# Missing Data Project (Final version before analysis 2021-02-08)

## Aim

Determine how to handle missing training load data in load(-injury?) research. Suggested research questions:

1) Is deletion of rows of missing data (case-wise/listwise deletion) a suitable method, or must imputation of some kind be used?

2) If imputation is necessary, which imputation method(s) are able to replicate real life values most accurately, and retain the properties of the original data?  
  
3) Is there any difference between imputing the sRPE and imputing GPS measures?

4) Would imputation affect study conclusions? More concretely, does the imputation affect the results of a logistic regression model (coefficients), if it does, how and to what degree?

## Study Design

Experimental simulation study. Will be based on a real cohort of longitudinal sports data.

## Data

Using the Norwegian Premier Division data will provide both sRPE and GPS measures. It’s a small cohort, but followed for a whole year, meaning there are many load values. Most studies that reported having missing data were studies using GPS measures.

# Simulation process

## Questions 1-3)

### Step 1) Data preparation

Obtain sRPE as one dataset and the Total Distance in another. Total distance is chosen as it is the GPS measure with most consistent definition of all the GPS measures (Maupin et al., 2020, Dwyer and Gabbett, 2012).

Delete all rows in the sRPE dataset and GPS dataset that have missing data. The dataset with omitted missing will be the baseline for the simulation (0% missing data), and will be our baseline for true values that our missing methods will attempt to estimate.

Create 6 datasets with the following amount of missing:

1. 5%
2. 10 %
3. 20%
4. 40%
5. 60%
6. 80%

Of articles that have reported the amount of missing, the mean amount was 6.2%. That is why both 5% and 10% are included among the tested missing amounts.

A fake longitudinal correlation between the remaining load values will be created to represent dependency of values on the same person.

Missing at Random is more common in medical research (Janssen et al., 2010), but it’s reasonable to believe that GPS data missing due to erroneous measures, and intermittent missing values (i.e., not missing for the whole team a whole day, but perhaps for 1 player that day), are Missing Completely at Random.

It would be fruitful to compare methods under these two different scenarios to see if some perform better under certain conditions, or if simpler methods can be used with safety for GPS data.

Under one scenario, Missing Completely at Random, missing will be introduced through random sampling. Under another scenario, Missing at Random, missing-probability will be weighted according to a fake variables. Ideally, fake age (continuous), sex (logical) and training date (date) to represent different types of predictors.

With the different missing scenarios, this will mean altogether 12 datasets with unequal level of missing and correlation structure.

Since we are testing both sRPE and GPS, a total of 24 datasets.

### Step 2) Imputation

Different methods for handling the missing data will be compared. The figure below shows the methods reported in 29 load-injury articles that have reported their methodology for handling missing data. This was used to determine methods for comparison.

1. Mean imputation is the most common method of imputation in load-injury research
2. Listwise deletion is a convenient method, and the default in most statistical programs
3. Median imputation may be a better method than the mean in skewed data
4. Regression imputation may perform better under MAR
5. Multiple imputation has been shown to perform well in other fields. Here, we will compare using predicted mean matching, as both sRPE and GPS data are continuous variables.

The five methods will be performed on the 24 datasets.



### Step 3) Calculate performance measures

The following performance measures are recommended in <https://stefvanbuuren.name/fimd/sec-evaluation.html#sec:evaluationcriteria> :

1. The Raw Bias (RB): the absolute difference between the imputed and the real value . RB should be close to zero. Will also be calculated as percent bias (PB), where the upper limit for acceptable performance is 5% (Demirtas et al., 2008).
2. Coverage: the proportion of confidence intervals that contain the true value. The actual rate should be equal to or exceed the nominal rate.
3. Average width (AW). The average width of the confidence interval is an indicator of statistical efficiency.
4. Root-Mean-Squared Error (RMSE). A compromise between bias and variance, and evaluates  on both accuracy and precision.

It may seem unnecessary with the average width when we also have coverage, but:  
“If all is well, then RB should be close to zero, and the coverage should be near 0.95. Methods having no bias and proper coverage are called randomization-valid (Rubin [1987b](https://stefvanbuuren.name/fimd/references.html#ref-RUBIN1987)). If two methods are both randomization-valid, the method with the shorter confidence intervals is more efficient.”

In addition, a visualization of the distribution of imputed vs. real data, and scatterplot of imputed vs. real data will show whether the imputation method retained the distribution and is capable of predicting outliers, and whether any impossible values (negative values) are imputed. This will have to be limited to, say, the GPS measure only, for 60% missing, to limit the number of figures for the article.

Lastly, if possible, create a so-called nested loop-plot that compares percent bias for all methods under all missing data scenarios (Rücker and Schwarzer, 2014).

1900 permutations has previously been calculated for the number required to estimate coverage with a Monte Carlo Standard Error set to 0.5.

The number of permutations needed for bias is calculated by:

Where is the sample variance (Morris et al., 2019). For an estimation of needed simulations, 100 permutations will be performed as a pilot simulation, and the number of total permutations needed will be calculated based on the variance of those 100 permutations and an MCSE set to 0.5.

## Question 4)

So far, the methods answer research question 1-3). For question 4) The hypothesis is that a proper imputation would only improve the certainty of model estimates (filling in load values will allow modeling more injuries, the limiting factor in load-injury research). The purpose of question is to reassure a potential user that an imputation method does not alter the model so that conclusions would change. If coefficients and other parameters do change so substantially that conclusions change, then perhaps imputation should not be recommended at all.

### Step 1) Data preparation

To test this, injuries will be added with a fake, linear relationship with sRPE in the 0% missing, baseline data. A logistic regression model with random effects will be run on the data, giving a baseline for an “ideal” model.

Then, missing will be added in the same manner as for the previous analysis, in step 1 data preparation. Altogether 12 datasets with 6 missing levels and 2 missing scenarios (MCAR and MAR).

### Step 2) Imputation

The best-performing imputation method + listwise deletion will be compared. This is mostly to limit the resources – coding, and simulation time. If simulations prove to be quick, more methods may be compared. Perhaps including the worst method may serve to illustrate the importance of choosing good methods.

### Step 3) Performance Measures

We can calculate bias and percentage bias by comparing the estimated Odds Ratio to the baseline, “true” Odds Ratio. This was also why we chose a linear model, non-linearity would have more terms and complicate this estimation. This will show if conclusions are altered.

Coverage of CI for the coefficient, and coverage for the predicted values, would show whether there is any improvement in certainty when imputing vs. deleting.

Similar to the non-linearity study, Brier Score and C-statistics could be included to show model fit and predictive ability.

## Appendix-level analysis

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

* Histogram showing distribution of sRPE and GPS data in the football premier division dataset.
* Dataset with articles used to deduct most commonly used missing data methods. In methods section, will explain how articles were found in Windt et al. (2018), Eckard et al. (2018) and Griffin et al. (2019) reviews, with the idea that these reviews reflect the most relevant studies in the field at the time etc.
* The figure used in this plan can be remade to just a table and can be referred to for methods choices.

BENSON, L. C., STILLING, C., OWOEYE, O. B. & EMERY, C. A. 2021. Evaluating Methods for Imputing Missing Data from Longitudinal Monitoring of Athlete Workload. *Journal of Sports Science and Medicine,* 20**,** 187-195.

DEMIRTAS, H., FREELS, S. A. & YUCEL, R. M. 2008. Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: a simulation assessment. *Journal of Statistical Computation and Simulation,* 78**,** 69-84.

DWYER, D. B. & GABBETT, T. J. 2012. Global positioning system data analysis: Velocity ranges and a new definition of sprinting for field sport athletes. *The Journal of Strength & Conditioning Research,* 26**,** 818-824.

ECKARD, T. G., PADUA, D. A., HEARN, D. W., PEXA, B. S. & FRANK, B. S. 2018. The relationship between training load and injury in athletes: a systematic review. *Sports medicine,* 48**,** 1929-1961.

GRIFFIN, A., KENNY, I. C., COMYNS, T. M. & LYONS, M. 2019. The Association Between the Acute: Chronic Workload Ratio and Injury and its Application in Team Sports: A Systematic Review. *Sports Medicine***,** 1-20.

JANSSEN, K. J., DONDERS, A. R. T., HARRELL JR, F. E., VERGOUWE, Y., CHEN, Q., GROBBEE, D. E. & MOONS, K. G. 2010. Missing covariate data in medical research: to impute is better than to ignore. *Journal of clinical epidemiology,* 63**,** 721-727.

MAUPIN, D., SCHRAM, B., CANETTI, E. & ORR, R. 2020. The Relationship Between Acute: Chronic Workload Ratios and Injury Risk in Sports: A Systematic Review. *Open Access Journal of Sports Medicine,* 11**,** 51.

MORRIS, T. P., WHITE, I. R. & CROWTHER, M. J. 2019. Using simulation studies to evaluate statistical methods. *Statistics in medicine,* 38**,** 2074-2102.

RÜCKER, G. & SCHWARZER, G. 2014. Presenting simulation results in a nested loop plot. *BMC medical research methodology,* 14**,** 129.

VAN BUUREN, S. 2018. *Flexible Imputation of Missing Data* [Online]. CRC Press. Available: <https://stefvanbuuren.name/fimd/> [Accessed 2021-02-12].

WINDT, J., ARDERN, C. L., GABBETT, T. J., KHAN, K. M., COOK, C. E., SPORER, B. C. & ZUMBO, B. D. 2018. Getting the most out of intensive longitudinal data: a methodological review of workload–injury studies. *BMJ open,* 8**,** e022626.

## Change log

1. **Imputation methods for comparison.** When analyses started, we realized that mean imputation can be calculated in different ways. If It performs poorly, it could be that we have chosen a poor variant. We therefore chose two different variations of mean imputation, to see the discrepancy in performance between the two. We used Benson et al. (2021) to guide our choices of means.   
     
   In adding another mean imputation, we removed median imputation. Few had used median imputation, and we suspected it would perform similarly to mean imputation.
2. **Product or factors?** During analysis, we realized that sRPE is a product of two factors, duration and RPE. In (Van Buuren, 2018) it wasn’t clear whether it was best to impute factors first, and then calculate sRPE, or impute after sRPE ahs been calculated. There wasn’t much evidence for either or. We therefore did a subanalysis to test this. The best procedure was used in the main simulation.
3. **More realistic GPS scenario.** GPS measures are collected simultaneously, and we thought it would be common for all GPS measures to be missing at the same time, if there is a signal error etc. We therefore assessed the case where all GPS variables were missing.
4. **Available variables.** Some studies collect both GPS and sRPE data. We tested whether imputation of GPS improved with access to sRPE. We also thought it unrealistic that the player’s playing position on the football field should not be available in a study. We tested whether imputations improved with the addition of this variable.
5. **Single vs. multiple imputation.** Only one multiple imputation method was compared: predicted mean matching. If it performs well, is it because we used a multiple imputation framework, or is predicted mean matching a good method on its own? We therefore did a sub-analysis of single vs. multiple imputation of predicted mean matching two determine how much of the performance stems from the multiple imputation framework, and how much from predicted mean matching itself.