Prediction of Running Injuries from Training Load: a Machine Learning Approach.

Talko Dijkhuis[1,3], Ruby Otter[2,4], Hugo Velthuijsen[1] and Koen Lemmink[4]
[1]Hanze University of Applied Sciences,Institute for Communication, Media & IT, Groningen, The Netherlands
[2]Hanze University of Applied Sciences,Institute of Sport Studies, Groningen, The Netherlands
[3]University of Groningen, Johann Bernoulli Institute for Mathematics and Computer Science, Groningen, The Netherlands
[4]University of Groningen, Center for Human Movement Sciences, University Medical Center Groningen, Groningen, The Netherlands
[1]{t.b.dijkhuis, t.a.otter, h.velthuijsen}@pl.hanze.nl
[4]{k.a.p.m.lemmink}@rug.nl

***Abstract*—The prediction of the running injuries based on self-reported training data on load is difficult. At present, coaches and researchers have no validated system to predict if a runner has an increased risk of injuries. We aim to develop an algorithm to predict the increase of the risk of a runner to sustain an injury. As a first step Self-reported data on training parameters and injuries from high-level runners (duration=37 weeks, n=23, male=16, female=7) were used to identify the most predictive variables for injuries, and train a machine learning tree algorithm to predict an injury. The model was validated by splitting the data in training and a test set. The 10 most important variables were identified from 85 possible variables using the Random Forest algorithm. To predict at an earliest stage, so the runner or the coach is able to intervene, the variables were classified by time to build tree algorithms up to 7 weeks before the occurrence of an injury. By building machine learning algorithms using existing self-reported training data can enable prospective identification of high-level runners who are likely to develop an injury. Only the established prediction model needs to be verified as correct.**

Keywords–Human Performance; Machine Learning; Predictive Analysis; Load; Injuries; Monitoring; Endurance Athletes

I. INTRODUCTION

In the context of the project 'Groningen Monitoring Athletic Performance Study'(MAPS) [1] two years of data was gathered about load, and tests on the performance of competitive athletes. The investigation into the factors that influence performance and injury risk of athletes lead to more insight. The effect of a change in load on injuries is difficult to predict. The knowledge on the relationship between load and the effect on injuries might be improved by matching self-reported training data and injury data using machine learning techniques. At present, Human Movement Researchers of the project MAPS have no published findings on the relation between a change in load and injuries on endurance athletes. To identify the predictors, a choice has to be made on which information is to be used for developing an algorithm for the prediction of injuries. There are two kinds of data available, the daily self-reported training log data and the injury data of high-level competitive runners. An injury is defined as any musculosketal problem of the lower extremity or back that lead to an inability to execute training or competition as planned for at least one week [2]. The training log data contains information about the training duration and the training intensity. The intensity times the duration is the workload of a training. The terms acute and chronic workload are used to descibe the intensity of the immediate window of training. Acute workload is the average workload of the last seven days, chronic workload is the average workload of the last 28 days. [3]. A study to sustain an injury risk of rugby league players was conducted with the acute:chronic ratio and concluded that there was a relation between High Chronical workload, spikes in the acute workload and the increased risk on injuries [4]. To find a relation between the workload and injuries for high-level runners, the data of the training data log is to be converted to predictors, which are based on the acute:chronic ratio, derived ratio's and the workload. Machine Learning is an appropriate manner for examination of all these predictors [5] because Machine Learning Techniques can discover complex high dimensional interactions between predictors and predict the label of injury/no injury.

II. METHODS

A. Study design

The study design stems out of the MAPS project. A prospective cohort study was used in which 23 high-level competitive runners (16 male, 7 female) were followed for 14 months. During this period they reported data on training parameters and injuries(Table I).

TABLE I. BASELINE CHARACTERISTICS OF RUNNERS

Characteristic	*Male*	*Female*	*Total*
Number	16	7	23
Age(years;mean +- SD)	22.5 +- 6.3	21.4 +- 4.4	22.2 +- 5.7
Height(cm;mean +- SD)	185 +- 5	172 +- 7	181 +- 8
Body weight(kg;mean +- SD)	68.6 +- 6.0	58.3 +- 4.0	65.4 +- 7.2
Perc. body fat(perc.;mean +- SD)	8.5 +- 2.3	17.6 +- 4.2	11.3+- 5.2
VO2max(ml/min/kg)	66.7 +- 5.9	62.7 +- 7.4	65.5+-6.5

B. Dataset description

1) Overall dataset: The runners kept a daily training log, in which, information about the training duration, the training intensity and the sustained injuries was reported. The dataset contains 208 training patterns of 7 weeks, 52 patterns with injuries and 156 patterns without any injury. To find a relation between the workload and injuries for high-level runners, the data of the training data log is converted to features, which are based on the acute:chronic ratio, monotony, strain and the workload. For a sliding window of 7 weeks before the injury, every single week the same ratio's and workload were calculated. The final step was to determine the percentage of change between the features of the identified weeks and add this to the dataset

2) Injuries: There were 22 runners, out of 23, injured during the observed period of 14 months. Only the complete patterns were selected for the dataset. These are the athletes who had no missing data in their daily training log in combination with the reported injuries. And only the injuries, which had a sliding window of 7 weeks of self-report data before the injury were used.

3) No injuries: The period before the injury has to be compared with the period in which the workload doesn't lead to injuries. Therefore the 7 week period without resulting into an injury, before the 7 weeks sliding window of the injury date, was taken for every individual athlete.

C. Model development

For the identification of the most important features Random Forest of Sklearn was used. Random Forest builds trees on subsets of the data, bagging, and there for applicable for relative small datasets. [6] Next, the relative importance of the features can be identified. To intervene as early as possible, single trees will be assembled based on the moment in time using the identified features.

D. Feature selection

The features within and between the sliding windows are correlated with each other. The following features are identified as the most important, sorted by decreasing importance.

1) average workload week 2
2) sum workload week 2
3) percentage change monotony between week 1 and 2
4) acute:chronic ratio 7 over 42 week 7
5) acute:chronic ratio 7 over 28 week 7
6) percentage change strain between week 1 and 2
7) percentage change workload between week 2 and 3
8) acute:chronic ratio 7 over 42 week 2
9) percentage change strain between week 2 and 3
10) strain 2

The Random Forest is also used to predict on the dataset and had predicted 67 percent of the injuries correct.

E. Predictive modelling building

The 10 features are used to build individual trees to be able to identify as early as possible the increase of risk. The dataset was divided random in a training and a test set in respectively 75% and 25%. Building a tree on the features 'acute:chronic ratio 7 over 42 week 7' and 'acute:chronic ratio 7 over 28' resulted in an accuracy of 75 %. With these ratios the prediction of the injury was 75% correct. Other trees are not built yet.

III. Results

We selected 10 predictors to predict the occurrence of an injury in the future over an sliding window of 7 weeks. The predictive modelling is promising but it is also a bit suspicious that an tree is more accurate using 20% of the features in comparison with the Random Forest.

IV. Conclusion

The predictive modelling using two steps in the process seems to be promising. But the accuracy of the tree on 20% of the feature set is very high, future work is to investigate the rationale behind the results. The established prediction model needs to be verified as being correct. When a robust and accurate prediction model is realized, the model will be added to an app. The app will get information about the workload per training session and, using the model, predicts the probability on sustaining an injury.

Acknowledgment

The authors would like to thank Henk van der Worp for identifying the injuries in the data of the runners and M. Aiello for suggesting on improving the paper.

References

[1] S. Raak-Pro, "Belasting en belastbaarheid van topsporters," 2011. [Online]. Available: http://www.sia-projecten.nl/projectenbank/project/belasting-en-belastbaarheid-van-topsporters

[2] S. W. Bredeweg, S. Zijlstra, and I. Buist, "The GRONORUN 2 study: effectiveness of a preconditioning program on preventing running related injuries in novice runners. The design of a randomized controlled trial," BMC Musculoskelet Disord, vol. 11, 2010, p. 196. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2936887/pdf/1471-2474-11-196.pdf

[3] T. J. Gabbett, B. T. Hulin, P. Blanch, and R. Whiteley, "High training workloads alone do not cause sports injuries: how you get there is the real issue," British Journal of Sports Medicine, vol. 50, no. 8, 2016, pp. 1–2. [Online]. Available: http://bjsm.bmj.com/lookup/doi/10.1136/bjsports-2015-095567

[4] B. T. Hulin, T. J. Gabbett, D. W. Lawson, P. Caputi, and J. a. Sampson, "The acute:chronic workload ratio predicts injury: high chronic workload may decrease injury risk in elite rugby league players," British Journal of Sports Medicine, vol. 50, no. 4, 2016, pp. 231–236. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26511006 5Cnhttp://bjsm.bmj.com/lookup/doi/10.1136/bjsports-2015-094817

[5] M. Kuhn and K. Johnson, Applied predictive modeling. Springer, 2013, vol. 26.

[6] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, 2001, pp. 5–32.