EpiSmokEr: An Epigenetic Smoking status Estimator

Institute for Molecular Medicine Finland
Nordic EMBL Partnership for Molecular Medicine

HiLIFE UNIT
UNIVERSITY OF HELSINKI

S BOLLEPALLI[1,2], T KORHONEN[1,3,4], J KAPRIO[1,2], M OLLIKAINEN[1,2] and S ANDERS[1,5]

[1]Institute for Molecular Medicine (FIMM), University of Helsinki, Finland; [2]Department of Public Health, University of Helsinki, Finland; [3]National Institute for Health and Welfare (THL), Finland; [4]Institute of Public Health and Clinical Nutrition, University of Eastern Finland, Finland; [5]Centre for Molecular Biology (ZMBH), University of Heidelberg, Germany.

Contact: sailalitha.bollepalli@helsinki.fi

Classification of individuals based on self-reported smoking status is prone to errors due to under-reporting and poor recall of long-term smoking history. Traditional biomarkers, such as cotinine, only measure short-term exposure. Cigarette smoking strongly influences DNA methylation, with **differential methylation patterns** detected among current, former and never smokers. We developed, **EpiSmokEr** a robust prediction tool to infer the smoking status based on the whole-blood DNA methylation profiles.

METHODOLOGY

- DNA methylation data from DILGOM (N=517 and Age range: 25-74) a sub-cohort of FINRISK was used as the training dataset.
- After stringent QC, quantile normalization was performed. **Quantiles** were later used to fit the distributions of validation datasets.
- Multinomial LASSO regression was performed to select a parsimonious set of CpG sites predictive of smoking status (Figure 1).
- Tuning parameter, **optimal lambda** was chosen through a tenfold internal cross-validation from a sequence of 100 lambda values.
- **121** methylation sites and the corresponding training coefficients selected from this regression were used to predict the smoking statuses in the three test datasets.
- Three independent and external test datasets were used to assess the performance.

RESULTS

- **EpiSmokEr showed** high accuracy for predicting current and never smokers in all the 3 test datasets.
- Figures 2 - 4 illustrate results from the Finnish Twin Cohort.
- Precise classification of self-reported former smokers from current and never smokers depends on the extent and duration of smoking prior to smoking cessation. **Longer cessation time leads to reversal of methylation profiles in former smokers making them indistinguishable from never smokers** (Figures 3 & 4).

ADVANTAGES of EpiSmokEr

- Considers three smoking statuses
- Implementation of quantile normalization on test datasets using quantiles from discovery dataset has allowed transferring prediction model from one study to another. Thereby allowing and improving cross-study performance.
- Serves as a robust predictor of biological smoking status based on methylation data.
- Globally applicable to all populations.
- **Available on GitHub:** https://github.com/sailalithabollepalli/EpiSmokEr
- Output provided as Smoking status labels, in HTML and CSV file formats.

1

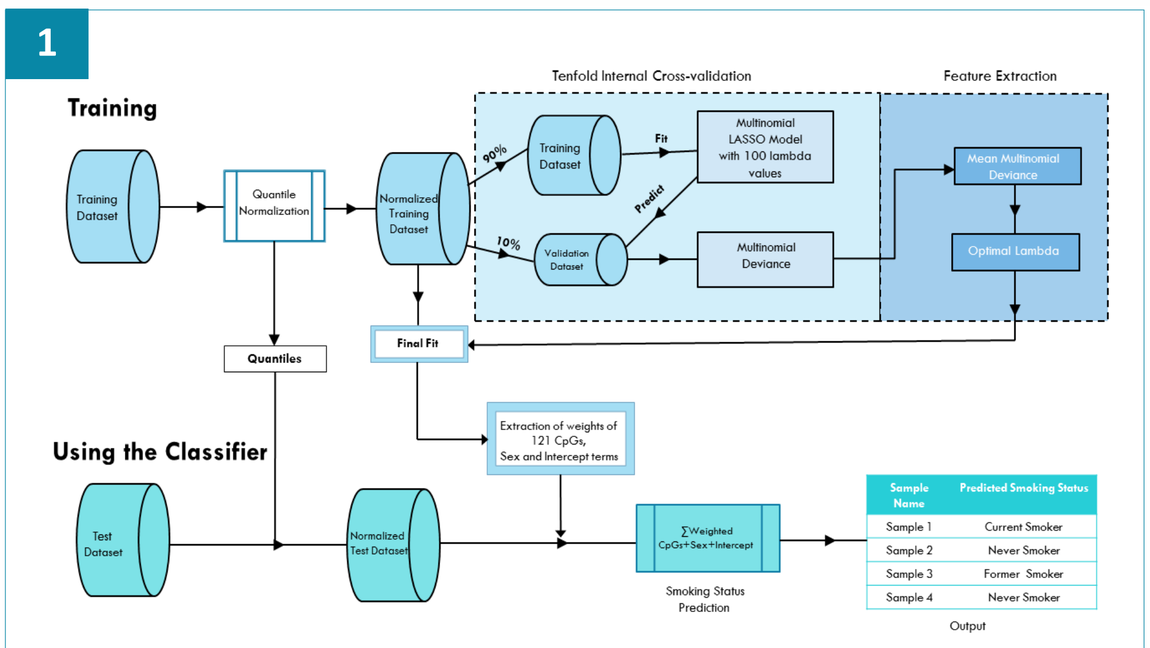

Figure 1: Schematic Representation of the Workflow of the EpiSmokEr

2

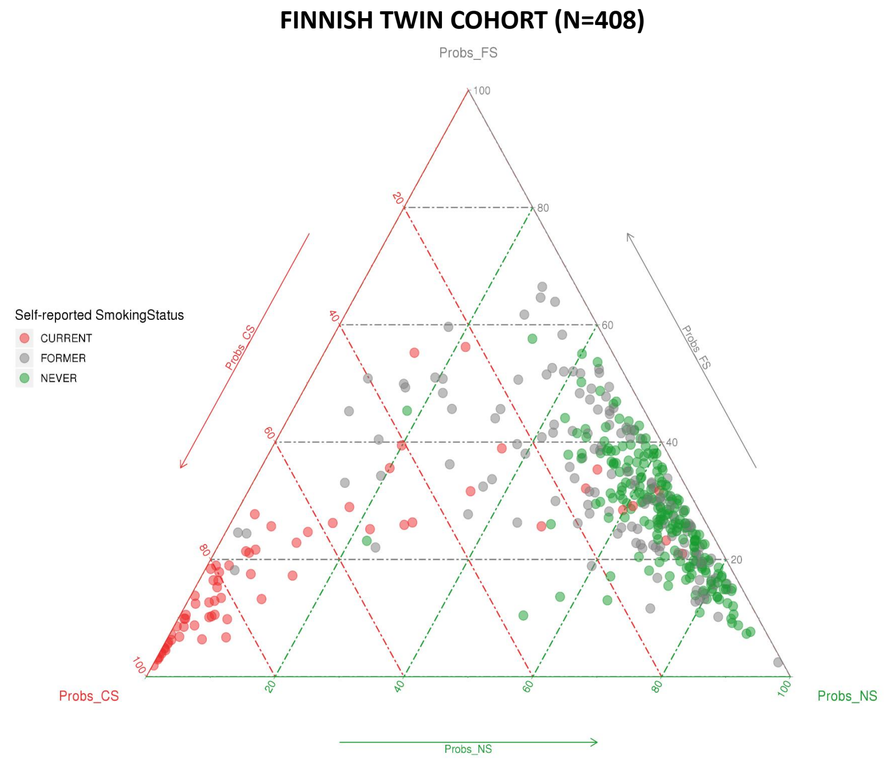

Figure 2: Ternary Plot Showing Estimated Smoking Probabilities in FTC

Ternary plot based on the predicted probabilities given by our classifier in the FTC dataset. The ternary plot is an equilateral triangle with each vertex corresponding to a smoking status. Here color of the points and grid lines corresponds to self-reported smoking status. Each point in the ternary plot represents an individual with the corresponding triplet of probabilities that adds up to 100%. So higher the probability for a class closer is the point to the corresponding corner

3

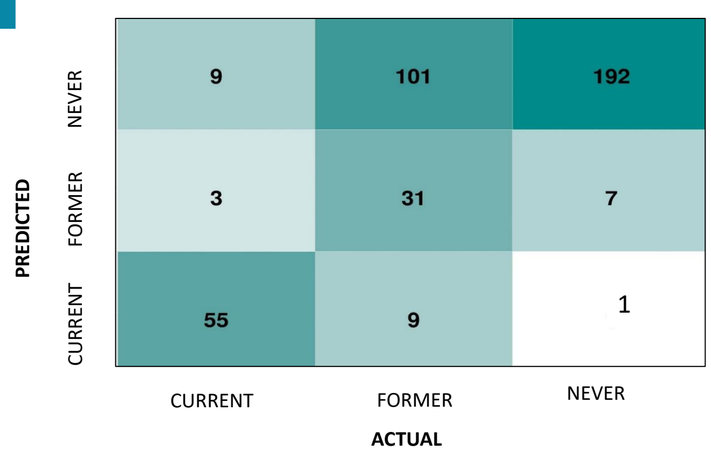

Figure 3: Confusion matrix comparing self-reported versus predicted smoking statuses

Self-reported smoking status is compared with the predicted smoking status determined by the classifier. Numbers along the antidiagonal (↗) indicate correctly predicted cases.

4

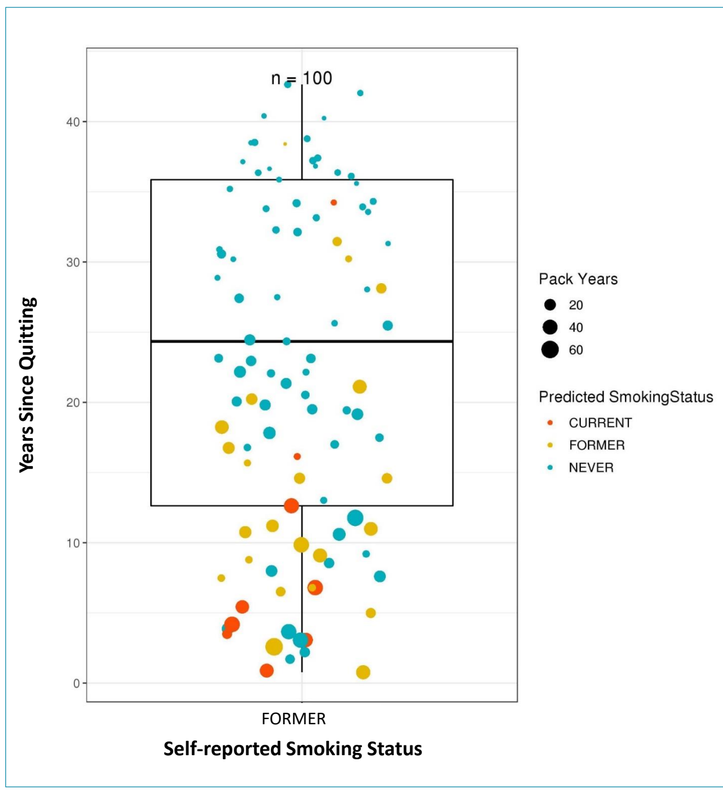

Figure 4: Box and whisker plot showing cessation time (years since quitting) of the self-reported former smokers in the FTC dataset. Here each point indicates a data point and the size of the point is proportional to the number of pack-years. Colors of the points correspond to smoking categories predicted by the classifier. Most of the former smokers with longer cessation time were predicted as never smokers by the classifier.

FIMM - Institute for Molecular Medicine Finland
P.O. Box 20, FI-00014 University of Helsinki, Finland
Biomedicum Helsinki 2U, Tukholmankatu 8, 00290 Helsinki, Finland

www.fimm.fi

In close partnership with:

DocPop Doctoral Programme in Population Health · HUS · National Institute for Health and Welfare · Nordic EMBL Partnership for Molecular Medicine · EMBL · eulife