1 **Modelling training loads and injuries: the dangers of discretization**

2

3   David L. Carey[1,2], Kay M. Crossley[1], Rod Whiteley[3], Andrea Mosler[1,3], Kok-Leong Ong[4],

4   Justin Crow[2], Meg E. Morris[1,5]

5

6   [1]La Trobe Sport and Exercise Medicine Research Centre, College of Science, Health and

7   Engineering, La Trobe University, Melbourne, Australia

8   [2]Essendon Football Club, Melbourne, Australia

9   [3]Rehabilitation Department, Aspetar Orthopedic and Sports Medicine Hospital, Doha, Qatar

10 [4]Research Centre for Data Analytics and Cognition, La Trobe University, Melbourne,

11 Australia

12 [5]Healthscope, Northpark Private Hospital, Melbourne, Australia

13

14 **Corresponding author:** David L. Carey (d.carey@latrobe.edu.au), La Trobe Sport and

15 Exercise Medicine Research Centre, College of Science, Health and Engineering, La Trobe

16 University, Plenty Road & Kingsbury Drive, Melbourne, Australia 3086.

17

18

19

ABSTRACT

Purpose: To evaluate common modelling strategies in training load and injury risk research when modelling continuous variables and interpreting continuous risk estimates; and present improved modelling strategies.

Method: Workload data were pooled from Australian football (n=2,550) and soccer (n=23,742) populations to create a representative sample of acute:chronic workload ratio observations for team sports. Injuries were simulated in the data using three pre-defined risk profiles (U-shaped, flat and S-shaped). One-hundred datasets were simulated with sample sizes of 1000 and 5000 observations. Discrete modelling methods were compared to continuous methods (spline regression and fractional polynomials) for their ability to fit the defined risk profiles. Models were evaluated using measures of discrimination (area under ROC curve) and calibration (Brier score, logarithmic scoring).

Results: Discrete models were inferior to continuous methods for fitting the true injury risk profiles in the data. Discrete methods had higher false discovery rates (16-21%) than continuous methods (3-7%). Evaluating models using the area under the receiver operator characteristic (ROC) curve incorrectly identified discrete models as superior in over 30% of simulations. Brier and logarithmic scoring was more suited to assessing model performance with less than 6% discrete model selection rate.

Conclusions: Many studies on the relationship between training loads and injury that have used regression modelling have significant limitations due to improper discretization of continuous variables and risk estimates. Continuous methods are more suited to modelling the relationship between training load and injury. Comparing injury risk models using ROC curves can lead to inferior model selection. Measures of calibration are more informative when judging the utility of injury risk models.

Keywords: *acute:chronic workload ratio; injury risk; ROC curves; calibration*

INTRODUCTION

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

Discretization is the practice of transforming continuous data into discrete categories, and is a prevalent methodology in studies of training load and injury risk (4-10). Discretization methods in sports medicine research include median splits (5, 7), percentiles (5, 6, 13), z-score categories (4, 7), and arbitrary bins (8-10). These methodologies have not been critically examined in the context of modelling training loads and injuries. Discretization of continuous covariates in risk models has been criticized in other fields (16-19). Discretization of a continuous risk factor into categories assumes that each individual within that category has equal risk. For example, if cumulative training load is split into low, medium and high categories using percentiles then it is assumed that each athlete in the high category has identical risk irrespective of how broad the category is (i.e. an athlete at the 67th percentile is considered to be at the same risk as one at the 99th percentile). This practice causes a loss of information because within-category variation is ignored (17). The loss of information lowers the statistical power of the study and may reduce the ability to detect relationships between variables, increasing the likelihood of a false negative result (17, 18, 20). Discretization can also lead to inflated false discovery rates (17, 18). It is common for studies using discretization to analyze results by choosing a reference category and making multiple comparisons to each other category, increasing the chance of finding a significant result (17). It has also been shown by Wainer et al. (21) that categorization of continuous variables can make trends appear in otherwise unrelated data if there is freedom to choose the boundaries of the categories. Modelling methods that allow risk factors to vary continuously such as cubic regression splines and fractional polynomials have therefore been advocated as appropriate alternatives to discretization for modelling non-linear risk profiles in epidemiology (16, 20).

78   The increase in studies investigating training load as a risk factor for injury has been

79   accompanied by an increase in studies exploring injury prediction (5, 11, 12, 14, 22, 23).

80   Injury prediction models have been evaluated and compared using metrics such as sensitivity,

81   specificity or area under the receiver operator characteristic curve (AUC) (5, 11, 12, 14, 22,

82   23). These scoring metrics are designed to evaluate binary predictions (i.e. injury or no-

83   injury) and look at how often the model predictions match the actual outcomes (24). In a

84   practical setting, where there is a clinician or coach to synthesize other sources of information

85   to make a contextualized judgement, a model would not be expected to make a yes/no

86   decision. In this scenario it could be more informative to evaluate injury risk models using

87   measures of calibration (19, 25). Calibration refers to how well a model is able to estimate

88   the probability of an event (24).

89

90   In this study we critically evaluated the modelling and evaluation methodologies found in the

91   existing literature on the relationships between training load and injury. Training load data

92   collected from Australian football (6) and soccer (26) were used to generate a set of

93   hypothetical datasets with known injury risk profiles (27). Discrete risk models using z-score,

94   percentile, and arbitrary binning methods (4-10) were compared to continuous methods;

95   regression splines and fractional polynomials (16, 20). Models were evaluated using

96   measures of discrimination (AUC) and calibration (Brier score) to assess which metrics were

97   the most informative for assessing the utility of risk models (24).

98

99   **METHODS**

100   **Training load data**

101   The acute:chronic workload ratio (ACWR) is a relative training load variable calculated by

102   dividing an athlete's acute workload (typically 1 week) by their chronic workload (typically 4

103   weeks) (2, 27). It is a bounded continuous variable that has been studied extensively as an

104   injury risk factor (2-14). ACWR data were pooled from two studies on separate male

105   populations; a two season study at a single Australian Football club (6) (n = 2550), and a two

106   season study of 17 soccer teams in the Qatar Stars League (26) (n = 23742). One week acute

107   and four week chronic periods (overlapping) were used for both data sets. Total distance was

108   used as the load variable in the Australian football data set and training/match duration in the

109   soccer data set (the only available load metric). Combined, these data had a mean and

110   standard deviation of 1.05 and 0.42; similar to values reported in previous studies (5, 7, 8)

111   (see Figure, Supplemental Digital Content 1, histogram of ACWR values). The pooling of

112 data from independent sources was done to ensure the distribution of values used in the
113 simulations was as representative as possible (i.e. it is a good approximation of what a
114 researcher could expect to collect in a hypothetical future study). Alternate ACWR
115 calculation methods that use exponentially weighted averages (28) or de-couple the acute and
116 chronic time windows (29) have been proposed. These modifications likely change the
117 distribution of ACWR values (e.g. de-coupling causes the ACWR to become unbounded).
118 Despite this, each method still produces a continuous variable and the investigation into the
119 effects of discretization in this study remain relevant irrespective of the ACWR calculation
120 method. Ethical approval for this study was obtained from the Shafallah Medical Genetics
121 Centre, Approval number: 2012–017 and the La Trobe University Faculty of Health Sciences
122 Human Ethics Committee (FHEC14/233). Informed consent was obtained from the
123 participating teams for the analysis of de-identified data.

124

125 **Injury risk profiles**
126 Traditional research designs collect data and build models in an attempt to estimate the true
127 relationship between variables of interest (e.g. ACWR and injury). In order to evaluate
128 different modelling approaches we have used a different strategy. Artificial injuries were
129 inserted into existing training load data based on pre-defined risk profiles. This enabled us to
130 compare different models based on how well they were able to recover the true relationship
131 in the data. Three pre-defined theoretical risk profiles were considered (figure 1).

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

134 - **Flat:** To represent the null hypothesis that ACWR does not influence injury risk
135   (every observation poses a uniform 5% injury risk).

136 - **S-shaped:** An alternative risk profile that has injury risk as constant (2%) for ACWR
137   < 1 then rises sharply to very high injury risk.

138 Details on the mathematical form of risk curves can be found in the supplementary text (see
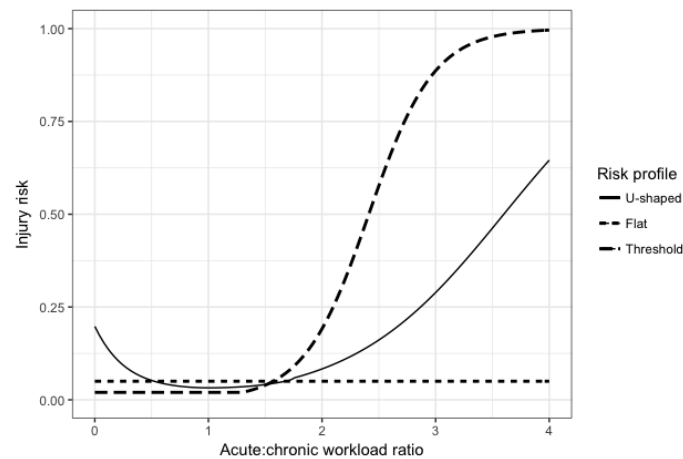139 text, Supplemental Digital Content 2, equations of risk profiles).

140

Figure 1: Theoretical risk profiles used to simulate injuries.

Simulating study data

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

- Choose a sample size (N_s) and randomly choose N_s observations of ACWR from the pooled data distribution (see Figure, Supplemental Digital Content 1, histogram of ACWR values).

- Assign an injury probability (p_i) to each observation using one of the pre-defined theoretical risk profiles (figure 1).

- Randomly generate injuries by treating each observation as a Bernoulli trial with probability of injury p_i. Simply, this means for an observation with injury risk of 20% we randomly assigned it an injury or no-injury label with probability 0.2 and 0.8 respectively.

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

Training load – injury models

Two types of modelling approach were considered; discrete and continuous. Discrete models were defined as those that applied a discretization strategy to the ACWR values *before*

167 modelling them against injury incidence. We considered three different discretization
168 methods to reflect those found in the existing literature (4, 5, 7-10, 13).

169 • **D1:** Normalize ACWR values (z-score) then split into 7 categories using cut-points:
170 $\{-\infty, -2, -1, 0, 1, 2, 3, \infty\}$ (4, 7).

171 • **D2:** Split into 5 quantiles (5, 6, 13).

172 • **D3:** Split the ACWR into 5 categories using the cut-points: $\{0, 1, 1.35, 1.5, 2, \infty\}$
173 (10).

174 Following discretization, ACWR was modelled against injury incidence using binary logistic
175 regression, with the central group used as the reference level. This method of analysis
176 replicates that commonly used in previous studies (8-10, 13).

177

178 To contrast the discrete models, two continuous modelling methodologies (C1 and C2) were
179 considered. The continuous methods apply a transformation to the independent variable
180 (ACWR) within the logistic regression. This allows for non-linear relationships that vary
181 continuously to be modelled.

182 • **C1: Restricted cubic splines** model relationships by subdividing the range of values
183 of the covariate (at locations called knots), and fitting a cubic polynomial between
184 each pair of knots. The polynomials are constrained to join smoothly at each knot and
185 to be linear in the two outermost regions (19). Restricted cubic splines are a common
186 method of analysis in epidemiological studies of nonlinear dose-response
187 relationships (16, 19, 20). Spline models were fitted in R using the *splines* package
188 (30). Spline regression models were constructed with three internal knots placed at
189 equally spaced percentiles (19, 20). The number of knots was chosen a-priori in this
190 study but in general can be chosen by comparing multiple options using an objective
191 criterion (e.g. Akaike information criterion (AIC)) (19).

192 • **C2: Fractional polynomials** are a flexible method of modelling nonlinear,
193 continuous relationships. Fractional polynomials consider a combination of candidate
194 functions and select a final model after a series of tests for nonlinearity and
195 complexity (31). A potential benefit of fractional polynomials over cubic splines is
196 that they are more interpretable. The final model can be described by a closed form
197 equation and offers potential insight into the underlying relationship. Their drawback
198 is that they are a global model (i.e. they fit the entire range of data with a single

199   function), and therefore cannot fit local features as well as splines. Fractional
200   polynomial models were fitted in R using the *mfp* package (30, 32).

201 Presently, few studies of the relationship between ACWR and injury have used modelling

202 methods that allow for non-linear trends and avoid discretization of the ACWR. An

203 implementation of each modelling method considered is included in the supplementary code

204 (see text, Supplemental Digital Content 3, simulation code).

205

206 Each of the models (discrete and continuous) were used to produce estimates of injury risk

207 for each ACWR observation in the simulated data sets. This replicates a study design from a

208 team sport environment where workload risk factor and injury outcomes are recorded daily.

209

210 **Evaluating injury models**

211 *Comparison between true and modelled risk curves*

212 A direct comparison can be made between the modelled risk profile and the true risk profile

213 in this study because the function used to simulate the injuries was pre-defined (i.e. it is

214 exactly known, as shown in figure 1). Root mean square error (RMSE) (24) was calculated

215 for the difference between the true risk and predicted risk for each observation in each

216 simulated study. This provides a measure of how well the modelling procedure was able to

217 recover the true risk profile used to generate the data.

218

219 *False discovery and false rejection rates*

220 The flat injury risk profile was used to estimate the false discovery rate for each modelling

221 approach (figure 1). Data simulated under the flat profile contained no association between

222 ACWR and injury risk. Therefore, any simulated study finding a significant relationship in

223 the data could be considered a false discovery (Type I error). Significance testing for discrete

224 models (D1, D2, D3) was performed by comparing the reference ACWR level to all other

225 levels in the discretized ACWR (4, 5, 7-10, 13). A simulation was deemed to have a

226 significant finding if any of the 95% confidence intervals for the odds ratios did not contain

227 1. Significance testing for spline regression (C1) and fractional polynomials (C2) was

228 performed by comparing to a null model using the likelihood ratio test with $\alpha=0.05$ (32).

229

230 False rejection rates (Type II error) were estimated for each model by counting the number of

231 times no significant result was found when the data were simulated with a U-shaped or S-

shaped risk profile. Discretizing continuous variables causes a decrease in statistical power (17, 18), potentially causing the false rejection rates of discrete models to increase.

Receiver operator characteristic

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

Calibration

Calibration is a measure of how well a model is able to estimate the probability of an event. It can be assessed visually by constructing calibration curves (19, 33). Calibration curves show how closely the predicted probabilities match the observed event rates (i.e. for observations estimated to have injury risk of 20% - was the actual injury incidence rate on those days around 20%?). Calibration can also be assessed quantitatively by computing the Brier score or logarithmic scoring rule (19). In the case of a binary outcome variable the Brier score is calculated as the square of the probability assigned to the incorrect class (e.g. if the model predicts injury with probability 0.2, and there wasn't an injury, the Brier score would be 0.2^2 = 0.04, but if there *was* an injury then the score would be 0.8^2 = 0.16). A lower Brier score indicates a better model. The logarithmic scoring rule is evaluated by taking the natural logarithm of the probability assigned to the correct class (e.g. a predicted injury probability of 0.2 and no injury would score log(0.8) = -0.22, and if there was an injury would score log(0.2) = -1.61). A higher score indicates better probability estimates. Logarithmic scoring may be more appropriate than the Brier score in the case of rare event estimation (34). Brier and logarithmic scores were estimated for each model and simulated study using 10-fold cross validation (24).

Longitudinal models of training loads and injury

For clarity of message in the previous sections we have simulated ACWR and injury data with no correlation structure and used logistic regression assuming independence of

266   observations to illustrate the effects of discretization. However, training load monitoring data

267   collected from sporting teams often consists of repeated measurements taken from the same

268   athletes. It is therefore possible that the observations from the same athletes will be

269   correlated. To investigate the effects of this correlation on injury risk modelling we simulated

270   longitudinal training load data sets using the *SimCorMultRes* package (35) (see text,

271   Supplemental Digital Content 3, simulation code). Injuries were simulated in the data by

272   defining a marginal risk profile and specifying a within-subject correlation strength (35).

273   Four longitudinal data sets were simulated (100 times each) to investigate the effects of

274   different sample sizes, within-subject correlations and marginal risk profiles. The first

275   simulated 50 observations from 20 participants with a U-shaped marginal risk (figure 1) and

276   a within-subject correlation of 0.1. The second increased the correlation strength to 0.7. The

277   third considered a larger sample size of 100 observations from 50 participants. The fourth

278   considered the effect of reducing the strength of the marginal risk by reducing the injury risk

279   by a factor of ½ for each ACWR value.

280

281   Each longitudinal data set was analyzed using naïve logistic regression (i.e. assuming

282   independence of observations) and generalized estimating equations (GEE) (36). GEE

283   models have been used in previous studies of training load and injury (5, 12, 14). GEE

284   models were fitted using the R package *geepack* (37) using a binomial link and exchangeable

285   working correlation structure. Both analysis methods allowed for the relationship between

286   ACWR and injury risk to vary continuously using restricted cubic splines (as previously

287   described). Modelling approaches were compared for their ability to recover the pre-defined

288   marginal effect of ACWR on injury risk using RMSE. Additionally, significance testing was

289   performed for each simulated study result by comparing to a null model using a likelihood

290   ratio test (see text, Supplemental Digital Content 3, simulation code).

291

292   **RESULTS**

293   **Simulated studies**

294   Details of the simulated studies (injury summary statistics) are found in supplementary table

295   1 (see Table, Supplemental Digital Content 4, simulated injury statistics).

296

297   **Comparison between true and modelled risk profiles**

298   A visual inspection of how well each modelling procedure was able to recover the true risk

299   profile used to generate the data is shown in figure 2. It is clear that models that discretized

300 the ACWR (figure 2(D1-3)) are unable to fully capture the U-shaped relationship.

301 Continuous modelling methods C1 (spline regression) and C2 (fractional polynomials) fared

302 much better at fitting the true risk profile (see Figures, Supplementary Digital Content 5, S-

303 shaped risk, and Supplemental Digital Content 6, flat risk).
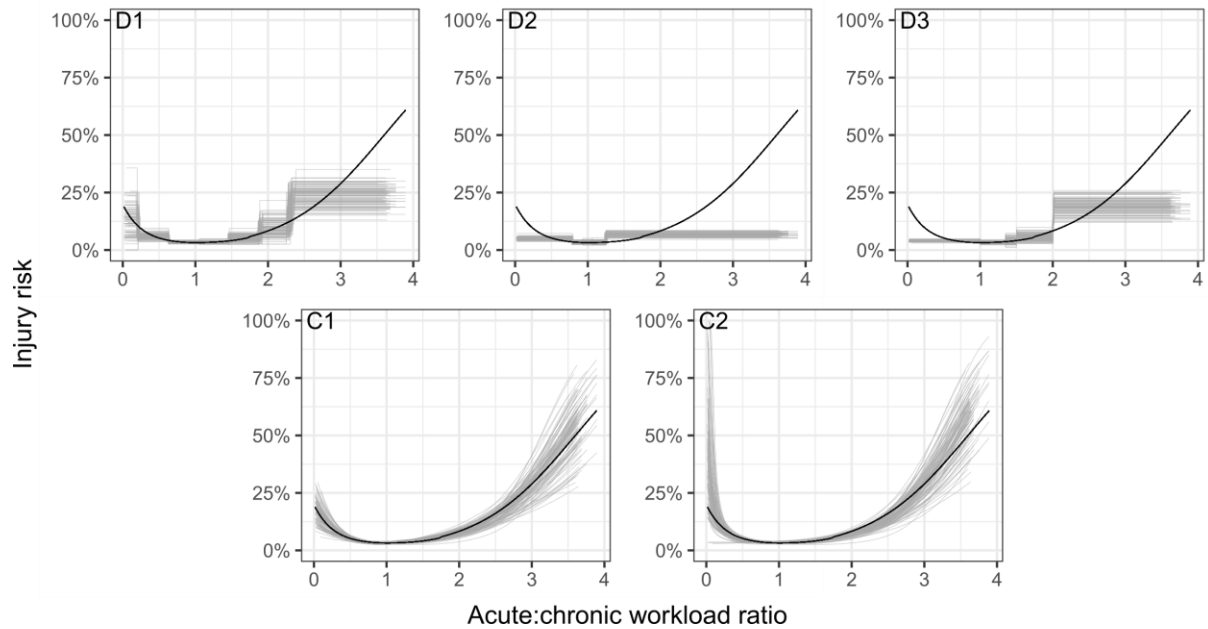
304

305

306 Figure 2: Comparison of 100 simulated study results (N_s = 5000 and U-shaped risk) analyzed using

307 discrete models (D1, D2, D3) and continuous models (C1, C2). Solid line represents the true risk

308 profile used to generate the data and each grey line represents one simulated study result.

309

310 The RMSE performance of each modelling strategy under the different simulation parameters

311 is shown in figure 3. Continuous modelling methods (C1, C2) had noticeably lower RMSE

312 for data generated using a U or S-shaped injury risk profile (particularly in larger simulated

313 studies with N_s = 5000). The difference between discrete and continuous methods was less

314 pronounced for the flat injury risk profile. In general, the total error and variance in error for

315 each model tended to decrease when the simulated sample size increased from 1000 to 5000
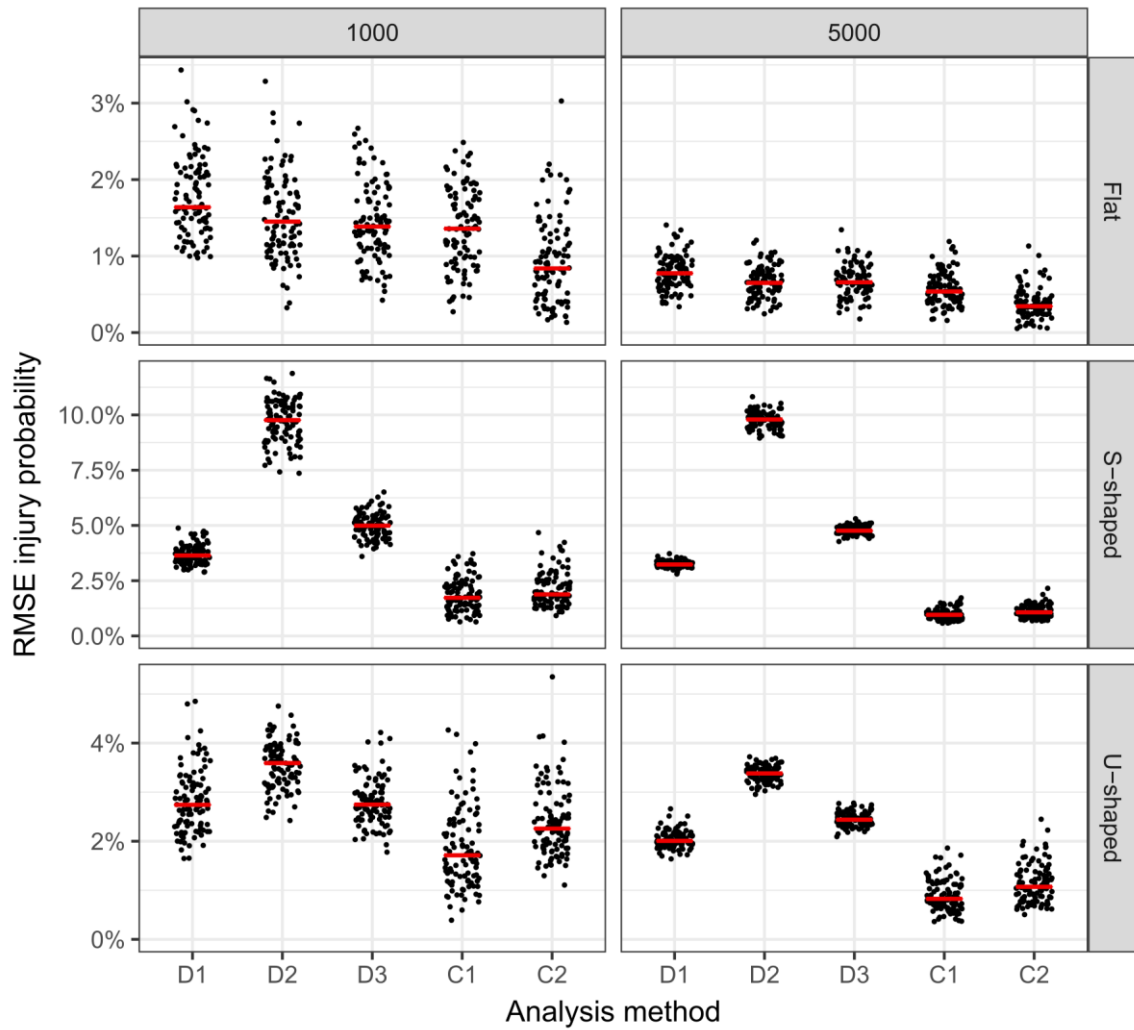
316 observations.

317

11

Figure 3: Root mean square error of model probability estimates for 100 trials of each theoretical risk profile and sample size (red bar = median).

False discovery rates

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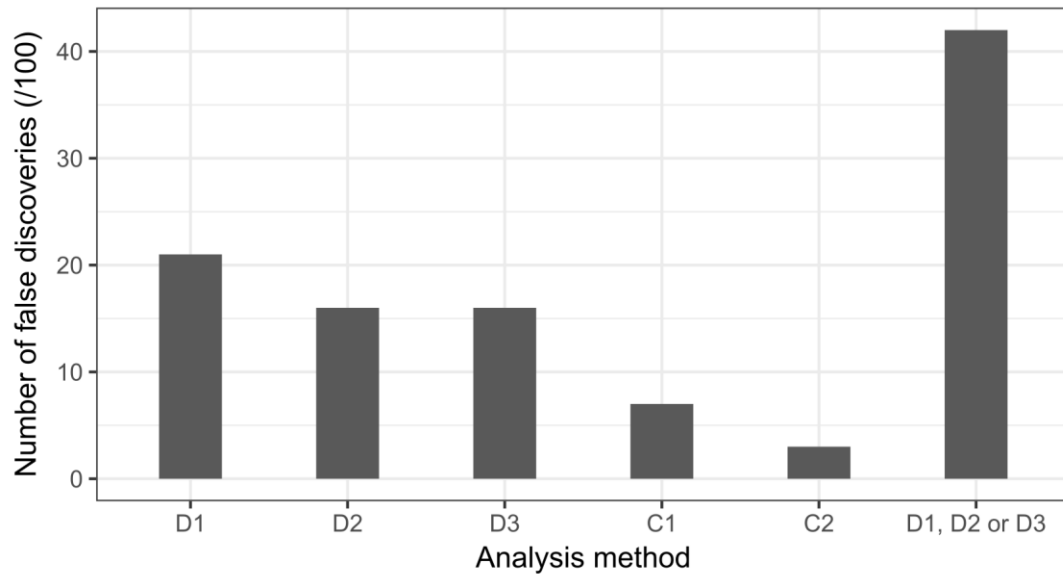
Figure 4: False discovery rates (out of 100 simulated studies with $N_s = 5000$ and flat risk profile).

Choice of reference level

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

False rejection rates

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

Receiver operator characteristics

13

354   Area under the ROC curve was estimated for each of the 100 simulated studies using 10-fold

355   cross validation (24). The continuous analysis methods had higher median AUC values but

356   did not clearly outperform discrete methods under this evaluation metric. If AUC was used to

357   select the best performing model in each simulation we found that one of the discrete models

358   was chosen on 38/100 and 31/100 occasions for U-shaped and S-shaped risk respectively

359   (table 1).

360

361   Table 1: Comparison of model selection rates using area under the ROC curve (AUC), Brier score and

362   logarithmic scoring as the evaluation metric ($N_s = 5000$).

Method ID	Number of times selected as best model (/100 simulated studies)					
	U-shaped risk			S-shaped risk		
	AUC	Brier	Logarithmic	AUC	Brier	Logarithmic
D1	28	6	3	15	0	1
D2	2	0	0	0	0	0
D3	8	0	0	16	0	0
C1	35	80	70	26	73	75
C2	27	14	27	43	27	24

363

364   **Calibration**

365   To compare with ROC curves, Brier and logarithmic scores were estimated for each model

366   using 10-fold cross validation (19, 24). When the Brier score was used to select the best

367   performing model in each simulation, discrete models were chosen on only 6/100 and 0/100

368   occasions for U-shaped and S-shaped risk respectively (table 1). When logarithmic scoring

369   was used the rates were 3/100 and 1/100. Brier and logarithmic scoring favored the

370   continuous methods far more than evaluation with ROC curves.

371

372   Calibration curves offer a way to visually evaluate injury risk models (figure 5). A calibration

373   curve shows the relationship between the predicted probabilities and actual event occurrence

374   rate (perfect calibration is represented by the diagonal line). An exemplar set of calibration

375   curves are shown in figure 5 (one simulated study with U-shaped risk and $N_s = 5000$). Ideally

376   a well calibrated risk model will have a curve that is close to the diagonal line and covers a

377   large range of probabilities (i.e. has confidence in identifying both high and low risk

378   scenarios). Discrete model D2 provided little information other than the baseline injury rate.
379   Model D3 did not appear to be well calibrated. Models D1, C1 and C2 were well calibrated
380   (close to diagonal line) however the continuous methods covered a much larger range of
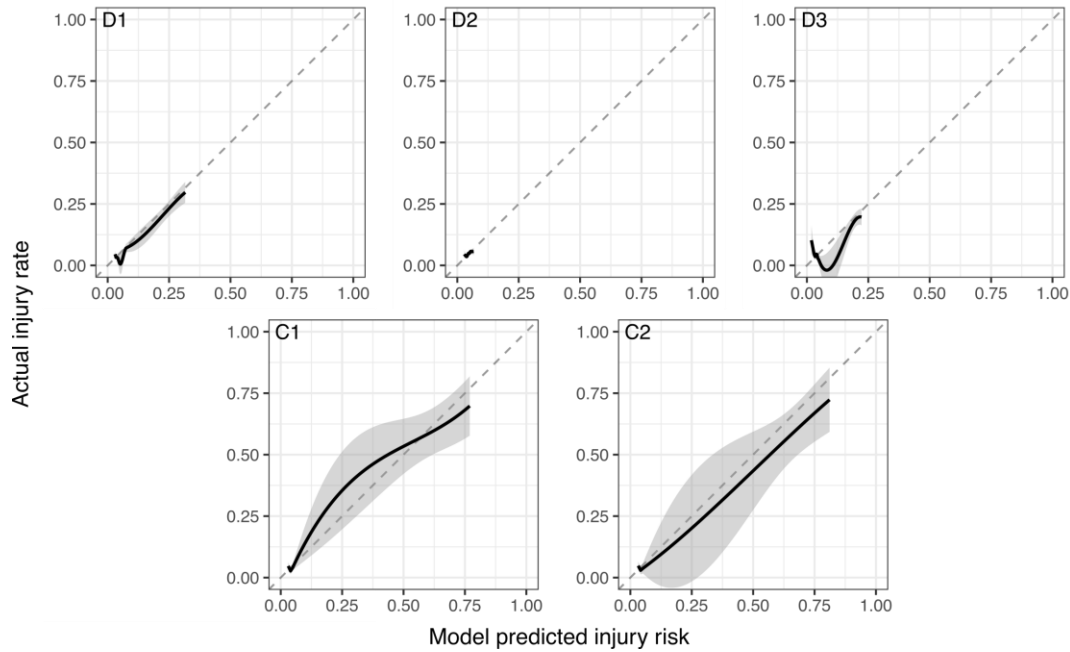381   probabilities.

382

383
384   Figure 5: Comparison of cross-validated calibration curves from a single simulated study ($N_s = 5000$
385   and U-shaped risk) analyzed using discrete (D1, D2, D3) and continuous models (C1, C2). Diagonal
386   line represents perfect calibration and shaded area represents 95% CI.

387

Longitudinal data models

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

394

395   Table 2: Comparison of logistic regression and GEE modelling for longitudinal data.

Marginal risk profile	Study size (observations x participants)	Correlation strength	Significant results (/100)		Median RMSE (IQR)	
			Logistic	GEE	Logistic	GEE

U-shaped	50 x 20	0.1	84	95	0.020 (0.013-0.025)	0.021 (0.014-0.025)
U-shaped	50 x 20	0.7	86	97	0.030 (0.024-0.037)	0.028 (0.021-0.035)
U-shaped	100 x 50	0.1	100	100	0.009 (0.007-0.012)	0.009 (0.007-0.011)
U-shaped (dilated ½)	50 x 20	0.1	53	82	0.013 (0.011-0.017)	0.014 (0.011-0.017)

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

DISCUSSION

Discrete versus continuous modelling strategies

Discrete models showed limited ability to capture the risk profiles used to generate the simulation data (figure 2-3). Discretization forced the models to fit an unrealistic and discontinuous step profile to the data (figure 2 and Supplemental Digital Content 5, S-shaped risk). This illustrates how discretization of continuous risk factors can lead to inaccurate estimation of effects (17, 20). Figure 2 shows how using percentile splits (method D2) groups a large range of ACWR values together and provides an inaccurate estimated effect that is far lower than the true risk for ACWR values greater than 2. Similarly, the ACWR categories used in method D3 assume homogeneity of risk over the range 0 to 1 leading to an estimated effect that cannot capture the rise in risk seen for small ACWR values. Our simulations suggest that the discrete methods found in the current literature (4-8, 10, 13) are unsuited to modelling the continuous U-shaped risk profile between ACWR and injury proposed in the literature (2, 27).

420 Continuous modelling methods (spline regression and fractional polynomials) were better
421 suited to fitting the non-linear risk profiles (U-shaped and S-shaped) and provided more
422 accurate estimated effects. This was demonstrated by lower RMSE scores (figure 3) and also
423 confirmed visually by the 100 simulations shown in figure 2. These findings align with
424 recommendations from other fields that continuous modelling methods are preferable to
425 discretization (17, 20). Future studies may benefit from using continuous modelling methods
426 instead of discretizing continuous training load variables when analyzing their relationship to
427 injury.
428

429 **False discovery rates**
430 Data generated under the assumption that ACWR had no relationship to injury risk (figure 1,
431 flat risk profile) was used to estimate the false discovery rate for each modelling approach.
432 False discovery rates were inflated by using discrete models (16-21%) (figure 4). Splitting
433 the ACWR into multiple categories before modelling leads to multiple comparisons between
434 groups and may explain the higher false discovery rates (17, 18). Discrete method D1 used
435 the most categories (7 groups) and had the highest false discovery rate (21%). A secondary
436 issue was the choice of reference level when categorical predictors are used in generalized
437 linear models (e.g. logistic or Poisson regression). Discrete model D1 had 21/100 false
438 discoveries when the central ACWR category was used as the reference but only 11/110 if
439 the highest was used.
440

441 There is currently no consensus in the literature regarding the discretization strategy or choice
442 of reference level when modelling ACWR and injury risk (5, 8, 10, 13). The apparent
443 freedom of choice of discretization and reference level may have caused highly inflated false
444 discovery rates in previous studies (38). When a choice of only three methods was considered
445 in our simulations false discovery rates were as high as 42% (figure 4). Continuous modelling
446 methods do not require choosing a reference level and do not suffer from multiple
447 comparisons between predictor categories. Spline regression and fractional polynomials had
448 substantially lower false discovery rates (7% and 3%).
449

450 **False rejection rates**
451 Models that transformed the continuous ACWR into discrete categories showed higher false
452 rejection rates in the simulated studies (see Table, Supplemental Digital Content 7, false
453 rejection rates). This aligns with findings from other studies that discretization lowers

17

454   statistical power (17, 18, 20). Simulations using a larger sample size (N_s=5000) were not as

455   prone to false rejections, highlighting the benefits of larger sample sizes. The negative

456   consequences of discretization on statistical power are particularly relevant for research in

457   elite sport cohorts where sample sizes are often constrained.

458

459   **Evaluating injury risk models**

460   *ROC curves*

461   Comparing models using the area under the ROC curve did not always identify that

462   continuous methods were better fits to the risk profiles (table 1). RMSE scores showed that

463   continuous methods were clearly superior when modelling U-shaped or S-shaped risk profiles

464   when a sample size of 5000 observations was used (figure 3). Despite this, AUC incorrectly

465   identified discrete methods as superior in 38 and 31/100 simulations for U and S-shaped risk

466   (table 1). This suggests that using AUC as the sole evaluation metric when selecting injury

467   prediction models (11, 12, 14) runs the risk of selecting an inferior model.

468

469   A ROC curve is constructed by sampling through the possible decision thresholds that could

470   be applied (i.e. cut points where the models makes an injury or no-injury prediction). This

471   may not realistically represent the purpose of the model if it to be used for risk estimation. If

472   the output of the model is used along with context and clinical judgement, and not required to

473   make a binary decision, then AUC may not be an appropriate evaluation metric (25). ROC

474   curves also assume that false positive errors and false negative errors are of equal

475   consequence (39). This is likely not the case when a false negative means an injured athlete

476   and a false positive may be a modified or missed training session. We suggest that ROC

477   curves in isolation are insufficient to evaluate the performance of injury prediction models.

478

479   *Probabilistic scoring and calibration curves*

480   Evaluating the modelling strategies with Brier scores and logarithmic scoring strongly

481   favored the continuous models (table 1). Discrete models were selected in only 6/100 and

482   0/100 simulations using the Brier score and 3/100 and 1/100 when using logarithmic scoring.

483   These provide a much closer reflection of the RMSE scores (ground truth) than evaluating

484   models with AUC. The Brier and logarithmic scores are probabilistic scoring rules designed

485   to evaluate probability estimates (19) and are therefore better suited to assessing injury risk

486   models. We suggest that Brier scores, logarithmic scoring, or another comparable

487   probabilistic scoring rule (34) be included in future studies to compare injury risk models.

18

488

489   Calibration curves (figure 5) provided an informative visualization of the performance of

490   injury risk models (33). They showed how closely the risk estimates of each model matched

491   the observed injury rates and how well each model discriminated between high and low risk

492   instances. Figure 5 clearly shows that continuous models gave more informative probability

493   estimates (closer to the observed event rates and over a larger range of values) than the

494   discrete models. Calibration curves show absolute risks and thus may be a more important

495   result for clinicians and decision makers (40).

496

497   **Longitudinal models**

498   Extending the simulation study to include correlated within-individual observations showed

499   the negative effects of incorrectly assuming independence between repeated measurements.

500   The naïve logistic regression approach had higher false rejection rates than a GEE approach

501   (table 2). Assuming independence can cause the standard errors for time varying covariates to

502   be overestimated (41) and may have been the cause of the inflated false rejection rates.

503   When the strength of the 'signal' in the data was decreased the difference between logistic

504   and GEE approaches became more pronounced, and the naïve logistic approach had very

505   high false rejection rate (47/100). This highlights the importance of accounting for correlated

506   observations when modelling longitudinal training load data, particularly if the expected

507   strength of signal in the data is small.

508

509   Both longitudinal modelling approaches showed similar ability to recover the true marginal

510   risk profile. This is likely because the parameter estimates from logistic regression and GEE

511   models are generally very similar (41). In all simulations, larger sample sizes improved the

512   accuracy of model estimated effects, suggesting the potential benefits of collaborative studies

513   with large sample sizes.

514

515   **Limitations and extensions**

516   Restricted cubic spline regression and fractional polynomials were considered as the

517   alternative modelling methods in this study. Although they are common approaches for

518   modelling non-linear relationships (16, 20) they are not the only possible approaches. A

519   number of other non-parametric and semi-parametric methods may have been suitable (e.g.

520   locally weighted regression, generalized additive models, and smoothing splines) (19).

521

522   We did not consider multivariable modelling and used only a single covariate (ACWR) in our

523   simulations. This was done for clarity of the message. The issues caused by discretization are

524   equally problematic in multivariable modelling. Spline and fractional polynomial techniques

525   can still be used when there is more than one covariate to allow for proper modelling of

526   continuous variables (19, 32). For example, recent studies have investigated the effect of

527   ACWR on injury risk moderated by absolute chronic workload dichotomized using a median

528   split (4, 5, 7). This dichotomization removes a significant amount of variation in the data,

529   leading to decreased statistical power and inaccurate estimation of effects (17, 20). It is

530   possible, and we would suggest more appropriate, to avoid discretization and model both risk

531   factors continuously using a technique such as restricted cubic surfaces (19). This study did

532   not consider time-to-event approaches for modelling training loads and injury (e.g. survival

533   analysis and cox regression (42)). Discretization of baseline or time-varying covariates can

534   have similar consequences on statistical power and estimated effects in these contexts and

535   continuous approaches are advised (19).

536

537   **CONCLUSION**

538   Modelling methods that discretize continuous risk factors are inappropriate for studying the

539   relationship between training loads and injuries. Discrete models have inflated false

540   discovery and false rejection rates and are unsuited to fitting non-linear risk profiles. Strong

541   justification is required for research that chooses a discrete approach and we suggest avoiding

542   discretization and modelling relationships with continuous methods such as spline regression

543   or fractional polynomials. Accounting for correlated observations in longitudinal training

544   data decreases the risk of false rejection. Evaluating injury risk models using ROC curves

545   may not reflect their practical use and may lead to inferior model selection. Probabilistic

546   scoring methods such as Brier scores, logarithmic scoring and calibration curves may be

547   more informative when assessing injury models.

548

554

555   **REFERENCES**

20

556 1.   Soligard T, Schwellnus M, Alonso J-M, et al. How much is too much?(Part 1)
557 International Olympic Committee consensus statement on load in sport and risk of injury. *Br*
558 *J Sports Med*. 2016;50(17):1030-41.

559 2.   Gabbett TJ. The training-injury prevention paradox: should athletes be training
560 smarter and harder? *Br J Sports Med*. 2016;50(5):273-80.

561 3.   Drew MK, Finch CF. The Relationship Between Training Load and Injury, Illness
562 and Soreness: A Systematic and Literature Review. *Sports Med*. 2016;46(6):861-83.

563 4.   Bowen L, Gross AS, Gimpel M, Li FX. Accumulated workloads and the
564 acute:chronic workload ratio relate to injury risk in elite youth football players. *Br J Sports*
565 *Med*. 2017;51(5):452-9.

566 5.   Colby MJ, Dawson B, Peeling P, et al. Multivariate modelling of subjective and
567 objective monitoring data improve the detection of non-contact injury risk in elite Australian
568 footballers. *J Sci Med Sport*. 2017;20(12):1068-74.

569 6.   Carey DL, Blanch P, Ong KL, Crossley KM, Crow J, Morris ME. Training loads and
570 injury risk in Australian football-differing acute: chronic workload ratios influence match
571 injury risk. *Br J Sports Med*. 2017;51(16):1215-20.

572 7.   Hulin BT, Gabbett TJ, Lawson DW, Caputi P, Sampson JA. The acute:chronic
573 workload ratio predicts injury: high chronic workload may decrease injury risk in elite rugby
574 league players. *Br J Sports Med*. 2016;50(4):231-6.

575 8.   Murray NB, Gabbett TJ, Townshend AD, Hulin BT, McLellan CP. Individual and
576 combined effects of acute and chronic running loads on injury risk in elite Australian
577 footballers. *Scand J Med Sci Sports*. 2017;27(9):990-8.

578 9.   Malone S, Owen A, Newton M, Mendes B, Collins KD, Gabbett TJ. The acute:chonic
579 workload ratio in relation to injury risk in professional soccer. *J Sci Med Sport*.
580 2017;20(6):561-5.

581 10.   Malone S, Roe M, Doran DA, Gabbett TJ, Collins KD. Protection Against Spikes in
582 Workload With Aerobic Fitness and Playing Experience: The Role of the Acute:Chronic
583 Workload Ratio on Injury Risk in Elite Gaelic Football. *Int J Sports Physiol Perform*.
584 2017;12(3):393-401.

585 11.   Ruddy JD, Shield AJ, Maniar N, et al. Predictive Modeling of Hamstring Strain
586 Injuries in Elite Australian Footballers. *Med Sci Sports Exerc*.
587 2017;10.1249/MSS.0000000000001527.

588   12.   Thornton HR, Delaney JA, Duthie GM, Dascombe BJ. Importance of Various
589   Training Load Measures on Injury Incidence of Professional Rugby League Athletes. *Int J*
590   *Sports Physiol Perform*. 2016:1-17.

591   13.   Malone S, Owen A, Mendes B, Hughes B, Collins K, Gabbett TJ. High-speed running
592   and sprinting as an injury risk factor in soccer: Can well-developed physical qualities reduce
593   the risk? *J Sci Med Sport*. 2018;21(3):257-62.

594   14.   Carey DL, Ong K-L, Whiteley R, Crossley KM, Crow J, Morris ME. Predictive
595   modelling of training loads and injury in Australian football. *International Journal of*
596   *Computer Science in Sport*. Forthcoming 2018;17.

597   15.   Bourdon PC, Cardinale M, Murray A, et al. Monitoring Athlete Training Loads:
598   Consensus Statement. *Int J Sports Physiol Perform*. 2017;12(Suppl 2):S2161-S70.

599   16.   Greenland S. Dose-response and trend analysis in epidemiology: alternatives to
600   categorical analysis. *Epidemiology*. 1995;6(4):356-65.

601   17.   Bennette C, Vickers A. Against quantiles: categorization of continuous variables in
602   epidemiologic research, and its discontents. *BMC Med Res Methodol*. 2012;12:21.

603   18.   Altman DG, Royston P. The cost of dichotomising continuous variables. *BMJ*.
604   2006;332(7549):1080.

605   19.   Harrell FE. Regression modeling strategies: with applications to linear models,
606   logistic regression, and survival analysis. New York: Springer; 2001. xxii, 568 p.

607   20.   Kahan BC, Rushton H, Morris TP, Daniel RM. A comparison of methods to adjust for
608   continuous covariates in the analysis of randomised trials. *BMC Med Res Methodol*.
609   2016;16:42.

610   21.   Wainer H, Gessaroli M, Verdi M. Visual Revelations. *Chance*. 2013;19(1):49-52.

611   22.   Gabbett TJ. The development and application of an injury prediction model for
612   noncontact, soft-tissue injuries in elite collision sport athletes. *J Strength Cond Res*.
613   2010;24(10):2593-603.

614   23.   Rossi A, Pappalardo L, Cintia P, Fernandez J, Iaia FM, Medina D. Who is going to
615   get hurt? Predicting injuries in professional soccer. In: *Proc the Machine Learning and Data*
616   *Mining for Sports Analytics workshop (MLSA'17), ECML/PKDD*; 2017 Sep 18; Skopje
617   (Macedonia). CEUR Workshop Proceedings; 2017. p. 21-30.

618   24.   Kuhn M, Johnson K. Applied Predictive Modeling. New York, NY: Springer New
619   York; 2013. 600 p.

620   25.   Cook NR. Use and misuse of the receiver operating characteristic curve in risk
621   prediction. *Circulation*. 2007;115(7):928-35.

622   26.   Mosler AB, Weir A, Eirale C, et al. Epidemiology of time loss groin injuries in a

623   men's professional football league: a 2-year prospective study of 17 clubs and 606 players. *Br*

624   *J Sports Med*. 2018;52(5):292-7.

625   27.   Blanch P, Gabbett TJ. Has the athlete trained enough to return to play safely? The

626   acute: chronic workload ratio permits clinicians to quantify a player's risk of subsequent

627   injury. *Br J Sports Med*. 2016;50(8):471-5.

628   28.   Williams S, West S, Cross MJ, Stokes KA. Better way to determine the acute:chronic

629   workload ratio? *Br J Sports Med*. 2016;10.1136/bjsports-2016-096589:bjsports-2016-096589.

630   29.   Lolli L, Batterham AM, Hawkins R, et al. Mathematical coupling causes spurious

631   correlation within the conventional acute-to-chronic workload ratio calculations. *Br J Sports*

632   *Med*. 2017;10.1136/bjsports-2017-098110.

633   30.   R Core Team. R: A Language and Environment for Statistical Computing. Vienna,

634   Austria2014.

635   31.   Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model

636   continuous risk variables in epidemiology. *Int J Epidemiol*. 1999;28(5):964-74.

637   32.   Ambler G, Benner A. mfp: Multivariable Fractional Polynomials. R package version

638   1.4. 9. 2010.

639   33.   Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration

640   of logistic regression models by using loess smoothers. *Stat Med*. 2014;33(3):517-35.

641   34.   Benedetti R. Scoring Rules for Forecast Verification. *Mon Weather Rev*.

642   2010;138(1):203-11.

643   35.   Touloumis A. Simulating Correlated Binary and Multinomial Responses under

644   Marginal Model Specification: The SimCorMultRes Package. *The R Journal*. 2016;8(2):79-

645   91.

646   36.   Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models.

647   *Biometrika*. 1986;73(1):13-22.

648   37.   Halekoh U, Højsgaard S, Yan J. The R package geepack for generalized estimating

649   equations. *Journal of Statistical Software*. 2006;15(2):1-11.

650   38.   Simpson D, Whiteley R, Carey D. Workload and Gaelic Football injury risk–have we

651   been fooled by spurious correlations or is Gaelic Football really different gravy? *Phys Ther*

652   *Sport*. 2017;28:e25.

653   39.   Holmberg L, Vickers A. Evaluation of prediction models for decision-making:

654   beyond calibration and discrimination. *PLoS Med*. 2013;10(7):e1001491.

655   40.   Nielsen RO, Bertelsen ML, Verhagen E, et al. When is a study result important for
656   athletes, clinicians and team coaches/staff? *Br J Sports Med*. 2017;51(20):1454-5.
657   41.   Twisk JWR. Applied longitudinal data analysis for epidemiology : a practical guide.
658   Cambridge, UK ; New York: Cambridge University Press; 2003. xvi, 301 p. p.
659   42.   Moller M, Nielsen RO, Attermann J, et al. Handball load and shoulder injury rate: a
660   31-week cohort study of 679 elite youth handball players. *Br J Sports Med*. 2017;51(4):231-
661   7.
662

663   **Supplemental digital content**
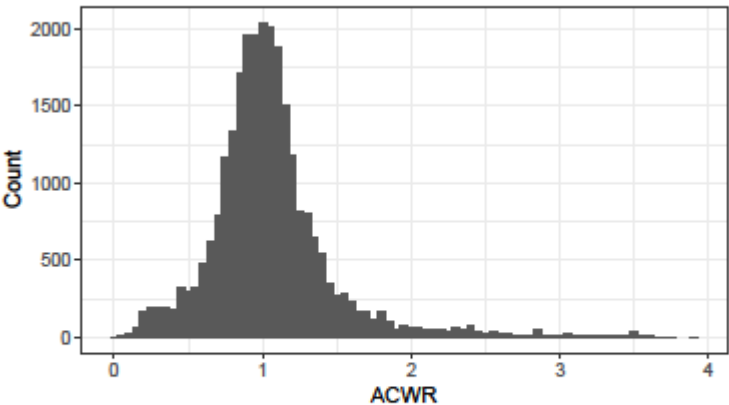
SUPPLEMENTAL DIGITAL CONTENT 1 - Histogram showing distribution of
ACWR values (pooled from two existing studies) used to simulate training load
datasets.

664

665   **[SDC1 - Supplemental Figure 1.pdf]** Supplemental figure 1: Histogram showing

666   distribution of ACWR values (pooled from two existing studies) used to simulate training

667   load datasets.

668

669 **SUPPLEMENTAL DIGITAL CONTENT 2:** Equations of theoretical risk profiles.

670

671 *U-shaped:*

672
$$\Pr(Y_i = 1 | R_i = r_i) = logistic(\begin{cases} -3.4 + 2 \cdot (1 - r_i)^2, & r_i < 1 \\ -3.4 + (1 - r_i)^2, & 1 \le r_i < 1.7 \\ 1.5 \cdot r_i - 5.4, & r_i \ge 1.7 \end{cases})$$

673 Where r is the ACWR, Y is an indicator variable for injury occurrence and *logistic* is the

674 standard function:

675
$$logistic(x) = \frac{1}{1 + e^{-x}}$$

676 *Flat:*

677
$$\Pr(Y_i = 1 | R_i = r_i) = 0.05$$

678 *S-shaped:*

679
$$\Pr(Y_i = 1 | R_i = r_i) = \begin{cases} 0.02, & r_i < 1.3 \\ logistic(3.5 \cdot (r_i - 1.3) - 3.89), & r_i \ge 1.3 \end{cases}$$

680

681

682 **[SDC2 - Supplemental Text.docx]** Supplementary text: Equations of theoretical risk

683 profiles.

684

685 **SUPPLEMENTAL DIGITAL CONTENT 4**

686 Supplementary Table 1: Results of 100 trials for each combination of sample size and risk

687 profile.

Sample size	Risk Profile	Mean number of simulated injuries (range)
	U-shaped	43.47 (29 – 66)
1000	Flat	50.43 (33 – 72)
	Threshold	42.00 (26 – 60)
	U-shaped	217.73 (172 – 247)
5000	Flat	251.37 (214 – 283)
	Threshold	211.50 (177 – 251)

688

689 **[SDC4 - Supplemental Table 1.docx]** Supplementary table 1: Results of 100 trials for each

690 combination of sample size and risk profile.

691
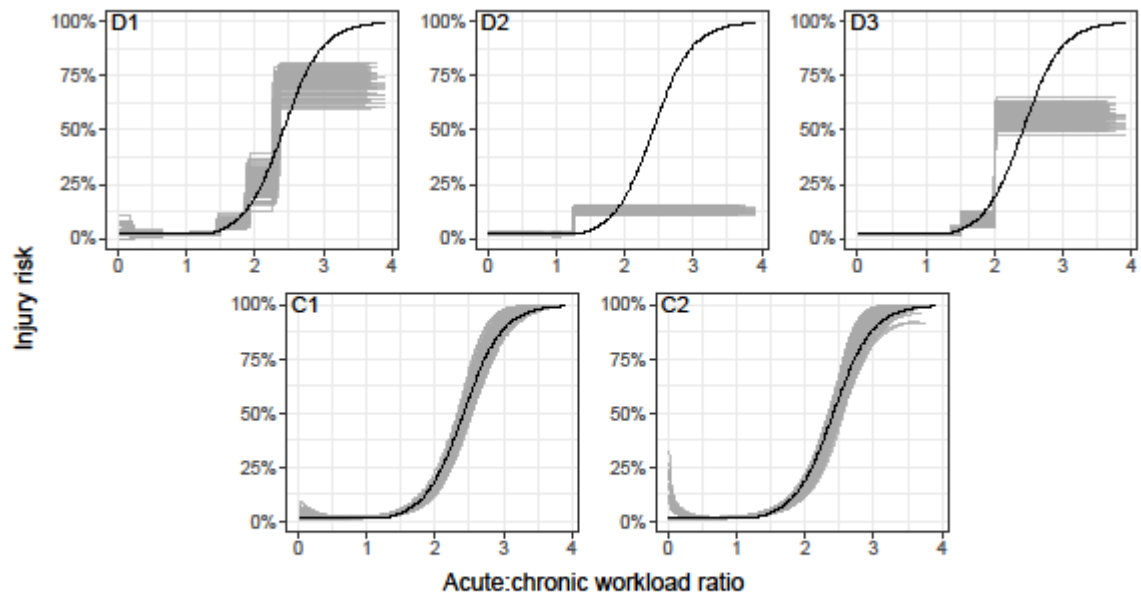
SUPPLEMENTAL DIGITAL CONTENT 5 - Comparison of 100 simulated study results (Ns= 5000 and a S-shaped risk profile) analysed using discrete models (D1, D2, D3) and continuous models (C1, C2). Solid line represents the true risk profile used to generate the data.

692

693 **[SDC5 - Supplemental Figure 2.pdf]** Supplementary figure 2: Comparison of 100 simulated

694 study results (N_s = 5000 and S-shaped risk profile) analysed using discrete models (D1, D2,

695 D3) and continuous models (C1, C2). Solid line represents the true risk profile used to

696 generate the data.

697
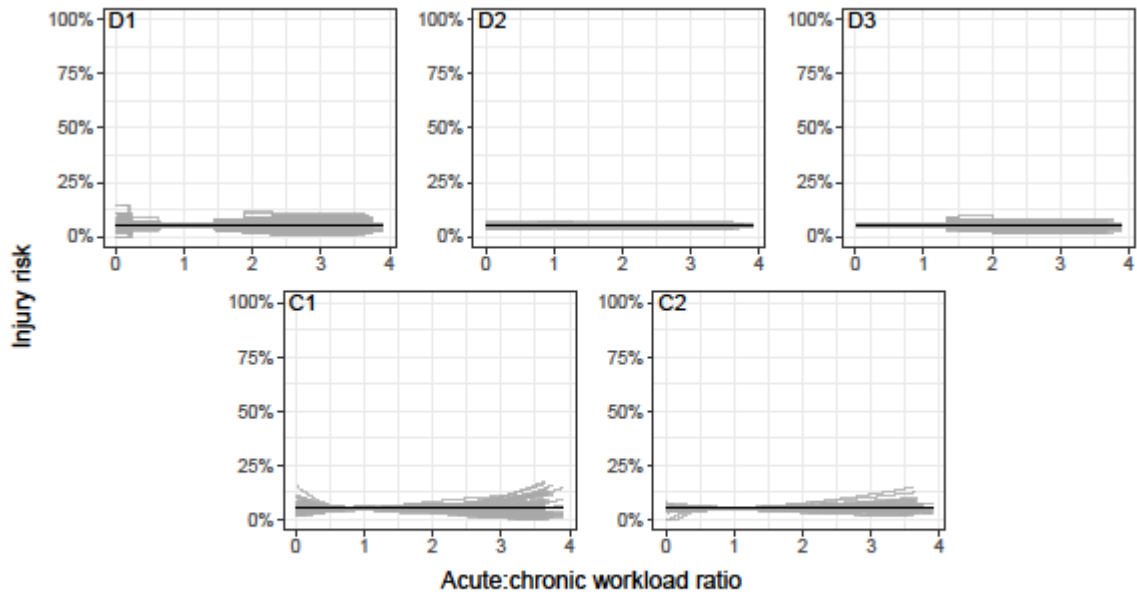
SUPPLEMENTAL DIGITAL CONTENT 6 - Comparison of 100 simulated study results (Ns= 5000 and a flat risk profile) analysed using discrete models (D1, D2, D3) and continuous models (C1, C2). Solid line represents the true risk profile used to generate the data.

698

699   **[SDC6 - Supplemental Figure 3.pdf]** Supplementary figure 3: Comparison of 100 simulated

700   study results (N_s = 5000 and a flat risk profile) analysed using discrete models (D1, D2, D3)

701   and continuous models (C1, C2). Solid line represents the true risk profile used to generate

702   the data.

703

704 **SUPPLEMENTAL DIGITAL CONTENT 7**

705 Supplementary Table 2: False rejection rates for discrete and continuous modelling

706 approaches.

Sample characteristics	False rejection rate (/100)				
	D1	**D2**	**D3**	**C1**	**C2**
U-shaped (N_s=1000)	5	59	57	12	19
U-shaped (N_s=5000)	0	4	0	0	0
S-shaped (N_s=1000)	0	2	16	0	0
S-shaped (N_s=5000)	0	0	0	0	0

707

708 **[SDC7 - Supplemental Table 2.docx]** Supplementary table 2: False rejection rates for

709 discrete and continuous modelling approaches.

710