# Subproject 1

Final version created 2020-04-30 before study analysis.

## Training load measures

U19 and handball data only have sRPE. Analyzing intensity, time and intensity\*time (all modified by splines or fractional polynomials) in the same model may be unfeasible due to number of injuries, especially while adjusting for age, sex and within-individual random effects. No matter the load measure, all will have a version with the following modifications:

* load unmodified (not aggregated over time periods, risk of injury is on same day as load. “baseline” method.)
* load ACWR 7:21 (uncoupled). Mean load across 7 day acute period / EWMA of 21 days chronic period.
* load ACWR 1:3 match periods (uncoupled). Mean load of 1 match period / EWMA load of 3 next match periods.
* mean(micro-cycle) – mean(previous micro-cycle) / mean(previous micro-cycle)

## Injury definition

Consider all health problems classified as “injury” as an event of interest, regardless of degree, time-loss and location, but consider “disease” as no injury.

Assume independence between different injuries. We will not adjust for it so far - this is something will bring back in subproject 3.

For analysis with match periods, evaluate the association between acute load in a match period and injury in the *next* match period.

For training load values calculated using day as time period, we will look at risk of injury 4 days later in time. We can choose to adjust this to 3 days etc.

## Non-simulated Analysis

Aims (to keep in mind!):   
1) Is the relationship between load and injury non-linear, as suggested by Gamble 2013 and Gabbett 2016?  
2) What is the shape of the relationship between load and injury in different sport populations?  
3) If the relationship is non-linear, what would we erroneously conclude by assuming linearity?   
  
Step 1) Multiple imputation.   
1 set of multiple datasets with imputed load values from multiple imputation (to improve statistical power in complex analyses). Variables in the imputation model in accordance with chapter 6.3 in <https://stefvanbuuren.name/fimd/sec-knowledge.html> , that is, all variables since we have less than 15. Derived load measures (such as ACWR) are not included in the imputation model but calculated after imputation in each dataset (in accordance with chapter 6.4). Any models in step 2 are run on each imputed dataset and then pooled (as described in chapter 5.1). By default, the method uses pmm, predictive mean matching on numeric data (load). “Many missing value imputation methods work by estimating joint distributions on all variables, so they fill in missing values of all columns recursively to better estimate a joint distribution of all values. The imputation is better if you do it on all columns. If you want to drop the imputed values, just save a dataframe of missing values, and you can reset imputed values to missing in columns you want to keep missing”

Step 2) Adjusting for age and sex in all models, perform the following:

* GLMM with binomial distribution with load unmodified (linearity assumed) with random intercept and random slope
* GLMM with binomial distribution with load modified by splines, with random intercept (k number of knots decided by AIC)
* GLMM with binomial distribution with load modified by splines, with random intercept and random slope
* If model doesn’t converge with random effects, and age and sex appear to be inconsequential, remove adjustment
* Compare the estimates of the above models – did effect sizes change dramatically by inclusion of random effects? Compare AIC/BIC of the different models. Which was consistently best fit? Was the linearity assumption violated in the first model?
* From best-fitting model, create a figure showing the predicted OR-values vs. each level of load  
    
  Repeat step 2) for each load measure, for each population. Figures are presented in results section. Final models reported in a table.

## Simulated Analysis

The dataset deemed to have best data quality is used for all simulations.  
Aims:   
1) If the relationship between load and injury is non-linear, which method(s) discover the true relationship a reliable amount of the time (> 95%)? Which method(s) ascertain the true relationships most accurately?

2) If there is no relationship between load and injury (Null hypothesis is true), does any method have an increased (> 5 %) risk of detecting spurious relationships?

3) If the relationship is non-linear, how often will methods that assume linearity provide statistically significant results?

4) Are all methods equally robust if the relationship is in fact linear? (do we have anything to lose from using non-linear methods)?

5) Which method is most understandable – both to understand how the method works, and how interpretable the results are.

Step 1) Create 4 datasets with 4 levels of n load values: 1) original load values, 2) n = 100, 3) n = 1000, 4) n = 10 000. What are the sample sizes in the original data?

Step 2) For each dataset created in dataset 4, create datasets with fake injuries by pre-determined probability distributions:   
- Injuries from a linear relationship with load  
- Injuries from a quadratic relationship with load  
- No association (flat), injuries are added from a uniform distribution  
- One or more probability functions discovered in the non-simulated analysis  
- For ACWR-load only, injuries from Gabbett’s J-shape-function  
Repeat for each load measure.

Step 3) Perform the following X times (1 000?) on each dataset created in steps 1-2 (100)

* Logistic regression (linearity assumed)
* Logistic regression with load categorized (by quartiles)
* Logistic regression with load as a quadratic term (quadratic regression)
* Logistic regression with load modified with splines
* Logistic regression with load modified with fractional polynomials

For each method, plot in comparison to the created relationship (similar to <https://cdn-links.lww.com/permalink/mss/b/mss_2018_06_19_carey_18-00196_sdc5.pdf>), calculate n and percentage that load is statistically significant (false discovery and false rejection rate), report mean model parameters (OR, confidence interval, p-values) and root-mean-square deviation (RMSE) injury probability.   
To assess validity of simulation, report Monte Carlo standard errors.

## Appendix-level analysis

Results that might be interesting, applicable or necessary to offer in an appendix.

* Data quality tables with an overview of the amount of missing data and the timeliness (mean number of days from a player received a survey form until they answered) for each population.
* Maybe appendix—but maybe results: Figures showing distribution of load during match periods, maybe in comparison to weekly loads, or to daily loads. Maybe for a handful of players as examples. Maybe some in results and some in appendix. A lot of options.
* Plots and maybe descriptive statistics showing difference between imputed and non-imputed load values. Are the distributions similar? Is skewness, outliers etc. retained in the imputed dataset? Perhaps imputation-evalutation statistics (chapter 2.5 in Van Buuren).
* Plot of 1) acute vs. chronic values and 2) ACWR vs. chronic values to see if the following assumption is true: “If there *is* a relationship between the numerator and the denominator, a ratio will be effective only when the relationship between the numerator and the denominator is a straight line that intersects the origin”, <https://journals.physiology.org/doi/full/10.1152/advan.00053.2013> .
* Table with all model results from non-simulated analysis (not just final model), including effect sizes, confidence intervals, p-values and AIC/BIC.

# Deviations from the protocol since 2020-04-30

**1.** **“Simulated analysis, Step 1)”** Comparing 4 different sample sizes. We ended up comparing two, and neither of those listed.

The strength of the simulated relationship determines the number of injuries that will appear in the data. What defines a small and large sample size is largely determined by the number of injury events (Riley et al., 2019), not the number of training load values. For a small sample, the strength of the relationship would therefore have to be weaker, so that fewer injuries are drawn. However, the relationships have to be strong enough to show clear non-linearity. We therefore stepped away from the idea of comparing performance of methods in different sample sizes.

However, we still included two examples of training load studies, one with the original training load observations, and one resampled to 22 500 observations, as representations of two types of studies.

**2.** **“Simulated analysis, Step 2)”** Comparing 5 different relationship shapes. We did not include the “flat”/no relationship variant. The idea was to see if false discovery rates in (Carey et al., 2018) could be reproduced in our population, but it was dropped for two reasons.

The first, was that the methods aren’t comparable when it comes to Type I or II error rates, because the hypothesis tests are on different grounds. The null hypothesis “Athletes who are within x and y zones of training load do not have an increased risk of injury compared to z and w training load zones” is very different from “The relationship between training load and risk of injury is linear.”

Secondly, the different methods output an unequal number of p-values. Should you count categorization as significant if at least 1 of 3 tests were significant, or do all have to be significant to count? What about both terms for the fractional polynomials? I’m not sure what would be correct here.

**3.** **“Simulated analysis, Step 2)”** “One or more probability functions discovered in the non-simulated analysis” was not included either. The discovered relationships were too weak with our sample sizes of 8 494 and 22 500 training load observations, respectively.

**4. “Simulated analysis, Step 3)”** Instead of choosing 1000 simulations, a bit arbitrary, we used equations in (Morris et al., 2019) the number of simulations.

**5.** **“Simulated analysis, Step 3)”** To determine difference in performance between subjective (data-informed) and data-driven approaches, categorization and restricted cubic splines was compared by specifying them in two different ways.

**6. “Appendix-level analyses”.** “Maybe appendix—but maybe results: Figures showing distribution of load during match periods, maybe in comparison to weekly loads, or to daily loads. Maybe for a handful of players as examples.”

We chose not to include this, as we already had a lot of results to convey in a single paper. It also did not mean the aim of the current study.

Carey, D. L., Crossley, K. M., Whiteley, R., Mosler, A., Ong, K.-L., Crow, J., & Morris, M. E. (2018). Modeling Training Loads and Injuries: The Dangers of Discretization. *Medicine and science in sports and exercise*, *50*(11), 2267-2276.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine*, *38*(11), 2074-2102.

Riley, R. D., Snell, K. I., Ensor, J., Burke, D. L., Harrell Jr, F. E., Moons, K. G., & Collins, G. S. (2019). Minimum sample size for developing a multivariable prediction model: PART II‐binary and time‐to‐event outcomes. *Statistics in medicine*, *38*(7), 1276-1296.