Project: Learning from Data

Statistical Analysis Plan – Version 2.0
Full study title: Effect of Smoking on the Pulmonary Function Among Children Aged 3 to 19 Years: A Cross-Sectional Study

Group 2: D. Byabazaire, E. Sacla Aide, G. Aga Hirko, M. Sazegar, Q.L. Khuong

1. Background and Objective

Smoking is known to have a negative effect on overall health, particularly on the lungs. In this project, we aim to (1) assess the effect of smoking cigarettes and (2) evaluate the effect of parental smoking on the FEV (Forced Expiratory Volume) of the children.

We hypothesize that smoking and parental smoking can affect FEV of children.

2. Study Design

This study is the cross-sectional subset of a longitudinal observational survey among East Boston area families (USA)[1]. Only one child per family will be included in the study. The sample size is 654 children.

Variables

Outcomes

The outcome of this study is the pulmonary function of the child, measured as the FEV (in liters), which is the volume of air an individual can exhale in the first second of a forceful breath.

Exposures

This study includes two exposure variables according to two research objectives. The first exposure is the smoking status of the child, which is the binary variable (non-smoker/smoker). The secondary exposure variable is the smoking status of the parents (non-smoker/smoker).

Potential Covariates

Potential covariates include age of the child (year), gender (male/female), height (m), BMI (kg/m^2), social-economic status (SES) of the family (low/middle/high), physical activity (average number of hours per week of sports activities and average number of days of sports activities per week), school results of previous school year (poor/average/good), type I diabetes (yes/no), color blind (yes/no), lung disease of the child (yes/no), and the mother education (secondary school/high school/university).

3. Statistical Analysis

Identifying the Set of Confounders

The set of confounders will be selected based on the Directed Acyclic Graphs (DAG)[2], which is the conceptual diagram representing causal relationships of all variables in the network. Based on DAGs, the set of confounders needed to block all confounding paths of the relationship between exposure and outcome, will be obtained. Thus, providing unbiased estimates of the relationship of interest. We construct the DAG for the primary and secondary analysis separately. The DAGs are shown in Figure 1.

Descriptive Statistics

Descriptive statistics will be used to summarize the data, with frequency analysis to describe categorical variables, means (standard deviations) to describe continuous variables which are normally distributed and medians (interquartile range) for skewed distributed continuous variables. Histogram will be used to check

[1] Tager, I., Weiss, S., Rosner, B., and Speizer, F. (1979). Effect of Parental Cigarette Smoking on the Pulmonary Function of Children, American Journal of Epidemiology, 110(1), 15-26.

[2] Shrier, I., & Platt, R. W. (2008). Reducing bias through directed acyclic graphs. BMC Medical Research Methodology, 8(1), 70.

marginal normal distribution of the continuous variables. The percentage of missing values will also be described. The descriptive statistics will be conducted for the overall sample and by groups (smoking status and parental smoking status). Typos and unrealistic values will be checked, unrealistic values will be excluded from the analysis.

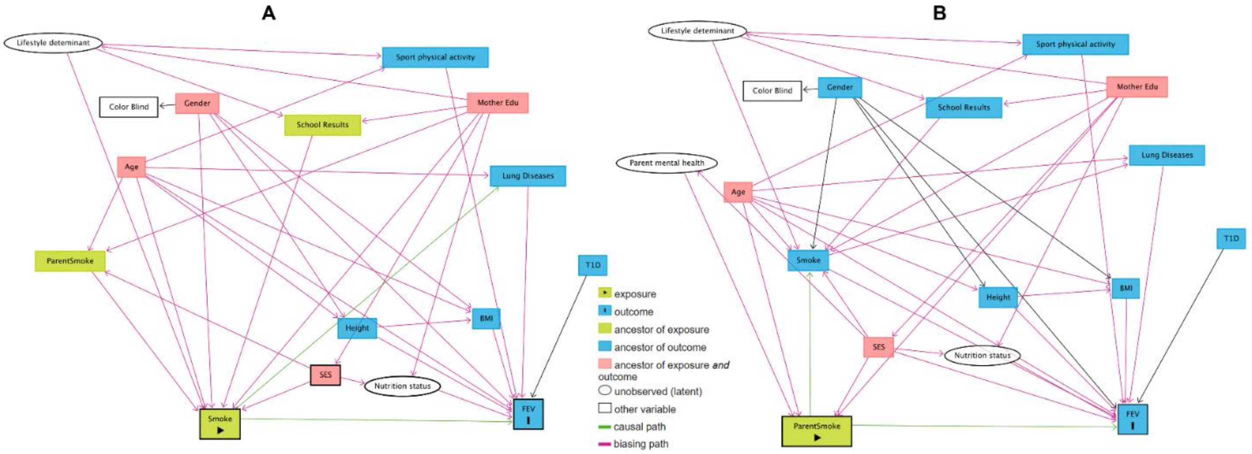

Fig 1: The causal diagrams for relationships between smoking/parental smoking and FEV of children

A: DAG for the relationship between smoking status and FEV of children. A set of confounders includeage, gender, sport activity of the child (average hours per week over previous year), mother education, and household SES; **B**: DAG for the relationship between parental smoking status and FEV of children. A set of confounders include age of the child, mother education, and household SES.

Handling of Missing Data

We will use the margin plot – a basic visualization method to explore the patterns of missing values. The complete case method will be applied to deal with missing values. The child will be excluded from the analysis, if the data for this child is completely missing.

Effects of Smoking and Parental Smoking on the FEV of Children

The multiple linear regressions will be used to assess the effect of smoking and effect of parental smoking on the FEV of the children. The models will be adjusted for set of confounders identified by DAGs.

The results of the regression models will be reported with parameter estimates, 95% confidence interval, and associated p-values. The two-sided hypotheses for the parameter estimates will be tested using t-tatistics. Unadjusted estimates will also be produced for comparison.

Since gender may have differences in the effect of smoking, we will consider the interaction between gender, and smoke in the model for the effect of smoking on FEV of children. If the effect of the interaction term is significant, it will be retained in the model. No interaction term will be considered in the model for the effect of parental smoking on the FEV of children.

Regarding model's assumptions, normality of residuals will be checked using Q-Q plot; linearity and homoscedasticity will be checked using residual versus regressors plot. When the model's assumptions are not met, log transformation will be applied for the regressors which are continuous variables.

We will check outliers using residual plot. Since the models included categorical regressors,the multi-collinearity will be assessed using the Generalized Variance Inflation Factor (GVIF). To make GVIFs comparable across dimensions, the $GVIF^{(1/(2\times DF))}$ criterium will be used. The variables with $[GVIF^{(1/(2\times DF))}]^2 > 10$ are considered to have multicollinearity and excluded from the models[3]. The significant level will be set at 5%. All analyses will be conducted using R version 4.1.2 (The R Foundation for Statistical Computing, Vienna, Austria). The DAGs were constructed using online software DAGitty, available at https://www.dagitty.net.

[3] Fox,John, and Georges Monette. 1992. "Generalized Collinearity Diagnostics." Journal of the American Statistical Association 87 (417): 178–83

Appendix

A list of changes from version 1.0 to version 2.0:

- Added administrative information in the heading
- Added hypothesis statemenet in the *"Background and Objective"* section
- Added the distribution checking step for continuous variables to decide how the descriptive statistic should be reported (i.e., mean (SD) or median (IQR)) in "*Descriptive Statistics"* section
- Replaced outlier checking by typos and unrea;istic values in "*Descriptive Statistics"* section
- Clarified two-sided hypotheses testing using t statistics in "*Effects of Smoking and Parental Smoking on the FEV of Children"* section
- Changed type of multicollinearity control from VIF to GVIF in "*Effects of Smoking and Parental Smoking on the FEV of Children"* section
- Added outlier checking method using residual plots in "*Effects of Smoking and Parental Smoking on the FEV of Children"* section