Title page

Authors and affiliation

Fraser Philp[1,2] ORCID ID https://orcid.org/0000-0002-8552-7869

Ahmad Al-Shallawi [1,2]

Theocharis Kyriacou [3]

Dimitra Blana [2]

Anand D Pandyan[1,2]

[1] School of Health and Rehabilitation, Keele University, UK

[2] Institute of Science and Technology in Medicine, Keele University, UK

[3] School of Computing and Mathematics, Keele University, UK

Corresponding author's details

Dr Fraser Philp

f.d.philp@keele.ac.uk

School of Health and Rehabilitation, MacKay Building, Keele University, Keele, ST5 5BG

Title: Improving predictor selection for injury modelling methods in male footballers.

Short title: Improving predictor selection in football injury modelling.

Abstract

This study evaluated whether combining existing methods of Elastic net for zero-inflated Poisson and zero-inflated Poisson regression methods could improve real life applicability of injury prediction models in football. Predictor selection and model development was conducted on a pre-existing dataset, from a single English football teams' 2015/2016 season. The Elastic Net for zero-inflated Poisson penalty method was successful shrinking the total number of predictors in the presence of high levels of multicollinearity. It was additionally identified that easily measurable data, i.e. mass and body fat content, training type, duration and surface, fitness levels, normalised period of "no-play" and time in competition could contribute to the probability of acquiring a time loss injury. Further, prolonged series of match play and increased in-season injury reduced the probability of not sustaining an injury. For predictor selection, the Elastic net for zero-inflated Poisson penalised method in combination with the use of ZIP regression modelling for predicting time loss injuries have been identified appropriate methods for improving real life applicability of injury prediction models. These methods are more appropriate for datasets subject to multicollinearity, smaller sample sizes and zero-inflation known to affect the performance of traditional statistical methods. Further validation work is now required.

Keywords

Football, injury prediction, modelling, predictor selection, validation, variable selection,

Contributorship statement: All authors in this study have been involved in the planning, conduct, and reporting of the work described in the article. All authors have seen an approved the final draft of this article.

Acknowledgements: NIL

Competing interests: NIL

Funding details: No funding was received for this study

Ethical approval information: Granted by the Keele University Ethical Review Panel (ERP1237)

Disclosure statement: None of the authors have any financial interests or benefits that have arisen from the direct application of their research.

Data availability statement: The supporting dataset and code will be uploaded to a recognised data repository or made available once accepted

1. Introduction

Statistical models for injury prediction lack clinical applicability and have not been routinely adopted for use in clinical practice. Several predictor selection and modelling methods have been advocated for prospective injury modelling, including clinical movement scales (Kiesel et al., 2007), laboratory based algorithms (Myer et al., 2011) and statistical models (Bahr and Holme, 2003, Hagglund et al., 2005, Hägglund et al., 2006, Venturelli et al., 2011). Within football, injury reporting, recording(Fuller et al., 2006) and predictor selection methods are informed by existing frameworks which advocate the use of multivariate statistical models (Bahr and Holme, 2003, Hagglund et al., 2005). Multivariate modelling, at the level of an individual club, is likely to have little clinical value as these methods tend to require large sample sizes or expensive and complex measurements which are not easily attainable. Furthermore, existing models for injury prediction have been developed using posteriori datasets, i.e. the injury outcome is already known and associations between the variables and the known outcome is estimated.

These models have limited clinical applicability for the following reasons (1) the models are often 'black-boxes' that provide no physiological explanation for the predictor variables, sports and exercise medicine practitioners may have an inherent distrust of complex models, in which the results cannot be explained (Philp, 2018) (2) instability in model performance as a result of small sample size combined with a large number of correlated independent variables (3) a lack of external validation for currently proposed models. There are therefore two gaps that need to be addressed. The first is to explore if the traditional predictor selection methods can be replaced with modern methods. The second is to externally validate models that have been developed. In this study we will be addressing the first of the two of the problems.

Traditional methods are considered appropriate for datasets in which there is a random sample of the complete population, adequate sample size relative to the number of predictors and a low level of multicollinearity between variables. Given that most variables within football are related, the existence of multicollinearity is probable. Despite this, previous research has neglected to report and manage the existence of multicollinearity between variables (Bahr and Holme, 2003, Hägglund et al., 2006, Venturelli et al., 2011). Multicollinearity results in increased variance and an inability to identify the independent effect of a single predictor. This therefore, renders traditional methods less suitable and requires use of penalised methods for predictor selection e.g.

elastic net of zero-inflated Poisson. Additionally, datasets within football are also likely to be inflated by a high level of zero values given that more severe injuries, resulting in time lost from participation, are arguably rare events relative to the number of training or match play events that are injury free. Zeroes are additionally derived from instances in which players are injured but the injury is not reported as it is not considered significant enough. This makes injury prediction difficult and if not accounted for, predictor selection and model performance may be negated.

A range of modern modelling methods (e.g. Elastic net) have been developed with the potential to overcome some the limitations presented by traditional methods. These newer methods advantages over traditional methods, namely that they are able to select predictors in the presence of small sample sizes despite the existence of multicollinearity and have the ability to reduce the prediction error by shrinking unrelated predictors. Additionally, these methods can be integrated into models capable of managing datasets affected by zero-inflation (Desjardins, 2016).

Therefore, the first aim of this study was to explore whether penalised methods are more effective than traditional methods for predictor selection for injury in football. The second aim of this study was to develop a model, based on evaluation of the dataset and identified predictors, for prospective injury modelling in football.

2. Methodology

2.1. Study design

Ethical approval was granted and a pre-existing dataset collected as part of a prospective observational longitudinal study was used in this study (Philp et al., 2018). Model development consisted of two stages. In the first stage, an appropriate modern statistical method was used for shrinkage and predictor selection. Results of the modern statistical method were then evaluated against traditional methods. In the second stage, using traditional statistical methods, a model was developed based on the identified variables, allowing for estimations of bias, standard deviations and confidence intervals for injury prediction. Model development and predictor selection was carried out on an existing database from a prospective observational longitudinal study, set up in accordance with the consensus statement for data collection and injury reporting in football (Fuller et al., 2006, Philp et al., 2018). Additional personal training activities not planned by the team's coaching or fitness staff were recorded within the database alongside measures of fitness and workload. Approval was given to access an existing dataset and extract anonymous data related to injuries, screening, training/match play and demographic data.

2.2. Dataset characteristics

The data from one season (September 2015–May 2016) informed this study and contained variables related to a total of 24 male participants from a single football team competing in the British Universities and College Sports (BUCS) league. A total of 44 separate injury episodes were included in the dataset. Injury characteristics relating to severity have been outlined in table 1. The dependant variable selected for prediction was time loss injuries i.e. injuries with a severity of one day or more (n=33).

Table 1. Number of injuries for match and training according to injury severity categories.

		Number of injuries	
Injury severity	**Severity Category**	**Match**	**Training**
≤ 1day	slight	7	4
>1 day and < 3 days	minimal	5	2
>3 days and < 7 days	mild	5	1
> 7 days and < 28 days	moderate	10	8
> 28 days	severe	0	2
	career ending	0	0

The dataset contained a total of 34 variables (table 2).

Table 2. Variables contained within the dataset

Category	Number	Input
Position	**1**	Attacker
	2	Midfielder
	3	Defender
	4	Goalkeeper
Anthropometric	**5**	Kicking Leg
	6	Height
	7	Weight
	8	Sum of 4 sites skinfold thickness (biceps, triceps, subscapular suprailiac)
	9	Activity duration
Activity type	**10**	Match
	11	Training
	12	Futsal
	13	Conditioning
Surface type	**14**	Sand Astroturf
	15	Natural grass
	16	Artificial Astroturf (3G)
	17	Wooden
Injuries	**18**	Previous injuries
	19	In-season injuries
	20	Cumulative number of injuries (to case)
Variables related to training / match activities / fitness	**21**	Acute to Chronic workload ratio
	22	Cumulative match load
	23	Cumulative match grass load
	24	Total match Artificial Astroturf (3G) load
	25	Total training (all types) load
	26	Total training load (excluding futsal and conditioning)
	27	Total training grass load (excluding futsal and conditioning)
	28	Total training Sand Astroturf load (excluding futsal and conditioning)
	29	Total training Artificial astroturf (3G) load (excluding futsal and conditioning)
	30	Total training futsal load
	31	Total training load (with futsal) excluding conditioning
	32	Total training conditioning load
	33	Cumulative Match and Training load (22 + 23)
	34	Yo-Yo fitness score

*load refers to time in minutes

2.3. Data pre-processing and model selection

A summary of processes and results for the model and predictor selection stages have been outlined in figure 1. As a primary stage of the model development process, characteristics of the dependant variable (injury severity) were evaluated in order to identify the most appropriate modelling method. Sports and exercise medicine practitioners looking to reduce the risk of injury for individual players need to decide whether a team member's condition means they are safe to train and play in the squad on a sessional basis. Given that practitioners are concerned with the likelihood of injury over time and the number of days missed through injury, the dependant variable was considered as count data which follows a Poisson distribution. On further analysis it was identified that the count outcome of the dependant variable suffered from over-dispersion and excess zeros (table 3). Therefore, the zero-inflation Poisson (ZIP) regression model for prediction in the second stage was selected.

(ZIP) represents framework for the analysis the counts data that have distribution with high level of score zero counts than is expected for the Poisson distribution. ZIP assumes that the population involve of two kinds of individual. The first type represents the source of a Poisson-distributed count, which may be zero, whereas the second type always represents the source of a zero count. The distribution has two parameters, the mean of the Poisson distribution (λ) and the part of individuals that are of the second type (p). The mixture distribution formula is (Lambert, 1992):

$$P(y_i = k) = \begin{cases} p_i + (1 - p_i)e^{-\lambda_i} & if\ k = 0 \\ (1 - p_i)\dfrac{e^{-\lambda_i}{\lambda_i}^k}{k!} & if\ k = 1,2, \ldots. \end{cases}$$

where the parameters p_i, and λ depend on (vectors of) covariates xi and z_i, respectively through the logit and log links as:

$$logit(p_i) = x_i\gamma$$

$$\log(\lambda_i) = Z_i\theta$$

The dataset was structured to reflect the way in which the model would therefore be used in clinical practice i.e. each time a player trained or played a match, it was established as a separate episode (sample) in which injury could occur. The dataset therefore contained a total of 2784 episodes for potential injury. Missing data was handled using a multi-imputation method, in which uncertainty about the missing data is accounted for by generating multiple imputed datasets, the results of which are then combined (Sterne et al., 2009). Multi-

imputation methods are known to be more effective for reducing standard error and accounting for uncertainty related to the missing data (Sterne et al., 2009).

2.4. Predictor selection

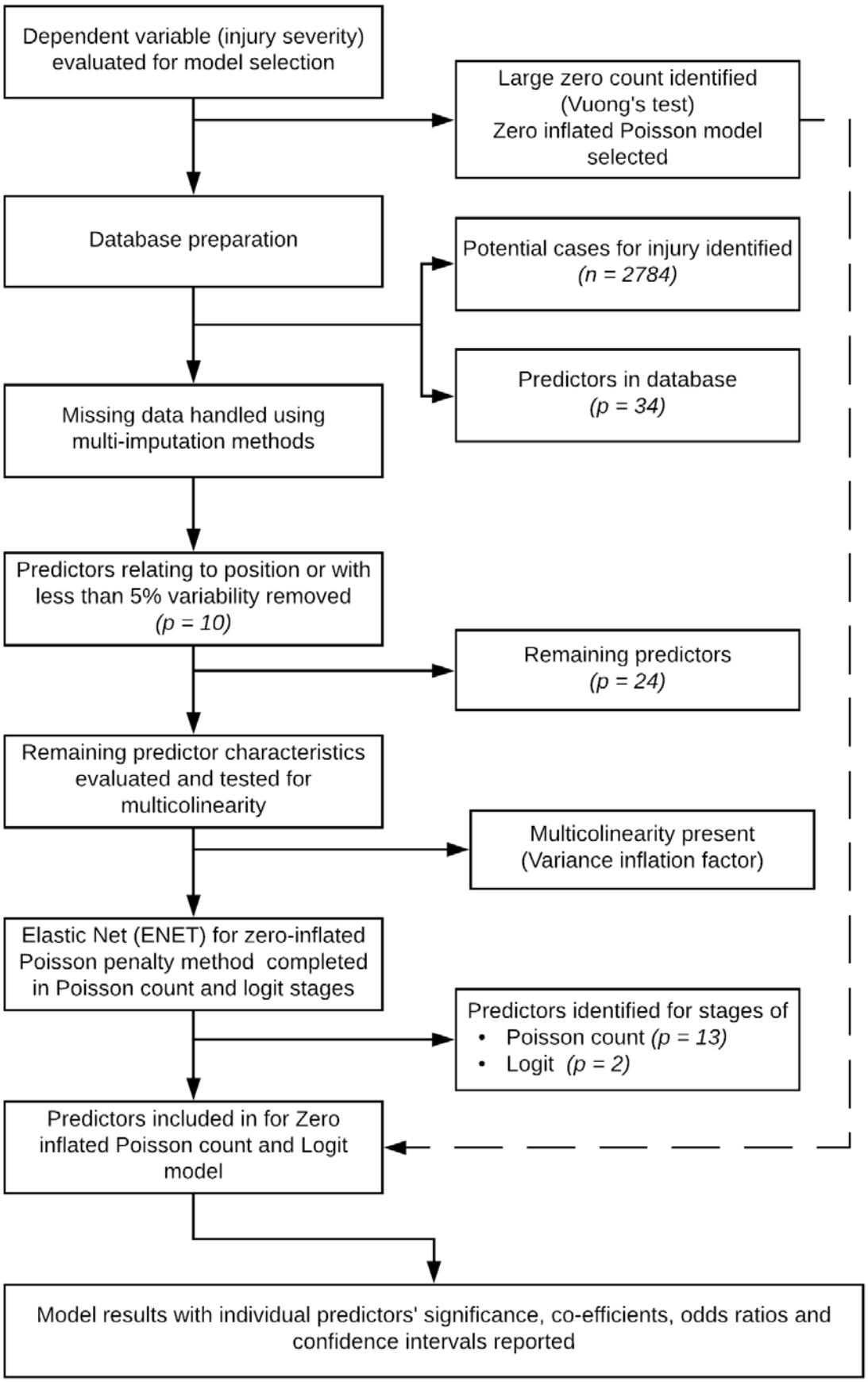
During the data preparation stage, any variables with variability less than 5% were removed as they have no discriminatory ability. The variables of **kicking leg**, surface types **artificial turf 3G and wooden,** and the activity of **futsal** were therefore removed. Additionally, 4 variables relating to player position were excluded as these categories are mutually exclusive to the respective position. In the next step, the variance inflation factor test (VIF) was used to determine if multicollinearity was present between the remaining independent variables. The method of predictors selection was then determined after evaluating characteristics of the independent variable. The Elastic Net (ENET) for zero-inflated Poisson penalty method was initially used.

ENET is used to improve selection predictors process when groups of predictors are highly correlated Zou and Hastie (2005). The elastic net of ZIP is defined by the formula(Tang et al., 2014):

$$\hat{\beta}_{ENet} = \arg min\{-L(\beta)\} + \lambda_1 \left\{ \sum_{j=1}^{p} |\theta_j| + \sum_{j=1}^{p} |\gamma_j| \right\} + \lambda_2 \left\{ \sum_{j=1}^{p} {\theta_j}^2 + \sum_{j=1}^{p} {\gamma_j}^2 \right\}$$

Cross-validation was then used to determine a penalty for shrinkage, established through a trade-off between bias and variance, whereby high levels of bias result in omission of appropriate predictors and high levels of variance result in inclusion of inappropriate predictors (Fan and Tang, 2013). Boot-strapping was not used. The process of shrinkage does not provide estimates of bias, standard deviations and confidence intervals. Therefore, these require estimation by integrating the identified predictors into the classical selected modelling method.

Figure 1. Summary of results for the model and predictor selection processes



3. Results

3.1. Results following evaluation of dependant variable for model selection

Vuong's test confirmed the presence of zero-inflation, with zero values determining more than 85% of the injury outcomes (table 3). The zero-inflated Poisson model was therefore selected.

Table 3. Vuong's test for the presence of zero-inflation

	Vuong z-statistic	Model comparison	p-value
Raw	-4.982125	model2 > model1	p < 0.001 ***
Akaike information criterion-corrected	4.978704	model2 > model1	p < 0.001***
Bayesian information criterion-corrected	-4.968557	model2 > model1	p < 0.001***

*** p < 0.001

3.2. Results for multicollinearity testing

Multicollinearity was identified between the independent variables following the VIF test, with scores > 10 indicating significant multicollinearity requiring correction (table 4).

Table 4. Variance inflation factor testing results for multicollinearity

Predictor label	**VIF value**
Weight	1.9922
SKF	2.2473
Time in activity	1.3662
Match	**32.2477◊**
Training	**39.6572◊**
Conditioning	**22.1688◊**
Sandastro	**27.3124◊**
Grass	**31.6022◊**
Artificial turf 3G	**14.9991◊**
Previous injuries	1.3860
In-season injuries	1.5483
Cumulative match volume	**103.6455◊**
Cumulative match Grass volume	**110.9217◊**
Total all training	8.2872
Total training volume excluding futsal and conditioning	**12.8611◊**
Total training Artificial turf 3G volume excluding futsal and conditioning	3.4659
Total training futsal volume	2.6395
Total training Grass volume excluding futsal and conditioning.	9.0142

◊ indicates high level of multicollinearity > 10

3.3. Results for predictor selection

Due to a high level of multicollinearity, the Elastic Net (ENET) for zero-inflated Poisson penalty method was initially used. As an inherent feature of the ENET zero-inflated Poisson (ZIP) regression model, the variable selection process is completed in two stages, namely the Poisson count (e.g. dependant variables increasing in a count fashion 1,2,3 etc.) and logit (dependant variables of a zero value) stages for predicting excess zeros (Desjardins, 2016). This is done to identify which predictors account for the count and zero dependant variables respectively. Despite the existence of multicollinearity, predictor selection was also carried out using the traditional ZIP model for comparative purposes. The predictors from both approaches are identified as per table 5.

Table 5. Results for the modern ENET for ZIP penalty method and traditional ZIP method

Modern ENET for ZIP penalty method *(n = 13)* Variable selection of count part Predictors name	**Traditional ZIP method *(n =11)* Variable selection of count part Predictors name**
Weight	
Sum of 4 sites skin fold thickness	Sum of 4 sites skin fold thickness
Time in activity	Time in activity
Training	Match
Conditioning	Grass
Artificial turf 3G	Artificial turf 3G
Previous injuries	Previous injuries
Acute to Chronic workload ratio	Cumulative number of injuries
Total time match play (3G)	Total time match play (3G)
Total time trained (grass)	Total time trained (grass)
Total time (futsal)	Total time (futsal)
Total time (conditioning)	Total time (conditioning)
Yo-Yo fitness score	

Modern ENET for ZIP penalty method (*n = 2)* Variable selection of Zero part Predictor name	**Traditional ZIP method *(n =5)* Variable selection of Zero part Predictor name**
Match	Sum of 4 sites skin fold thickness
In-season injuries	Sandastro
	Previous injuries
	Cumulative number of injuries
	Total time (conditioning)

The ENET for ZIP penalty method was more successful in shrinking the total number of predictors when compared to the traditional ZIP, with 15 and 16 predictors identified for each method respectively. Use of traditional methods resulted in selection of variables that were nonsensical for the zero and count parts of the model e.g. an increase in the number of previous injuries increased the odds of getting both a more severe injury and no injury.

3.4. Results for the ZIP model based on predictors identified using ENET

The predictors were then integrated into the ZIP model, results of which have been presented in table 6. The predictors of weight, training, artificial turf (3G), total time match play (3G), total time trained (grass), Yo-Yo fitness score, and previous injury were identified as being positively related with the count outcome of injury. Previous injury was however not identified as being statistically significant. Sum of 4 sites skin fold thickness, time in activity, acute to chronic workload ratio, total time in futsal, total time conditioning and the activity of conditioning were identified as being negatively related with the count outcome of injury. The activity of conditioning was however not identified as being statistically significant. For predictors relating to the zero part of the model, it was identified that the sources of zero-inflation within the dependant variable stemmed from the variables of match and in-season injury. Both predictors were negatively related with the zero outcome of the model i.e. for a one-unit increase in the identified variable, the likelihood of a zero outcome decreases by the respective value, assuming all other variables are constant, and this relation was statistically significant.

Table 6. Results for Regularized Zero-inflated Poisson Regression model
For the count outcome of the model:

$$logit(P_i) = \exp(0.022 * weight - 0.008 * sum\ of\ 4\ sites\ skin - 0.006 * time\ in\ activity + 0.35 * training - 0.94 * conditioning + 094 * \text{Artificial turf 3G} + 0.04 * \text{Previous injuries} - 0.295 * \text{Acute to Chronic workload ratio} + 0.001 * \text{Total time match play (3G)} + 0.002 * \text{Total time trained (grass)} - 0.005 * \text{Total time (futsal)} - 0.0004 * \text{Total time (conditioning)} + 0.182 * \text{Total time (conditioning)})$$

- Predictors with positive coefficients were identified as being positively related with the count outcome of injury i.e. for a one-unit increase in the identified variable, the likelihood of injury *increases* by the respective value, assuming all other variables are constant.
- Predictors with negative coefficients were identified as being negatively related with the count outcome of injury i.e. for a one-unit increase in the identified variable, the likelihood of injury *decreases* by the respective value, assuming all other variables are constant.

Variable selection of count part

Name of Predictor	Estimated coefficient	Standard deviation	Calculated value	P-value	Odds ratio	2.5%	97.5%
Weight	0.0220	0.0110	1.97	0.04*	1.0223	1.0002	1.0449
Sum of 4 sites skin fold thickness	-0.0080	0.0040	-2.33	0.01*	0.9913	0.9840	0.9986
Time in activity	-0.0060	0.0020	-3.2	0.001**	0.9936	0.9897	0.9975
Training	0.3490	0.1000	3.1	0.001**	1.4174	1.1372	1.7667
Conditioning	-0.94	0.5000	-1.9	0.06	0.3905	0.1480	1.0306
Artificial turf 3G	0.9410	0.2000	4.7	< 0.001***	2.5636	1.7365	3.7847
Previous injuries	0.0410	0.1000	0.4	0.68	1.0418	0.8560	1.2680
Acute to Chronic workload ratio	-0.2950	0.0800	-3.6	< 0.001***	0.7446	0.6337	0.8748
Total time match play (3G)	0.0010	0.0005	2.1	0.02*	1.0010	1.0001	1.0020
Total time trained (grass)	0.0020	0.0003	6.02	< 0.001***	1.0017	1.0012	1.0023
Total time (futsal)	-0.0050	0.0010	-5.0	< 0.001***	0.9948	0.9928	0.9968
Total time (conditioning)	-0.0004	0.0001	-8.2	< 0.001***	0.9996	0.9995	0.9997
Yo-Yo fitness score	0.1820	0.0080	2.3	0.01*	1.1993	1.0298	1.3968

For the zero outcome of the model:

$$\log(\lambda_i) = \exp(-1.25 * Match - 0.76 * in\ season\ injury)$$

- Predictors with negative coefficients were identified as being negatively related with the zero outcome i.e. for a one-unit increase in the identified variable, the likelihood of not getting an injury *decreases* by the respective value, assuming all other variables are constant.

Variable selection of Zero part

Name of Predictor	Estimated coefficient	Standard deviation	Calculated value	P-value	Odds ratio	2.5%	97.5%
Match	-1.2500	0.3500	-3.62	< 0.001***	0.2870	0.1460	0.5639
In-season injury	-0.7600	0.1300	-5.63	< 0.001***	0.4694	0.3608	0.6106

Significance codes: * p < 0.05, ** p < 0.01, *** p < 0.00

4. Discussion

4.1. Selection of methods for modelling processes

The novelty and strengths of this paper for application in sports injury modelling are the use of ZIP regression for dependant variables subject to zero-inflation, statistical testing for multicollinearity between independent variables and use of penalised methods (ENET) for predictor selection to reduce the confounding influence of multicollinearity in variable selection. Datasets relating to injury in football are likely to suffer from zero-inflation and a level of multicollinearity. On evaluation of the existing literature, these assumptions are not routinely tested, nor corrected for, and may explain limitations of existing models for identifying appropriate predictors and explaining their relationship to injury(Hägglund et al., 2006, Venturelli et al., 2011). Whilst some of our variables identified are consistent with the literature, a direct comparison is not possible. This is due to studies having non-identical variables within their datasets. Additionally, variations in methodology are likely to account for some differences observed given the limitations of existing traditional methods in the presence of multicollinearity.

4.2. Multicollinearity and predictor selection

The ENET for ZIP penalty method was more successful in shrinking the total number of predictors when compared to the traditional ZIP, with 15 and 16 predictors identified for each method respectively. It is important to note that whilst the difference may appear small, the traditional approach identified some predictors as having contradictory associations with injury i.e. the same predictor was found to both increase

and decrease injury risk. As previously discussed, the selection of variables that are nonsensical or lacking in physiological explanation limit clinical applicability due to distrust by practitioners. Given that traditional methods are unable to handle high levels of multicollinearity this may account for the observed results. Within our study it was identified that when determining variable coefficients using traditional methods, some variable values reached infinity. The use of traditional approaches for predictor selection in the presence of multicollinearity can be therefore be complex as selection of predictors in these cases is based on non-objective methods. The ENET penalised method for shrinkage therefore provides an objective statistical solution for predictor selection in the presence of multicollinearity.

As most variables within football are related, the existence of multicollinearity is probable. This step has not been routinely reported within other studies looking to identify risk factors or predictors for injury in football (Hägglund et al., 2006, Venturelli et al., 2011). Use of the ENET penalised method, following testing for multicollinearity, is therefore is a strength of our study. The phenomenon of multicollinearity results in increased variance and an inability identify the independent effect of a single predictor on the dependant variable. This renders traditional methods of predictor selection less suitable, as these are more appropriate in the absence of multicollinearity and when there is an adequate sample size relative to the number of predictors. Penalised methods may therefore be more appropriate for predictor selection in datasets containing a smaller sample sizes and levels of multicollinearity.

4.3. Predictors positively related with injury severity (count part)

As it is not possible to completely eliminate multicollinearity, there is still likely to be an interaction effect between variables. This is evident in the count part of the model. Surface type of **artificial turf 3G** was found to have the largest positive effect on injury (OR 2.6 95%CI 1.7 to 3.8) and this variable is consistent with previous studies (Ekstrand et al., 2011a, Ekstrand et al., 2006, Fuller et al., 2007b, Fuller et al., 2007a, Kristenson et al., 2013). However, it has been identified that increased duration on surface types and not surface type alone are linked to increased injury risk (Kristenson et al., 2013, Aoki et al., 2010). There is therefore an interaction effect between surface type and variables of **training, total time training on grass, total time match play on artificial turf 3G** and **Yo-Yo IR2,** with these variables being positively associated with injury. It is to be expected that

increased participation, facilitated by increased cardiovascular capacity, in football increases the risk of injury. Therefore, practitioners wishing to mitigate injury risk may consider the frequency and duration of activity on different surface types. Furthermore, this should be considered within the capacity of the player. Other studies have identified positive relationships between previous injury and subsequent injuries(Hägglund et al., 2006, Venturelli et al., 2011). This relationship is consistent with our study although statistical significance was not reached, possibly owing to the use of self-reported injury history, which is subject to recall bias and underestimation of injuries (Junge and Dvorak, 2004). Therefore, accurate injury records obtained by practitioners from previous seasons are required if previous injury is to be used for prospective injury modelling (Hägglund et al., 2006). A further interaction effect was also found between the variables of weight and skin fold thickness, having positive and negative relations to injury respectively. Within the literature no consistent anthropometric traits are associated with injury (Frisch et al., 2011, Arnason et al., 2004, Gajhede-Knudsen et al., 2013, Fousekis et al., 2011) although similar results to our study have been identified for players of a lower lean mass having increased risk of hamstring injuries (Henderson et al., 2010). It may be hypothesised that increased body fat, up to a point, has a protective effect against injury giving the sustained demands on players throughout the season, accounting for the observed results.

4.4. Predictors negatively related with injury severity (count part)

As a result of the interaction effect, not all predictors related to time in activity resulted in increased injury risk. The predictors of **time in activity** related to activities of training, match play, acute to chronic workload ratio, futsal and conditioning were found to have a negative relation with injury. This possibly indicates that players who have an ability to engage in these activities without getting injured are less likely to sustain a severe injury and are better conditioned as a result. For example, it is known that undertaking resistance exercise has been linked to a reduction in injury with higher levels of severity (van der Horst et al., 2015). Whilst the predictors of **conditioning, total time (conditioning)** and **total time (futsal)** fall outside of the consensus statement (Fuller et al., 2006), it was recognised that within our study, any forms of additional resistance, skill development or fitness training needed to be included as these would likely be conducted outside of formal training. It is acknowledged that time is not the only determinant related to load or forms of technical, resistance or cardiovascular training. There may therefore be other determinants within these variables linked to injury which need to be considered

alongside the complex nature of injury. For example, it is recognised that the acute to chronic workload is known to have a non-linear association with injury risk (Malone et al., 2017, Hulin et al., 2016) and has been applied to multiple metrics of performance (Hulin et al., 2016, Hulin et al., 2014, Møller et al., 2017). It is therefore unknown how the predictor selection and modelling process would be affected should the index be based on alternate measures of performance e.g. total distance. However, for clinical application, these results support increased time (up to a point) for engaging in activities relating to load, resistance and skill development sessions for injury risk reduction.

4.5. Predictors related with zero part

The zero component of the ZIP model identifies factors contributing to either an increased or decreased odd of getting a zero i.e. no injury or injury severity of less than one day. The variables of in-season injury and match play were found to have negative relations with this outcome. Therefore, for a single unit increase in the events of a match or in-season injury, players were less likely to get a zero i.e. not sustain a time-loss injury (OR 0.2870 and 0.4690 respectively). The larger effect was seen for in-season injury. This predictor, used as a cumulative total, comprised of time-loss and non-time-loss injuries. Within the existing literature more severe injuries are known to be preceded by less severe injuries (Ekstrand and Gillquist, 1983, Gajhede-Knudsen et al., 2013). This would therefore result in a more severe time-loss injury, reducing the presence of zeros, for which our model gives support. Injuries sustained during the season may lower the overall functional capacity of the player, resulting from pain or decreased conditioning. As a result of this, and possibly coupled by the existence of an injury which has not been fully rehabilitated, players may go beyond their functional capacity resulting in more severe time-loss injures. Therefore, it is important to establish any limitations associated with in-season injuries, for both time-loss and non-time-loss injuries, as identification of these factors may reduce the occurrence of a more severe time-loss injury.

Match play was also found to have a negative relation with the zero outcome. In comparison to training, matches are known to have a higher rate and number of injuries (Ekstrand et al., 2011b), possibly explained by the fact that the functional demands of a match are higher. This is also supported by the injury characteristics within our dataset (table 1). As a result of the greater functional demands and competitive nature of matches, it may be expected that more injuries of greater severity will be sustained during match play, therefore, reducing the presence of zeros. In comparison to alternate models of injury(Hägglund et al., 2006, Venturelli et al., 2011),

where injury episodes are viewed as separate independent events owing to the nature of the modelling methods, our model assumes a cumulative risk of injury. Based on the results of the model's zero component, sports and exercise practitioners may modify or limit the number of consecutive matches in which a player competes in order to prevent a player sustaining a time-loss injury.

4.6. Limitations of the model

A strength of the methods in our study is that the count and zero outcomes were modelled independently though use of the ZIP model. This is in contrast to other studies which combine zero and count outcomes, possibly overlooking the presence of zero-inflation (Hagglund et al., 2005, Hägglund et al., 2006, Venturelli et al., 2011). This may provide some insight into the limitations of existing models, given that the nature of the data violates the premise of some models e.g. for a model assuming a Poisson distribution it is assumed that the variance equals the mean, however for zero-inflated datasets this is not the case. ZIP is also appropriate for studies looking to identify the sources of zero-inflation and in which a zero outcome may be derived from two sources or processes (Wang et al., 2015). Within our study zeros may have been derived from either the existence of no injury or the presence of a non-time-loss injury/ non-reported injury. A limitation of the modelling method, however is that it is not able to identify from which source the zero is derived.

Within our study, penalised methods for predictor selection, evaluated using cross-validation, have been identified as superior when compared traditional predictors selection methods. It is recognised that within our study we were unable to determine the accuracy of the final model on an unseen data given the small number of more severe injury cases. A larger dataset is therefore required to investigate the sensitivity and specificity of our model for comparison against existing models. Additionally, alternate datasets may have access to a greater number of teams over longer periods of time which may help in identification of smaller relations with injury (Arnason et al., 2004, Hägglund and Waldén, 2016, Hägglund et al., 2006, Bahr and Holme, 2003). However, if models are to be integrated routinely into clinical decision making, they should be clinically useful in football squads of a typical size and retain the function of prospectively identifying injury as the dataset is populated in real-time. It is also acknowledged that variables collected for measures of injury risk and performance will differ between teams for frequency, measures collected and units used to inform indexes such as the acute to chronic workload ratio (Hulin et al., 2014). Therefore, the predictors identified within this study are based on the

measures available to the researchers and discretion should be used when applying the model to other datasets comprised of different variables. This does not however detract from or negate the processes used for predictor and model selection.

5. Conclusion

Penalised methods for predictor selection and use of zero-inflated Poisson regression modelling for predicting time loss injuries have been identified as alternate and appropriate methods. These methods are more appropriate for datasets subject to multicollinearity and zero-inflation known to affect the performance of traditional statistical methods.

AOKI, H., KOHNO, T., FUJIYA, H., KATO, H., YATABE, K., MORIKAWA, T. & SEKI, J. 2010. Incidence of injury among adolescent soccer players: a comparative study of artificial and natural grass turfs. *Clin J Sport Med,* 20**,** 1-7.

ARNASON, A., SIGURDSSON, S. B., GUDMUNDSSON, A., HOLME, I., ENGEBRETSEN, L. & BAHR, R. 2004. Risk factors for injuries in football. *Am J Sports Med,* 32**,** 5s-16s.

BAHR, R. & HOLME, I. 2003. Risk factors for sports injuries—a methodological approach. *British journal of sports medicine,* 37**,** 384-392.

DESJARDINS, C. D. 2016. Modeling Zero-Inflated and Overdispersed Count Data: An Empirical Study of School Suspensions. *The Journal of Experimental Education,* 84**,** 449-472.

EKSTRAND, J. & GILLQUIST, J. 1983. The avoidability of soccer injuries. *Int J Sports Med,* 4**,** 124-8.

EKSTRAND, J., HAGGLUND, M. & FULLER, C. W. 2011a. Comparison of injuries sustained on artificial turf and grass by male and female elite football players. *Scand J Med Sci Sports,* 21**,** 824-32.

EKSTRAND, J., HÄGGLUND, M. & WALDÉN, M. 2011b. Injury incidence and injury patterns in professional football: the UEFA injury study. *British Journal of Sports Medicine,* 45**,** 553-558.

EKSTRAND, J., TIMPKA, T. & HÄGGLUND, M. 2006. Risk of injury in elite football played on artificial turf versus natural grass: a prospective two-cohort study. *British Journal of Sports Medicine,* 40**,** 975-980.

FAN, Y. & TANG, C. Y. 2013. Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 75**,** 531-552.

FOUSEKIS, K., TSEPIS, E., POULMEDIS, P., ATHANASOPOULOS, S. & VAGENAS, G. 2011. Intrinsic risk factors of non-contact quadriceps and hamstring strains in soccer: a prospective study of 100 professional players. *Br J Sports Med,* 45**,** 709-14.

FRISCH, A., URHAUSEN, A., SEIL, R., CROISIER, J. L., WINDAL, T. & THEISEN, D. 2011. Association between preseason functional tests and injuries in youth football: a prospective follow-up. *Scand J Med Sci Sports,* 21**,** e468-76.

FULLER, C. W., DICK, R. W., CORLETTE, J. & SCHMALZ, R. 2007a. Comparison of the incidence, nature and cause of injuries sustained on grass and new generation artificial turf by male and female football players. Part 1: match injuries. *Br J Sports Med,* 41 Suppl 1**,** i20-6.

FULLER, C. W., DICK, R. W., CORLETTE, J. & SCHMALZ, R. 2007b. Comparison of the incidence, nature and cause of injuries sustained on grass and new generation artificial turf by male and female football players. Part 2: training injuries. *Br J Sports Med,* 41 Suppl 1**,** i27-32.

FULLER, C. W., EKSTRAND, J., JUNGE, A., ANDERSEN, T. E., BAHR, R., DVORAK, J., HÄGGLUND, M., MCCRORY, P. & MEEUWISSE, W. H. 2006. Consensus statement on injury definitions and data collection procedures in studies of football (soccer) injuries. *British Journal of Sports Medicine,* 40**,** 193-201.

GAJHEDE-KNUDSEN, M., EKSTRAND, J., MAGNUSSON, H. & MAFFULLI, N. 2013. Recurrence of Achilles tendon injuries in elite male football players is more common after early return to play: an 11-year follow-up of the UEFA Champions League injury study. *Br J Sports Med,* 47**,** 763-8.

HÄGGLUND, M. & WALDÉN, M. 2016. Risk factors for acute knee injury in female youth football. *Knee Surgery, Sports Traumatology, Arthroscopy,* 24**,** 737-746.

HAGGLUND, M., WALDEN, M., BAHR, R. & EKSTRAND, J. 2005. Methods for epidemiological study of injuries to professional football players: developing the UEFA model. *British Journal of Sports Medicine,* 39**,** 340-346.

HÄGGLUND, M., WALDÉN, M. & EKSTRAND, J. 2006. Previous injury as a risk factor for injury in elite football: a prospective study over two consecutive seasons. *British Journal of Sports Medicine,* 40**,** 767-772.

HENDERSON, G., BARNES, C. A. & PORTAS, M. D. 2010. Factors associated with increased propensity for hamstring injury in English Premier League soccer players. *J Sci Med Sport,* 13**,** 397-402.

HULIN, B. T., GABBETT, T. J., BLANCH, P., CHAPMAN, P., BAILEY, D. & ORCHARD, J. W. 2014. Spikes in acute workload are associated with increased injury risk in elite cricket fast bowlers. *British Journal of Sports Medicine,* 48**,** 708-712.

HULIN, B. T., GABBETT, T. J., LAWSON, D. W., CAPUTI, P. & SAMPSON, J. A. 2016. The acute:chronic workload ratio predicts injury: high chronic workload may decrease injury risk in elite rugby league players. *British Journal of Sports Medicine,* 50**,** 231-236.

JUNGE, A. & DVORAK, J. 2004. Soccer injuries: a review on incidence and prevention. *Sports Med,* 34**,** 929-38.

KIESEL, K., PLISKY, P. J. & VOIGHT, M. L. 2007. Can Serious Injury in Professional Football be Predicted by a Preseason Functional Movement Screen? *North American Journal of Sports Physical Therapy: NAJSPT,* 2**,** 147.

KRISTENSON, K., BJORNEBOE, J., WALDEN, M., ANDERSEN, T. E., EKSTRAND, J. & HAGGLUND, M. 2013. The Nordic Football Injury Audit: higher injury rates for professional football clubs with third-generation artificial turf at their home venue. *Br J Sports Med,* 47**,** 775-81.

LAMBERT, D. 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics,* 34**,** 1-14.

MALONE, S., OWEN, A., NEWTON, M., MENDES, B., COLLINS, K. D. & GABBETT, T. J. 2017. The acute: chonic workload ratio in relation to injury risk in professional soccer. *Journal of science and medicine in sport,* 20**,** 561-565.

MØLLER, M., NIELSEN, R. O., ATTERMANN, J., WEDDERKOPP, N., LIND, M., SØRENSEN, H. & MYKLEBUST, G. 2017. Handball load and shoulder injury rate: a 31-week cohort study of 679 elite youth handball players. *British Journal of Sports Medicine,* 51**,** 231-237.

MYER, G. D., FORD, K. R., KHOURY, J., SUCCOP, P., HEWETT, T. E., MYER, G. D., FORD, K. R., KHOURY, J., SUCCOP, P. & HEWETT, T. E. 2011. Biomechanics laboratory-based prediction algorithm to identify female athletes with high knee loads that increase risk of ACL injury. *British Journal of Sports Medicine,* 45**,** 245-252.

PHILP, F. 2018. *Validating models of injury risk prediction in football players.* Doctorate of Philosophy, Keele University.

PHILP, F., BLANA, D., CHADWICK, E. K., STEWART, C., STAPLETON, C., MAJOR, K. & PANDYAN, A. D. 2018. Study of the measurement and predictive validity of the Functional Movement Screen. *BMJ Open Sport & Exercise Medicine,* 4.

STERNE, J. A., WHITE, I. R., CARLIN, J. B., SPRATT, M., ROYSTON, P., KENWARD, M. G., WOOD, A. M. & CARPENTER, J. R. 2009. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj,* 338**,** b2393.

TANG, Y., XIANG, L. & ZHU, Z. 2014. Risk factor selection in rate making: EM adaptive LASSO for zero-inflated poisson regression models. *Risk Analysis,* 34**,** 1112-1127.

VAN DER HORST, N., SMITS, D.-W., PETERSEN, J., GOEDHART, E. A. & BACKX, F. J. 2015. The preventive effect of the nordic hamstring exercise on hamstring injuries in amateur soccer players: a randomized controlled trial. *The American journal of sports medicine,* 43**,** 1316-1323.

VENTURELLI, M., SCHENA, F., ZANOLLA, L., BISHOP, D., VENTURELLI, M., SCHENA, F., ZANOLLA, L. & BISHOP, D. 2011. Injury risk factors in young soccer players detected by a multivariate survival model. *Journal of Science & Medicine in Sport,* 14**,** 293-298.

WANG, Z., SHUANGGE, M. & WANG, C.-Y. 2015. Variable selection for zero-inflated and overdispersed data with application to health care demand in Germany. *Biometrical journal. Biometrische Zeitschrift,* 57**,** 867-884.

ZOU, H. & HASTIE, T. 2005. Regularization and variable selection via the elastic-net. *Journal of the Royal Statistical Society,* 67**,** 301-320.