# Report on lung cancer problem

Karmel Teder

Laura Birgit Luitva

Saskia Kuusk

From the problem synopsis we got the implication that all the data were collected concurrently. Since in that case we could not make any causal inferences, we made some assumptions for the data analysis. The assumptions we made were the following.

1. None of the people included in the sample had lung cancer before inclusion.
2. Values of variables were determined at the time of inclusion.
3. Lung cancer incidence was recorded over a fixed time period.

In addition to making these assumptions, we asked the authors of the problem synopsis if gender existed in the world of this problem and learnt that it does not. Hence it won't be a confounder to any of the variables.

After these assumptions were made, we supposed the causal paths shown on the following DAG.

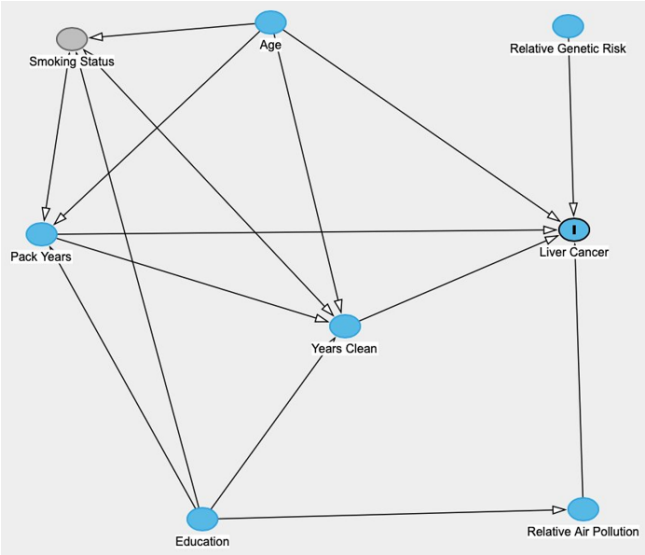


Figure: Visual representation of the assumed causal structure.

Goal of the analysis was to find out how much do individual's age, relative genetic risk score, education, smoking habits, and relative air pollution in their environment contribute to their risk of developing lung cancer. The estimation of direct effects of these previously listed risk factors can be done under the assumption that all biasing paths from these variables to the outcome of interest are blocked. For this DAG, the assumptions are met when including all variables except smoking status, given that in

this data generating process gender is not a confounder. In the DAG we assumed, that smoking status only affects cancer through pack years smoked and years clean from smoking and hence it is not necessary to block its path nor include it in our model. Our hypothesis is that education has no direct effect on developing lung cancer, but this hypothesis does not affect the process of estimating direct effects of the variables. From the model we can later assess whether there is a direct effect of education on lung cancer. Direct effects of all the variables of interest can be estimated from the following model:

*lung_cancer ~ age + relative_genetic_risk + education + pack_years_smoked + years_clean + relative_air_pollution*

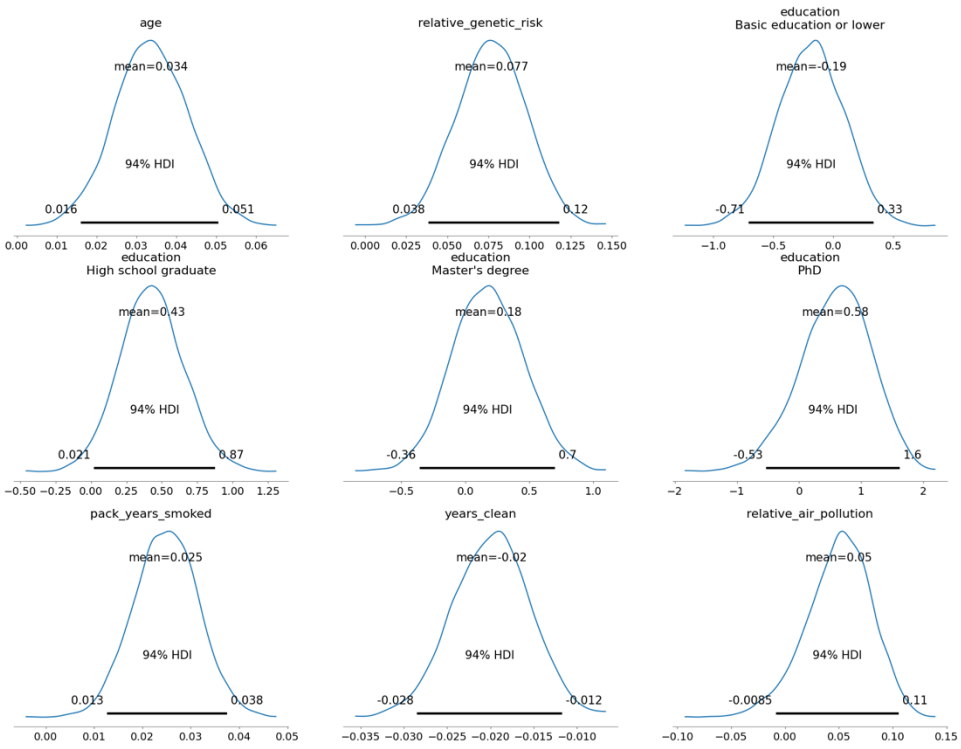


Figure: Posterior distributions for the effect estimates.

From the results we can see that age, relative genetic risk and pack years smoked all increase the risk of developing lung cancer, while years clean of smoking reduces it. Education does not seem to affect the risk systematically and it seems that there could be a positive direct effect of air pollution. It seems that there might be higher risk of developing lung cancer among those who have high school education as their highest obtained education compared to those who have bachelor's degree, but we think that this difference might not be systematic so our hypothesis of no direct effect from education could still hold. In conclusion, the estimates seem quite reasonable, and we are satisfied with our analysis. In the following table we present the summary of the final model.

Table: Summary of the final model.

| | mean | sd | hdi_3% | hdi_97% | mcse_mean | mcse_sd | ess_bulk | ess_tail | r_hat |
|---|---|---|---|---|---|---|---|---|---|
| **Intercept** | -6.307 | 0.506 | -7.240 | -5.324 | 0.007 | 0.005 | 5022.0 | 3230.0 | 1.0 |
| **age** | 0.034 | 0.009 | 0.016 | 0.051 | 0.000 | 0.000 | 4543.0 | 3272.0 | 1.0 |
| **relative_genetic_risk** | 0.077 | 0.021 | 0.038 | 0.118 | 0.000 | 0.000 | 6643.0 | 3142.0 | 1.0 |
| **education[Basic education or lower]** | -0.187 | 0.278 | -0.707 | 0.333 | 0.005 | 0.003 | 3607.0 | 3086.0 | 1.0 |
| **education[High school graduate]** | 0.429 | 0.226 | 0.021 | 0.874 | 0.004 | 0.003 | 3269.0 | 2829.0 | 1.0 |
| **education[Master's degree]** | 0.178 | 0.282 | -0.358 | 0.702 | 0.005 | 0.004 | 3515.0 | 3058.0 | 1.0 |
| **education[PhD]** | 0.584 | 0.576 | -0.531 | 1.618 | 0.008 | 0.006 | 5666.0 | 3079.0 | 1.0 |
| **pack_years_smoked** | 0.025 | 0.007 | 0.013 | 0.038 | 0.000 | 0.000 | 3807.0 | 3248.0 | 1.0 |
| **years_clean** | -0.020 | 0.004 | -0.028 | -0.012 | 0.000 | 0.000 | 3479.0 | 3043.0 | 1.0 |
| **relative_air_pollution** | 0.050 | 0.030 | -0.008 | 0.105 | 0.000 | 0.000 | 6071.0 | 2581.0 | 1.0 |