modelling without initially treating the data to extract orthogonal components (Weaving et al., 2017). The amount of redundancy in the predictor variables may also partially explain the inability to predict injuries with high accuracy. The inclusion of multiple GPS and accelerometer derived metrics provided little additional information and future studies may benefit from including different sources of information in modelling approaches (e.g. anthropomorphic measurements, injury histories or psychological profiles). The internal training load variable (sRPE) did not appear to provide different information to external loads (Figure 5). The strong correlations between different training load metrics may be a consequence of the particular demands of training and matches in Australian football, and it is not known if this finding generalizes to other athletic cohorts.

Possibilities for improving injury prediction models

The learning curves for different modelling approaches are shown in Figure 7. These provide some indicators for ways to potentially improve the performance of future injury prediction models. Underfitting models (Figure 7(a)) can be improved by increasing the model complexity or by increasing the number and variety of predictors (Kuhn & Johnson, 2013). This suggests that linear models such as logistic regression may not be well suited to modelling complex injury relationships, supporting the contention of Bittencourt et. al. (Bittencourt et al., 2016). The more complex models (RF and SVM) tended to over-fit the small number of injuries in the data (Figure 7(b)) and did not generalise well to future injury observations. Collecting more player monitoring and injury data (>10 team-seasons) may provide a way to construct practically useful prediction models.

Given the low injury per session rate (0.4% for hamstring to 3% for all non-contact) and the large number of possible risk factors, a probabilistic approach to load and injury risk models that allows for clinical judgement given context and other sources of information may improve outcomes. Future studies may see progress by taking this approach and not considering injury prediction as a binary classification problem.

Utility of training load monitoring for injury risk reduction

Results of this study suggested that training loads models were limited in their ability predict future injuries for an athlete on a given day. However this not does not rule out training load monitoring as a valid practice. There is strong evidence (Blanch & Gabbett, 2016; Carey et al., 2016; Hulin et al., 2015; Malone et al., 2016; Murray et al., 2016; Soligard et al., 2016) that rapid increases in load and spikes in acute:chronic workload ratio are associated with increases in team injury rates. For this reason, measuring absolute and relative training loads in team sports to monitor load progression and allow for informed modification of plans is still considered best practice (Drew & Finch, 2016; Soligard et al., 2016). Using individual training loads as a daily decision tool for athlete injury prediction has limited utility in this study.

Limitations

The professional sporting team participating in this study had comparatively high player turnover. This may have impacted on the accuracy of injury prediction models (Figure 6) and restricted the ability to build player specific models. The small injury sample size was a consequence of conducting this study within a single club. There were an insufficient number of injury records of a particular type (other than hamstrings) to create different injury specific models.

This study included a number of the commonly used training load measures (GPS and session-RPE) and derived metrics (cumulative load, acute:chronic workload ratio, monotony and strain). Running demands in relative speed zones, subjective wellness and fatigue markers and biomechanical screening data were not available. These variables may contain relevant information to improve predictive models.

In both the model training and testing seasons physical preparation staff were aware of the emerging body of research on training loads and injury risk (Colby et al., 2014; Gabbett, 2010, 2016; Hulin et al., 2015; Rogalski et al., 2013) and gave consideration to this when planning training. This may have influenced the distribution of training load variables recorded.

Practical applications

The results of this study highlighted the limitations of using training load based predictive models as a daily decision tool for practitioners in Australian football. We outlined how a practitioner’s judgement on the “cost” of an injury relative to a modified training session can influence how a predictive model is implemented in practice (Table 3). Improved predictive performance for hamstring specific models suggests that modelling specific injury outcomes instead of general non-contact injuries may be more informative for assessing injury risk in practice. The findings relating to insufficient data set size and poor model performance address an issue likely encountered by practitioners operating in a team-sport environment.

CONCLUSION

Models of training load, age and session type showed limited ability to predict future injury in Australian footballers. Hamstring specific injury models showed better predictive performance than general non-contact models. Future studies may benefit by considering specific injury types as outcomes variables. Predictive performance also improved with increasing quantity of data. This highlighted the limitations of predictive modelling attempts conducted within a single team and the need for collaboration or larger cohorts. Training load may be an important injury risk factor to monitor, yet considering it in isolation as a daily decision tool has limited ability to predict injury.

REFERENCES

Bittencourt, N., Meeuwisse, W., Mendonça, L., Nettel-Aguirre, A., Ocarino, J., & Fonseca, S. (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*(21), 1309-1314.

Blanch, P., & Gabbett, T. J. (2016). 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. *British Journal of Sports Medicine, 50*(8), 471-475.

Bourdon, P. C., Cardinale, M., Murray, A., Gastin, P., Kellmann, M., Varley, M. C., . . . Cable, N. T. (2017). Monitoring Athlete Training Loads: Consensus Statement. *Int J Sports Physiol Perform, 12*(Suppl 2), S2161-S2170. doi:10.1123/IJSPP.2017-0208

Boyd, L. J., Ball, K., & Aughey, R. J. (2011). The reliability of MinimaxX accelerometers for measuring physical activity in Australian football. *International Journal of Sports Physiology and Performance, 6*(3), 311-321.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software, 45*(3).

Carey, D. L., Blanch, P., Ong, K.-L., Crossley, K. M., Crow, J., & Morris, M. E. (2016). Training loads and injury risk in Australian football—differing acute: chronic workload ratios influence match injury risk. *British Journal of Sports Medicine*, bjsports-2016-096309. doi:10.1136/bjsports-2016-096309

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research, 16*, 321-357.

Colby, M. J., Dawson, B., Heasman, J., Rogalski, B., & Gabbett, T. J. (2014). Accelerometer and GPS-derived running loads and injury risk in elite Australian footballers. *The Journal of Strength & Conditioning Research, 28*, 2244–2252.

Drew, M. K., & Finch, C. F. (2016). The Relationship Between Training Load and Injury, Illness and Soreness: A Systematic and Literature Review. *Sports Medicine*. doi:10.1007/s40279-015-0459-8

Duhig, S., Shield, A. J., Opar, D., Gabbett, T. J., Ferguson, C., & Williams, M. (2016). Effect of high-speed running on hamstring strain injury risk. *Br J Sports Med, 50*(24), 1536-1540. doi:10.1136/bjsports-2015-095679

Foster, C., Florhaug, J. A., Franklin, J., Gottschall, L., Hrovatin, L. A., Parker, S., . . . Dodge, C. (2001). A new approach to monitoring exercise training. *The Journal of Strength & Conditioning Research, 15*, 109–115.

Gabbett, T. J. (2010). The development and application of an injury prediction model for noncontact, soft-tissue injuries in elite collision sport athletes. *The Journal of Strength & Conditioning Research, 24*, 2593–2603.

Gabbett, T. J. (2015). Relationship Between Accelerometer Load, Collisions, and Repeated High-Intensity Effort Activity in Rugby League Players. *J Strength Cond Res, 29*(12), 3424-3431. doi:10.1519/JSC.0000000000001017

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

Hägglund, M., Waldén, M., Magnusson, H., Kristenson, K., Bengtsson, H., & Ekstrand, J. (2013). Injuries affect team performance negatively in professional football: an 11-year follow-up of the UEFA Champions League injury study. *British Journal of Sports Medicine, 47*(12), 738-742.

Hulin, B. T., Gabbett, T. J., Lawson, D. W., Caputi, P., & Sampson, J. A. (2015). The acute:chronic workload ratio predicts injury: high chronic workload may decrease injury risk in elite rugby league players. *British Journal of Sports Medicine*, bjsports-2015-094817. doi:10.1136/bjsports-2015-094817

Jolliffe, I. T. (1986). Principal Component Analysis and Factor Analysis *Principal component analysis* (pp. 115-128): Springer.

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. New York, NY: Springer New York.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*(1), 13-22.

Malone, S., Owen, A., Newton, M., Mendes, B., Collins, K. D., & Gabbett, T. J. (2016). The acute:chonic workload ratio in relation to injury risk in professional soccer. *Journal of Science and Medicine in Sport*. doi:10.1016/j.jsams.2016.10.014

McCall, A., Fanchini, M., & Coutts, A. J. (2017). Prediction: The Modern-Day Sport-Science and Sports-Medicine "Quest for the Holy Grail". *Int J Sports Physiol Perform, 12*(5), 704-706. doi:10.1123/ijspp.2017-0137

Murray, N. B., Gabbett, T. J., Townshend, A. D., & Blanch, P. (2016). Calculating acute:chronic workload ratios using exponentially weighted moving averages provides a more sensitive indicator of injury likelihood than rolling averages. *British Journal of Sports Medicine*, bjsports-2016-097152. doi:10.1136/bjsports-2016-097152

Orchard, J. W., Seward, H., & Orchard, J. J. (2013). Results of 2 decades of injury surveillance and public release of data in the Australian Football League. *The American Journal of Sports Medicine*, 0363546513476270.

Rae, K., & Orchard, J. (2007). The Orchard sports injury classification system (OSICS) version 10. *Clinical Journal of Sport Medicine, 17*(3), 201-204.

Rampinini, E., Alberti, G., Fiorenza, M., Riggio, M., Sassi, R., Borges, T. O., & Coutts, A. J. (2015). Accuracy of GPS devices for measuring high-intensity running in field-based team sports. *International Journal of Sports Medicine, 36*(1), 49-53.

Rogalski, B., Dawson, B., Heasman, J., & Gabbett, T. J. (2013). Training and game loads and injury risk in elite Australian footballers. *Journal of Science and Medicine in Sport, 16*, 499-503. doi:10.1016/j.jsams.2012.12.004

Ruddy, J. D., Pollard, C. W., Timmins, R. G., Williams, M. D., Shield, A. J., & Opar, D. A. (2016). Running exposure is associated with the risk of hamstring strain injury in elite Australian footballers. *British Journal of Sports Medicine*, bjsports-2016-096777. doi:10.1136/bjsports-2016-096777

Soligard, T., Schwellnus, M., Alonso, J.-M., Bahr, R., Clarsen, B., Dijkstra, H. P., . . . Engebretsen, L. (2016). How much is too much? (Part 1) International Olympic Committee consensus statement on load in sport and risk of injury. *British Journal of Sports Medicine, 50*, 1030-1041. doi:10.1136/bjsports-2016-096581

Thornton, H. R., Delaney, J. A., Duthie, G. M., & Dascombe, B. J. (2016). Importance of Various Training Load Measures on Injury Incidence of Professional Rugby League Athletes. *International Journal of Sports Physiology and Performance*, 1-17.

Vapnik, V. N., & Vapnik, V. (1998). *Statistical learning theory* (Vol. 1): Wiley New York.

Weaving, D., Jones, B., Till, K., Abt, G., & Beggs, C. (2017). The Case for Adopting a Multivariate Approach to Optimize Training Load Quantification in Team Sports. *Front Physiol, 8*, 1024. doi:10.3389/fphys.2017.01024

Weaving, D., Marshall, P., Earle, K., Nevill, A., & Abt, G. (2014). Combining internal- and external-training-load measures in professional rugby league. *Int J Sports Physiol Perform, 9*(6), 905-912. doi:10.1123/ijspp.2013-0444

Williams, S., Trewartha, G., Cross, M., Kemp, S., & Stokes, K. (2016a). Monitoring What Matters: A Systematic Process for Selecting Training Load Measures. *International Journal of Sports Physiology and Performance*, 1-20.

Williams, S., West, S., Cross, M., & Stokes, K. (2016b). Better way to determine the acute:chronic workload ratio? *British Journal of Sports Medicine*, bjsports-2016-096589. doi:10.1136/bjsports-2016-096589

Williamson, D. S., Bangdiwala, S. I., Marshall, S. W., & Waller, A. E. (1996). Repeated measures analysis of binary outcomes: applications to injury research. *Accident Analysis & Prevention, 28*(5), 571-579.

Windt, J., & Gabbett, T. J. (2017). How do training and competition workloads relate to injury? The workload-injury aetiology model. *Br J Sports Med, 51*(5), 428-435. doi:10.1136/bjsports-2016-096040

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301-320.