# Sarcopenia Dataset

Writting Report: Statistical Analysis

Hanna Gloyna

The used data set contained initially 250 entries, where each entry represents a patient, with 84 features. After removing ID, 17 entries were deleted as they were duplicates. Furthermore 8 misses in *Marcha* (engl. walking speed) were imputed with the mean. The encoding of the categorical data was changed, so that afterwards 0 did always correspond to either *Not answered*, *Not required* or *Regular value*. To obtain comparable statistics a min-max-normalisation was applied, given by Equation 1.

$$\hat{x} = \frac{x - min(x)}{max(x) - min(x)} \tag{1}$$

Where $x$ corresponds to the vector of a feature, $min(x)$ the minimal values of this feature and $max(x)$ to the maximal value, respectively. Afterwards all features are within $[0, 1]$.

In accordance with specialists every representation of MMSE, Barthel, Norton, Lawton, and MNA, except the binary, was removed to reduce the dimensionality of the datset. Additionally MM was dropped, because of its high correlation to IMM.

We dropped further features because of too low variance or entropy. Additionally 7 features were deleted, because they contained a lot of misses or were included in another feature. With the remaining 53 features we performed a statistical analysis.

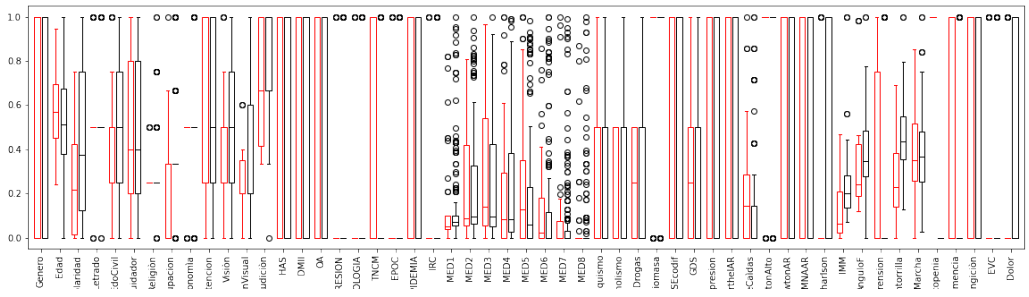


Figure 1: Boxplot of all remaining features, black boxes are from all mild cases and red from all serve, respectively