Est. 1841 | YORK ST JOHN UNIVERSITY

Philp, Fraser, Al-Shallawi, Ahmad, Kyriacou, Theocharis, Blana, Dimitra and Pandyan, Anand (2020) Improving predictor selection for injury modelling methods in male footballers. BMJ open sport & exercise medicine, 6 (1).

Downloaded from: https://ray.yorksj.ac.uk/id/eprint/13101/

The version presented here may differ from the published version or version of record. If you intend to cite from the work you are advised to consult the publisher's version: https://doi.org/10.1136/bmjsem-2019-000634

Research at York St John (RaY) is an institutional repository. It supports the principles of open access by making the research outputs of the University available in digital form. Copyright of the items stored in RaY reside with the authors and/or other copyright owners. Users may access full text items free of charge, and may download a copy for private study or non-commercial research. For further reuse terms, see licence terms governing individual outputs. Institutional Repositories Policy Statement

RaY

Research at the University of York St John

For more information please contact RaY at ray@yorksj.ac.uk

BMJ Open Sport & Exercise Medicine

Improving predictor selection for injury modelling methods in male footballers

Fraser Philp ,[1] Ahmad Al-shallawi,[2,3] Theocharis Kyriacou,[4] Dimitra Blana,[2] Anand Pandyan[1]

To cite: Philp F, Al-shallawi A, Kyriacou T, *et al*. Improving predictor selection for injury modelling methods in male footballers. *BMJ Open Sport & Exercise Medicine* 2020;**6**:e000634. doi:10.1136/bmjsem-2019-000634

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/bmjsem-2019-000634).

Accepted 15 December 2019

Check for updates

© Author(s) (or their employer(s)) 2020. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

[1]School of Health and Rehabilitation, Keele University, Keele, UK
[2]Institute of Science and Technology in Medicine, Keele University, Keele, UK
[3]The Engineering Technical College of Mosul, Northern Technical University, Mosul, Nineveh, Iraq
[4]School of Computing and Mathematics, Keele University, Keele, UK

Correspondence to
Dr Fraser Philp;
f.d.philp@keele.ac.uk

ABSTRACT

Objectives This objective of this study was to evaluate 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.

Methods Predictor selection and model development was conducted on a pre-existing dataset of 24 male participants from a single English football team's 2015/2016 season.

Results The elastic net for zero-inflated Poisson penalty method was successful in shrinking the total number of predictors in the presence of high levels of multicollinearity. It was additionally identified that easily measurable data, that is, 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. Furthermore, prolonged series of match-play and increased in-season injury reduced the probability of not sustaining an injury.

Conclusion 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.

What are the new findings?

- Modern penalised methods are superior to traditional methods for predictor selection in datasets with high levels of multicollinearity and zero inflation.
- Use of traditional predictor selection methods in datasets with high levels of multicollinearity may result in selection of variables with contradictory mechanisms or a lack of physiological explanation, limiting clinical application.

How might it impact on clinical practice in the future?

- Modern predictor selection methods may be used objectively to refine data collection processes in football datasets containing a large number of variables.
- Improvement of predictor selection processes may improve model stability for prospective injury modelling, further facilitating implementation and application for informing clinical decision-making.

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,[1] laboratory-based algorithms[2] and statistical models.[3–6] Within football, injury reporting, recording[7] and predictor selection methods are informed by existing frameworks which advocate the use of multivariate statistical models.[3 4] 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, that is, 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, and sports and exercise medicine practitioners may have an inherent distrust of complex models in which the results cannot be explained[8 9]; (2) instability in model performance, stemming from small sample sizes combined with large numbers of correlated independent variables; (3) a lack of external validation. There are therefore two gaps that need to be addressed: Firstly, to explore if traditional predictor selection methods can be replaced with modern methods. Secondly,

BMJ Open Sport & Exercise Medicine: first published as 10.1136/bmjsem-2019-000634 on 14 January 2020. Downloaded from https://bmjopensem.bmj.com on 2 December 2025 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

to externally validate models that have been developed. This study will address the first problem.

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. 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.[3 5 6] 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, for example, elastic net of zero-inflated Poisson. In addition, 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.

A range of modern modelling methods (eg, elastic net) have been developed with the potential to overcome some the limitations presented by traditional methods. These newer methods have 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. In addition, these methods can be integrated into models capable of managing datasets affected by zero-inflation.[10]

The first aim of this study was to explore whether penalised methods are more effective than traditional methods for predictor selection. 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.

METHODOLOGY

Study design

Source of data and participants

Ethical approval was granted and a pre-existing dataset collected as part of a prospective observational longitudinal study (set up in accordance with the consensus statement for data collection and injury reporting in football) was used in this study.[7 11] 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. 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 league. The mean age for participants was 19 years (range, 19–22) with a mean history of 12.13 years (SD 2.1) playing football. The mean number of self-reported previous injuries was 1.42 (SD 1.2). Participants had a mean standing height of 1.79 m (SD 0.06), mean weight of 77.75 kg (SD 9.7) and a mean skinfold thickness (sum of four sites: biceps, triceps, subscapular and anterior superior iliac spine) of 40.98 mm (SD 17.0). Twenty-two participants reported their preferred kicking leg (dominant leg) as being their right leg, and the remaining two participants reported their dominant leg as being their left leg. There were 4 attackers, 13 mid-fielders, 4 defenders and 3 goalkeepers. A total of 44 separate injury episodes were included in the dataset. Injury characteristics relating to severity have been outlined in table 1. Further information regarding injury reporting and recording methods and participant characteristics have been reported previously.[8]

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

Injury severity	Severity category	Number of injuries	
		Match	Training
≤1 day	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

Dependent variable/outcome

The count outcome/dependent variable selected for prediction was the number of days lost to injury, that is, time-loss injuries (injuries with a severity of 1 day or more)[7] (n=33). An assumption of our study was that all injury episodes were independent of each other, that is, when participants returned to training and match play, they had fully recovered from a previous injury and that the circumstances associated with each injury episode were unique to that injury episode. We acknowledge that there may be some serial dependency between some injury cases which violates an assumption of zero-inflation Poisson (ZIP) regression. However, we have selected this method as currently there is no systematic process to inform the circumstances under which previous injury would be causal of future injury. This would also more accurately reflect the way in which the model would be used in clinical practice, as during a progressive season it is likely that players will sustain more than one injury and therefore appear multiple times within a dataset.[12] In addition, there may be potential for the addition of new players throughout the season and attrition of players over multiple seasons who are in turn replaced. Under these circumstances, sports and exercise medicine practitioners need to make decisions regarding suitability to train and play during a progressive season, at any given time, with limited retrospective data, and this has to be unbiased.

Independent variables/predictors

The dataset contained a total of 34 variables (table 2). Further information regarding the methods of

BMJ Open Sport & Exercise Medicine: first published as 10.1136/bmjsem-2019-000634 on 14 January 2020. Downloaded from https://bmjopensem.bmj.com on 2 December 2025 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

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: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.

recording of the independent variables have been reported previously.[8]

Data pre-processing, predictor selection and model development

A summary of processes and results for the model and predictor selection stages have been outlined in figure 1. The dataset was structured to reflect the way in which the model would be used in clinical practice, that is, 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. All analysis was conducted in the software package R V.3.5.1[13] using the associated packages listed in online supplementary file 1. Missing data were handled using a multi-imputation method.[14 15] The multi-imputation method was undertaken based on the predictive mean matching for continuous predictors and multinomial logistic regression for categorical predictors. A single database was used for the predictor selection and model development processes as outlined below.

BMJ Open Sport & Exercise Medicine: first published as 10.1136/bmjsem-2019-000634 on 14 January 2020. Downloaded from https://bmjopensem.bmj.com on 2 December 2025 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

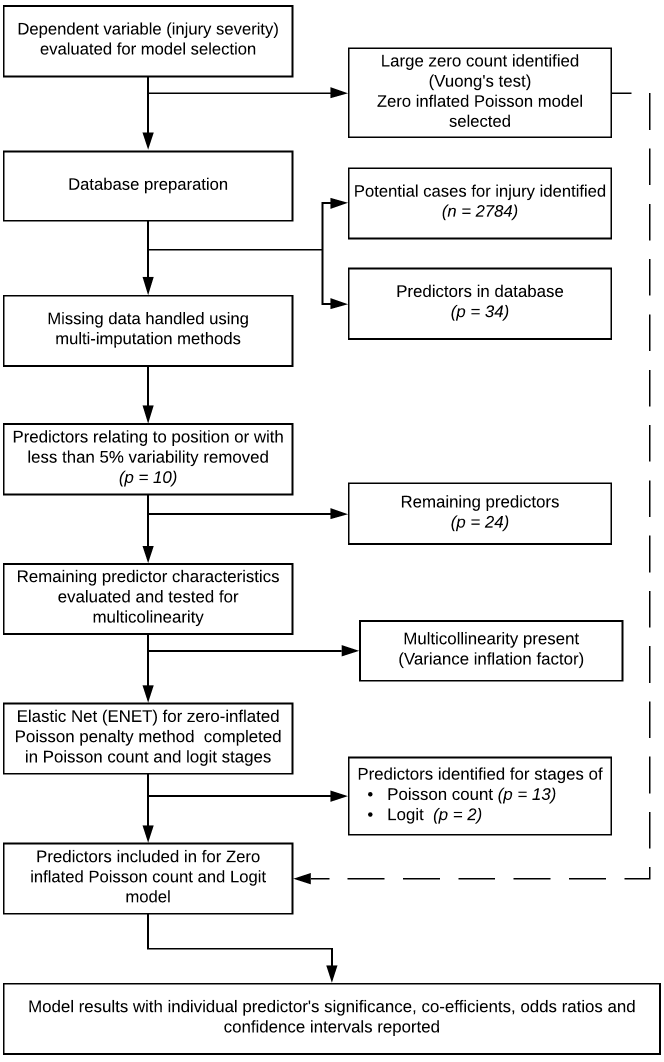

Figure 1 Summary of results for the predictor selection and modelling processes.

Characteristics of the dependent 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 dependent variable was considered as count data which follows a Poisson distribution. On further analysis, it was identified that the count outcome of the dependent variable suffered from overdispersion and excess zeros (table 3). Therefore, the ZIP regression model for prediction in the second stage was selected.[16]

Predictor selection

Any variables with less than 5% variability 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. In addition, four variables relating to player position were excluded as these categories are mutually exclusive to the respective position. The variance inflation factor test (VIF) was then used to determine if multicollinearity was present between the remaining independent variables. The method of predictor selection was then determined after evaluating characteristics of the independent variables.

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

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

***p<0.001.

For comparison of the predictor selection methods, the elastic net (ENET) for ZIP was evaluated against backward stepwise regression methods with a significance level alpha=0.01. A significance level alpha=0.01 was recommended for selecting the most important predictors when using the backward stepwise regression method.[17] Within our dataset, there were a limited number of injury cases and an assumption of our study was that injury episodes were independent of each other. We have acknowledged the limitations of these and the significance level alpha=0.01 was therefore selected in order to develop a conservative model with minimisation of type 1 errors. All model details can be found in online supplementary files 1 and 2. Performance of predictor selection process was evaluated using Akaike information criterion (AIC), Bayesian information criterion (BIC) and log likelihood.[18]

For the ENET predictor selection method (with cross-validation), the expectation maximisation (EM) algorithm was applied to find an optimal solution to the count and zero parts of the model in order to determine a penalty for shrinkage during ten-fold cross-validation.[19] The process of shrinkage does not provide estimates of bias, SD and CIs. Therefore, these require estimation by integrating the identified predictors into a classical modelling method.

For the ENET predictor selection method (without cross-validation), the optimal tuning parameter was calculated using BIC over the grid of candidate values. BIC has shown to be consistent in variable selection.[20]

RESULTS

Results following evaluation of dependent 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.

BMJ Open Sport & Exercise Medicine: first published as 10.1136/bmjsem-2019-000634 on 14 January 2020. Downloaded from https://bmjopensem.bmj.com on 2 December 2025 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Table 4 Variance inflation factor (VIF) testing results for multicollinearity

Predictor label	VIF value
Weight	1.9922
Sum of 4 sites skinfold thickness (biceps, triceps, subscapular suprailiac)	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.

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).

Results for predictor selection

Results for statistical comparison of ENET for ZIP and 'traditional' predictors selection methods are presented in table 5.

It was identified that for AIC and BIC, the modern ENET without cross-validation was superior to the full and backward stepwise regression models. For the log-likelihood criteria, ENET with cross-validation was superior to all models.

Due to a high level of multicollinearity, the ENET for ZIP penalty method was therefore selected.[21 22] As an inherent feature of the ENET ZIP regression model, the variable selection process is completed in two stages, namely the Poisson count (eg, dependent variables increasing in a count fashion 1, 2, 3 etc) and logit (dependent variables of a zero value) stages for predicting excess zeros.[10] Predictor selection was also carried out using the traditional backward stepwise regression for ZIP model for comparative purposes, despite the existence of multicollinearity (table 6).

During the modelling process, the full model traditional method was unable to handle the high levels of multicollinearity. It was identified that when determining variable coefficients using traditional methods, some variable values reached infinity. The variables of sum of four sites skinfold thickness (biceps, triceps, subscapular, suprailiac) (8), artificial Astroturf (3G) (16), cumulative number of injuries (to case) (20), cumulative match grass load (23) and total match artificial Astroturf (3G) load (24) therefore had to be excluded to find the full model so that the backward stepwise variable selection method could be applied with significance level alpha=0.01.

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

Results for the ZIP model based on predictors identified using ENET

The predictors were then integrated into the ZIP model (table 7). 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 skinfold thickness, time in activity, acute:chronic workload ratio, total

Table 5 Results for comparison of elastic net (ENET) for zero-inflated Poisson and 'traditional' predictors selection methods

	Full model	Backward stepwise regression	ENET without cross-validation	ENET with cross-validation
Akaike information criterion (AIC)	666.6	665.23	662.31	–
Bayesian information criterion (BIC)	909	793	750	–
Log likelihood	–929	–301.7	–298.0	–46.65

For AIC and BIC, lower values are indicative of better model performance while for log likelihood, a larger number is indicative of better model performance.
Calculation of AIC and BIC is not possible with ENET with cross-validation.

BMJ Open Sport & Exercise Medicine: first published as 10.1136/bmjsem-2019-000634 on 14 January 2020. Downloaded from https://bmjopensem.bmj.com on 2 December 2025 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Table 6 Results for the modern elastic net (ENET) for zero-inflated Poisson (ZIP) penalty method and traditional ZIP method

Modern ENET for ZIP penalty method (n=13) Variable selection of count part Predictor name	Traditional ZIP method (n=11) Variable selection of count part Predictor name
Weight	
Sum of 4 sites skinfold thickness	Sum of 4 sites skinfold thickness
Time in activity	Time in activity
Training	Match
Conditioning	Grass
Artificial turf 3G	Artificial turf 3G
Previous injuries	Previous injuries
Acute:chronic workload ratio	Cumulative no 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 skinfold thickness
In-season injuries	Sandastro
	Previous injuries
	Cumulative no of injuries
	Total time (conditioning)

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 dependent variable stemmed from the variables of match and in-season injury. Both predictors were negatively related with the zero outcome of the model, that is, 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.

DISCUSSION

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 dependent 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 as most variables within football are related. 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.[5 6] We did attempt to compare our model performance with the existing literature; however, due to the following challenges, a direct comparison was not possible: (1) the number and type of independent variables between datasets varied significantly[5 6 23]; (2) the methods for reporting the dependent variable varied significantly between studies,[3] for example, some studies used number of days lost to injury[1 6] whereas some used injury classification subtypes such as anatomical site[5 6] or tissue type[24]; (3) reported modelling methods had limited clinical applicability and explanatory validity[9] or (4) did not robustly manage the presence of zero inflation or multicollinearity between variables. While some of our variables identified are consistent with the literature, when evaluating these against existing studies, it is important that this is done within the context of the previously mentioned points.

Multicollinearity and predictor selection

Multicollinearity results in increased variance and an inability to identify the independent effect of a single predictor on the dependent 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. The ENET for ZIP penalty method was more successful in shrinking the total number of predictors when compared with the traditional ZIP within our study. While the difference in total number of predictors may appear small, the traditional approach identified some predictors as having contradictory associations with injury, that is, the same predictor was found to both increase and decrease injury risk. Contradictory predictors selection which lack physiological explanation cause distrust by practitioners and limit clinical applicability. Traditional methods are unable to handle high levels of multicollinearity and 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. Traditional approaches for predictor selection in the presence of multicollinearity is complex, as selection in these cases is based on non-objective methods. The ENET penalised method for shrinkage therefore provides an objective

BMJ Open Sport & Exercise Medicine: first published as 10.1136/bmjsem-2019-000634 on 14 January 2020. Downloaded from https://bmjopensem.bmj.com on 2 December 2025 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Table 7 Results for regularised zero-inflated Poisson regression model

For the count outcome of the model:

Count model or log(λi)=(0.022∗weight−0.008∗sum of 4 sites skinfold thickness−0.006∗time in activity+0.35∗training−0.94∗conditioning+094∗artificial turf 3G+0.04∗previous injuries−0.295∗acute:chronic workload ratio+0.001∗total time match-play (3G)+0.002∗total time trained (grass)−0.005∗total time (futsal)−0.0004∗total time (conditioning)+0.182∗total time (conditioning))

Name of predictor	Estimated coefficient	SD	Calculated value	P value	OR	2.50%	97.50%
Weight	0.022	0.011	1.97	0.04*	1.0223	1.0002	1.0449
Sum of 4 sites skinfold thickness	−0.008	0.004	−2.33	0.01*	0.9913	0.984	0.9986
Time in activity	−0.006	0.002	−3.2	0.001**	0.9936	0.9897	0.9975
Training	0.349	0.1	3.1	0.001**	1.4174	1.1372	1.7667
Conditioning	−0.94	0.5	−1.9	0.06	0.3905	0.148	1.0306
Artificial turf 3G	0.941	0.2	4.7	<0.001***	2.5636	1.7365	3.7847
Previous injuries	0.041	0.1	0.4	0.68	1.0418	0.856	1.268
Acute:chronic workload ratio	−0.295	0.08	−3.6	<0.001***	0.7446	0.6337	0.8748
Total time match-play (3G)	0.001	0.0005	2.1	0.02*	1.001	1.0001	1.002
Total time trained (grass)	0.002	0.0003	6.02	<0.001***	1.0017	1.0012	1.0023
Total time (futsal)	−0.005	0.001	-5	<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	0.182	0.008	2.3	0.01*	1.1993	1.0298	1.3968

Significance codes: *p<0.05, **p<0.01, ***p<0.001.
Predictors with positive coefficients were identified as being positively related with the count outcome of injury, ie, 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, ie, for a one-unit increase in the identified variable, the likelihood of injury *decreases* by the respective value, assuming all other variables are constant.

For the zero outcome of the model:

Zero-inflated model or logit(Pi)=(−1.25∗match−0.76∗in-season injury)

Name of predictor	Estimated coefficient	SD	Calculated value	P value	OR	2.5%	97.5%
Match	−1.25	0.35	−3.62	<0.001***	0.287	0.146	0.5639
In-season injury	−0.76	0.13	−5.63	<0.001***	0.4694	0.3608	0.6106

Significance codes: *p<0.05, **p<0.01, ***p<0.001.
Predictors with negative coefficients were identified as being negatively related with the zero outcome, ie, 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.

statistical solution for predictor selection in the presence of multicollinearity.

Predictors positively related with injury severity (count part)

An interaction effect between variables is likely, as it is not possible to eliminate multicollinearity. 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.[25–29] However, increased duration on surface type and not surface type alone are linked to increased injury risk.[29 30] 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 variables being positively associated with injury. It is expected that increased participation, facilitated by increased cardiovascular capacity, increases the risk of injury. Therefore, practitioners wishing to mitigate injury risk may consider the frequency and duration of activity on different surface types, alongside the capacity of the player. Other studies have identified positive relationships between previous and subsequent injuries.[5 6] 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.[31] Therefore, accurate injury records are required if previous injury is to be used for prospective injury modelling.[5] A further interaction effect was found between the variables of weight and skinfold thickness, having positive and negative relations to injury, respectively. No consistent anthropometric traits are associated with injury,[23 32–34] although similar results to our study have been identified for players of a lower lean mass having increased risk of hamstring injuries.[35] It may be

BMJ Open Sport & Exercise Medicine: first published as 10.1136/bmjsem-2019-000634 on 14 January 2020. Downloaded from https://bmjopensem.bmj.com on 2 December 2025 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

hypothesised that increased body fat, up to a point, has a protective effect against injury giving the sustained demands on players throughout the season.

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: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.[36] While the predictors of conditioning, total time (conditioning) and total time (futsal) fall outside of the consensus statement,[7] 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 linked determinants 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[37 38] and has been applied to multiple metrics of performance.[38–40] It is therefore unknown how the predictor selection and modelling process would be affected should the index be based on alternate measures of performance, for example, 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.

Predictors related with zero part

The zero component of the ZIP model identifies factors contributing to either an increased or decreased odds of getting a zero, that is, no injury or injury severity of less than 1 day. The variables of in-season injury and match-play were found to have negative relations with this outcome. For a single-unit increase in the events of a match or in-season injury, players were less likely to get a zero, that is, 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 time-loss and non–time-loss injuries. Within the existing literature, more severe injuries are known to be preceded by less severe injuries.[33 41] 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 more severe time-loss injuries.

Match-play was found to have a negative relation with the zero outcome. In comparison with training, matches are known to have a higher rate and number of injuries,[42] possibly explained by the functional demands of a match being higher. This is 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 with alternate models of injury,[5 6] 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.

Limitations of the model

Within our study, the count and zero outcomes were modelled independently through 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.[4–6] This may provide some insight into the limitations of existing models, given that the nature of the data violates the premise of some models, for example, 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.[18] 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.

An assumption of our model was that for the dependent variable, all injury episodes were independent of each other. Despite the absence of a systematic process for informing the circumstances under which previous injury would be causal of future injury, there may be cases of association between previous and future injuries. For some injury cases, this therefore violates an assumption of the ZIP regression model, namely that events are independent, and may explain some of the overdispersion observed. Justification for our selected method was based on the absence of a systematic process linking previous injury to future injuries and to reflect the clinical use of the model and real-world challenges faced by sports and

BMJ Open Sport & Exercise Medicine: first published as 10.1136/bmjsem-2019-000634 on 14 January 2020. Downloaded from https://bmjopensem.bmj.com on 2 December 2025 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

exercise practitioners, that is, the requirement to make unbiased objective sessional decisions around match and training suitability during a progressive season, for players with previous multiple injuries. The absence of a systematic framework for identifying injuries that are dependent or independent of each other remains a challenge for prospective injury modelling.

Within our study, penalised methods for predictor selection, evaluated using ten-fold cross-validation, have been identified as superior when compared with 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. Internal validation of our model was not possible owing to the limited number of count outcomes relative to the number of zeros. Therefore, when attempting to split the data for model development and internal validation purposes, there were an insufficient number of count outcomes within the internal validation set.[43–45] A larger dataset is therefore required to investigate the sensitivity and specificity of our model for comparison against existing models. In addition, 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.[3 5 23 46] 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:chronic workload ratio.[39] 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 composed of different variables. This does not, however, detract from or negate the processes used for predictor and model selection.

CONCLUSION

Penalised methods for predictor selection and use of ZIP 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.

Twitter Fraser Philp @fdphilp

Contributors 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.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Obtained.

Ethics approval Keele University Ethical Review Panel (ERP1237).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. Data are available in a public, open access repository—Code (R) (https://github.com/fraserphilp/Improving-predictor-selection-for-injury-modelling-methods-in-male-footballers).

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID iD

Fraser Philp http://orcid.org/0000-0002-8552-7869

REFERENCES

1 Kiesel K, Plisky PJ, Voight ML. Can serious injury in professional football be predicted by a preseason functional movement screen? *N Am J Sports Phys Ther* 2007;2:147–58.
2 Myer GD, Ford KR, Khoury J, *et al*. Biomechanics laboratory-based prediction algorithm to identify female athletes with high knee loads that increase risk of ACL injury. *Br J Sports Med* 2011;45:245–52.
3 Bahr R, Holme I. Risk factors for sports injuries—a methodological approach. *Br J Sports Med* 2003;37:384–92.
4 Hagglund M*et al*. Methods for epidemiological study of injuries to professional football players: developing the UEFA model. *Br J Sports Med* 2005;39:340–6.
5 Hägglund M, Waldén M, Ekstrand J. Previous injury as a risk factor for injury in elite football: a prospective study over two consecutive seasons. *Br J Sports Med* 2006;40:767–72.
6 Venturelli M, Schena F, Zanolla L, *et al*. Injury risk factors in young soccer players detected by a multivariate survival model. *J Sci Med Sport* 2011;14:293–8.
7 Fuller CW*et al*. Consensus statement on injury definitions and data collection procedures in studies of football (soccer) injuries. *Br J Sports Med* 2006;40:193–201.
8 Philp F. *Validating models of injury risk prediction in football players*. Keele University, 2018.
9 Rossi A, Pappalardo L, Cintia P, *et al*. Effective injury forecasting in soccer with GPs training data and machine learning. *PLoS One* 2018;13:e0201264.
10 Desjardins CD. Modeling zero-Inflated and overdispersed count data: an empirical study of school suspensions. *Int J Exp Educ* 2016;84:449–72.
11 Philp F, Blana D, Chadwick EK, *et al*. Study of the measurement and predictive validity of the functional movement screen. *BMJ Open Sport Exerc Med* 2018;4.
12 Ekstrand J, Krutsch W, Spreco A, *et al*. Time before return to play for the most common injuries in professional football: a 16-year follow-up of the UEFA Elite Club injury study. *Br J Sports Med* 2019. doi:10.1136/bjsports-2019-100666. [Epub ahead of print: 10 Jun 2019].
13 Team RC. *R: a language and environment for statistical computing*, 2013.
14 Sterne JAC, White IR, Carlin JB, *et al*. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009;338:b2393.
15 Sv B, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. *J Stat* 2010:1–68.
16 Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 1992;34:1–14.
17 Steyerberg EW. *Clinical prediction models*. Springer, 2009.
18 Wang Z, Ma S, Wang C-Y. Variable selection for zero-inflated and overdispersed data with application to health care demand in Germany. *Biom J* 2015;57:867–84.
19 Fan Y, Tang CY. Tuning parameter selection in high dimensional penalized likelihood. *J Royal Statistical Soc B* 2013;75:531–52.
20 Wang H, Li B, Leng C. Shrinkage tuning parameter selection with a diverging number of parameters. *J Royal Statistical Soc B* 2009;71:671–83.
21 Zou H, Hastie T. Regularization and variable selection via the elastic net. *J Royal Statistical Soc B* 2005;67:301–20.
22 Tang Y, Xiang L, Zhu Z. Risk factor selection in rate making: EM adaptive LASSO for zero-inflated Poisson regression models. *Risk Anal* 2014;34:1112–27.

BMJ Open Sport & Exercise Medicine: first published as 10.1136/bmjsem-2019-000634 on 14 January 2020. Downloaded from https://bmjopensem.bmj.com on 2 December 2025 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

23 Arnason A, Sigurdsson SB, Gudmundsson A, *et al*. Risk factors for injuries in football. *Am J Sports Med* 2004;32:5–16.
24 Engebretsen AH, Myklebust G, Holme I, *et al*. Intrinsic risk factors for hamstring injuries among male soccer players: a prospective cohort study. *Am J Sports Med* 2010;38:1147–53.
25 Ekstrand J, Hägglund M, Fuller CW. Comparison of injuries sustained on artificial turf and grass by male and female elite football players. *Scand J Med Sci Sports* 2011;21:824–32.
26 Ekstrand J, Timpka T, Hägglund M, *et al*. Risk of injury in elite football played on artificial turf versus natural grass: a prospective two-cohort study * commentary. *Br J Sports Med* 2006;40:975–80.
27 Fuller CW, Dick RW, Corlette J, *et al*. 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* 2007;41:i27–32.
28 Fuller CW, Dick RW, Corlette J, *et al*. 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* 2007;41:i20–6.
29 Kristenson K, Bjørneboe J, Waldén M, *et al*. 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* 2013;47:775–81.
30 Aoki H, Kohno T, Fujiya H, *et al*. Incidence of injury among adolescent soccer players: a comparative study of artificial and natural grass turfs. *Clin J Sport Med* 2010;20:1–7.
31 Junge A, Dvorak J. Soccer injuries: a review on incidence and prevention. *Sports Med* 2004;34:929–38.
32 Frisch A, Urhausen A, Seil R, *et al*. Association between preseason functional tests and injuries in youth football: a prospective follow-up. *Scand J Med Sci Sports* 2011;21:e468–76.
33 Gajhede-Knudsen M, Ekstrand J, Magnusson H, *et al*. 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* 2013;47:763–8.
34 Fousekis K, Tsepis E, Poulmedis P, *et al*. Intrinsic risk factors of non-contact quadriceps and hamstring strains in soccer: a prospective study of 100 professional players. *Br J Sports Med* 2011;45:709–14.
35 Henderson G, Barnes CA, Portas MD. Factors associated with increased propensity for hamstring injury in English Premier League soccer players. *J Sci Med Sport* 2010;13:397–402.
36 van der Horst N, Smits D-W, Petersen J, *et al*. The preventive effect of the Nordic hamstring exercise on hamstring injuries in amateur soccer players. *Am J Sports Med* 2015;43:1316–23.
37 Malone S, Owen A, Newton M, *et al*. The acute:chonic workload ratio in relation to injury risk in professional soccer. *J Sci Med Sport* 2017;20:561–5.
38 Hulin BT, Gabbett TJ, Lawson DW, *et al*. The acute:chronic workload ratio predicts injury: high chronic workload may decrease injury risk in elite rugby league players. *Br J Sports Med* 2016;50:231–6.
39 Hulin BT, Gabbett TJ, Blanch P, *et al*. Spikes in acute workload are associated with increased injury risk in elite cricket fast bowlers. *Br J Sports Med* 2014;48:708–12.
40 Møller M, Nielsen RO, Attermann J, *et al*. Handball load and shoulder injury rate: a 31-week cohort study of 679 elite youth handball players. *Br J Sports Med* 2017;51:231–7.
41 Ekstrand J, Gillquist J. The Avoidability of soccer injuries. *Int J Sports Med* 1983;04:124–8.
42 Ekstrand J, Hägglund M, Waldén M. Injury incidence and injury patterns in professional football: the UEFA injury study. *Br J Sports Med* 2011;45:553–8.
43 Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMC Med* 2015;13:1.
44 Riley RD, Hayden JA, Steyerberg EW, *et al*. Prognosis research strategy (progress) 2: prognostic factor research. *PLoS Med* 2013;10:e1001380.
45 Steyerberg EW, Moons KGM, van der Windt DA, *et al*. Prognosis research strategy (progress) 3: prognostic model research. *PLoS Med* 2013;10:e1001381.
46 Hägglund M, Waldén M. Risk factors for acute knee injury in female youth football. *Knee Surg Sports Traumatol Arthrosc* 2016;24:737–46.

BMJ Open Sport & Exercise Medicine: first published as 10.1136/bmjsem-2019-000634 on 14 January 2020. Downloaded from https://bmjopensem.bmj.com on 2 December 2025 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.