## Phenotype Preparation

1. **Transformation**: Please log transform smoking measured in pack years using the following transformation:

**log(pack\_years+1)**

* The constant of 1 is added to avoid undefined values

1. **Outliers:** Please trim outliers (points beyond 4 SDs of the mean) from the newly created variable. Specifically, please exclude observations that are greater than 4 SDs above the mean (i.e. > mean + 4\*SD) or less than 4 SDs below the mean (i.e. < mean – 4\*SD).

**outlierTrim(log(pack\_years+1))**

1. Please use the trimmed (i.e. outliers removed) and transformed variable in analyses

## Calculation of Scores

1. **Scores:** For reference, there are four scores to calculate:
2. A score based on Elastic Net regression (McCartney *et al.)*
3. A score based on CpGs identified in an EWAS by Zhang *et al.* (EpiSmokEr)
4. A score based on Bayesian penalised regression (models correlation structure between sites). Here, the CpG sites were identified by Trejo *et al.*
5. A score based on Elastic Net regression (mCigarette)
6. **Scripts:** There are two scripts to use in score generation, which are listed under scripts/ in the attached folder. The first script calculates scores a)-d) above. The second script automates tests for the correlation of all five scores with measured smoking (pack\_years).
7. 01\_generate\_scores.R
8. 02\_produce\_prediction\_metrics.R
9. **Example input files:** Example input files are included in the attached folder for reference under /examples.
10. **methylation\_df\_example.rds:** Fictitious methylation data are provided for 10 participants.

* These are methylation beta values. Ideally, you will have beta values in your analysis if possible but the scripts should convert M-values to beta values if not. However, it is worth double-checking that this step has worked.

1. **pheno\_example.csv:** Trimmed and transformed pack\_years values are also provided for the same 10 participants. The column names are ‘ID’ (needs to match participant IDs in the methylation file), ‘age’ (years), ‘sex’ (coded as 1-male, 2-female), ‘smoking\_cat’ (smoking category coded as 2 - current smokers, 1 - former smokers, 0 - never smokers) and ‘pack\_years’ (values have already been processed as per the phenotype preparations steps).

## Expected outputs

1. **Correlation:** Pearson’s correlation between trimmed and log-transformed pack years of smoking and each of the scores. Code to run and automate this step is included in script 02.
2. **AUC table:** AUC table comparing the predictive performance of scores a)-d). Again, code to run this is included in script 02.

Of note, all other files in the attached folder are either background files for analyses or example input files.