# Real Data Analysis Critique on “Lung Cancer Incidence Decreases with Elevation: Evidence for Oxygen as an Inhaled Carcinogen”

## 1 Summary of Study

In the paper “Lung cancer incidence decreases with elevation: evidence for oxygen as an inhaled carcinogen”, the authors Kamen P. Simeonov and Daniel S. Himmelstein studied the relationship between elevation and four different types of cancers (lung, breast, colorectal, and prostate), using the data from different counties in Western United States. They ran four hypothesis tests, each corresponding to one of the four cancers, and accounted for the multiple testing procedure by applying Bonferroni correction to have a significance level of 0.0125, which comes from familywise error rate (FWER) of 5%. The authors used two multivariate linear regression methods called best subset and Lasso regression; the latter was adopted to reduce collinearity among covariates and to mitigate overfitting caused by the exhaustive search method that best subset regression uses. Whereas the negative associations between elevation and nonrespiratory cancers disappeared when adjusted for variables such as demographics and other environmental factors, that between lung cancer and elevation was robust to such adjustments, leading the authors to reject the null hypothesis that assumes no relationship between elevation and lung cancer with a p-value of $10^{-16}$, and to fail to reject the three other hypotheses involving non-respiratory cancers.

This report addresses the statistical analysis performed in this paper, with a focus on issues that arise from multiple testing. We define the terms explicit and implicit issues of multiple testing in the corresponding sections, and we state the types of other statistical issues without defining them as they are self-explanatory. In the final section of this report, we discuss the validity of the conclusions that the authors make in their study. All direct quotations and figures come from the paper, which has not been cited as it is the only source of reference for this report.

## 2 Issues of Multiple Testing

*Multiple Testing procedure, as defined in class, is a map from a set of n p-values, denoted as S, to a subset of discoveries. In this report, we define multiple testing issues as issues that arise in a multiple testing procedure where this mapping is invalid, i.e. the type I error that is used to determine whether a test is to be rejected is incorrect, by failing to correctly account for the size of S in setting the significance level. Issues that arise in a multiple testing procedure can be either explicit or implicit.*

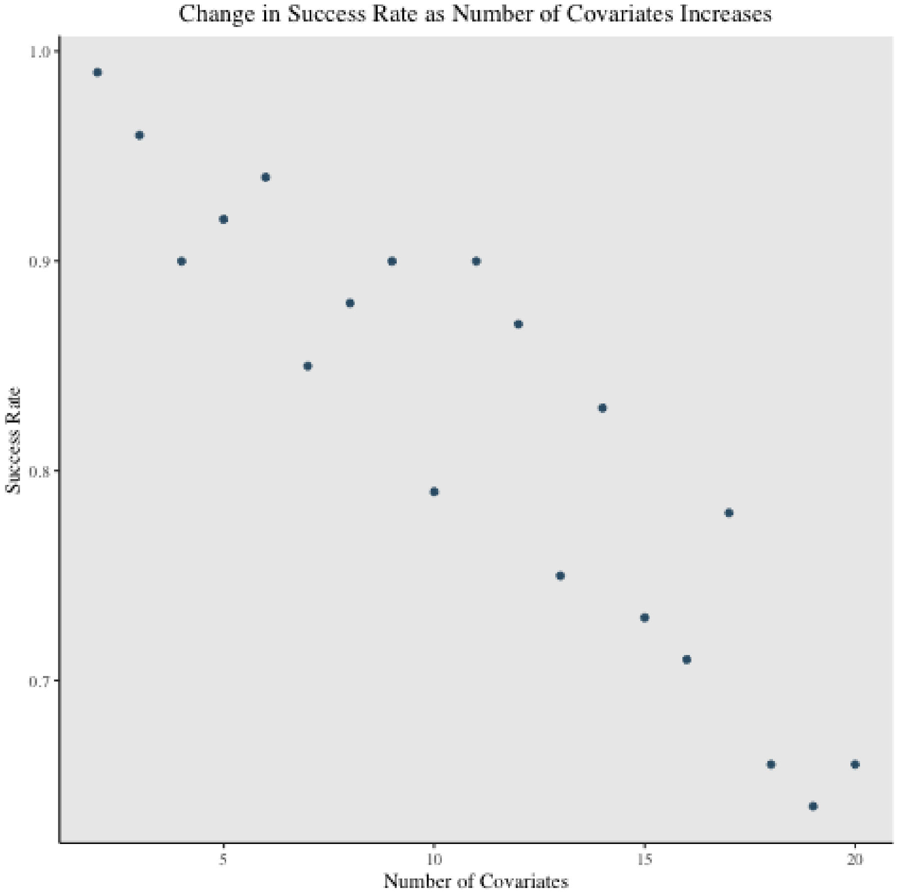
### 2.1 Explicit Issues of Multiple Testing

*In this report, we use the term explicit issues of multiple testing to refer to the issues that arise when the n tests in the set S are run but are not taken into account in setting the threshold false positive rate of the test that is being reported.*

As mentioned in the summary, to account for the issue of testing elevation with multiple cancers on the same dataset, a Bonferroni correction was applied with a threshold p-value of 0.0125. However, this cutoff might have been too conservative, since the different types of cancers could be correlated.

The use of best subset regression for finding a model that minimizes BIC gives rise to issues of explicit multiple testing since BIC does not correct for multiple comparisons. With up to $2^{12}$ model combinations (given that 12 is the maximum number of covariates in a model), there is a good chance that not all of the predictors that were selected were “true” predictors. A possible way to address this problem could be to incorporate cross validation into the procedure; cross validation implicitly accounts for multiple testing since the presence of false positives contributes to increased error on the validation set.

We show in a simulation that the best subset regression procedure used in the paper does not correct for multiple testings and performs worse as the number of covariate p increases. In the simulation, we generated data containing nulls and signals and created a function that runs the best subset regression procedure 100 times with different seeds. We then counted the number of times the procedure selects the correct model and repeated it for the number of covariates p from 2 to 20 while keeping the number of true covariates constant at 2. As can be seen in the plot below, there is a clear downward trend of the success rate as the number of covariates increases, suggesting that the best subset model is not robust to issues of multiple testing. We would like to mention that for this simulation, in order to replicate what the authors did in the paper as closely as possible, we referred to the procedure that they detailed in their codes in their GitHub repository, using the same built-in function called regsubsets from the leaps package with the same settings. Using the same built-in function, we ran the simulation with a toy data set that we generate in the code. The codes for this simulation are included in the appendix.



## 2.2 Implicit Issues of Multiple Testing

*Throughout this report, we use the term implicit issues of multiple testing to refer to the issues that arise when not all tests in the set S are run, and the tests that are not run are not taken into account in setting the threshold false positive rate of the test that is being reported. It is often the case that the existence of those tests that were not run is ignored. These implicit issues of multiple testing often occur when choices are made in a data-dependent way since it is often the case that the act of looking at the data leads to the decision to run particular tests in the set S and not run the others. It should be noted that the set S may be infinite in size.*

One of the most obvious sources of issues of implicit multiple testing in the paper is the possibility that the four tests that were run on the four different cancers could have come from a larger set of possible tests to run, such as a test that studies the relationship between elevation and skin cancer. Let's think of three hypothetical scenarios. In the first scenario, the authors have many different cancers to test in mind, and then look at the data, which guides their decision to test the four different cancers that they picked. In the second scenario, they have prior knowledge, stemming from their domain expertise, that the four different cancers they are studying could have a relationship with elevation. In this case, since the authors would not even have considered the other cancers as potential candidates for their hypothesis tests, their choice of the four tests would not incur any issue of multiple testing. In the third case, let's imagine that the authors incorporated both their prior knowledge and the data to shape their decision: they might have narrowed down the size of the space of possible tests based on their domain-specific knowledge, and then looked at the data to reduce it down to four, let's say, from ten possible tests. In that case, the way in which the authors chose the test would incur problems of multiple comparisons, but to a lesser extent than is the case in scenario 1.

Another example of implicit multiple testing can be found when the authors substitute elevation with each of the seven environmental variables to determine whether or not elevation gives the most likely model. By running best-subset regression with each of the substituted environmental variables and calculating the Bayes factor, the authors were able to conclude that elevation indeed produces the most likely model for lung cancer. Even though it is likely that they implemented this procedure only to justify their choice of using elevation as the environmental factor, the entire study could have been shaped differently had they found out that the other environmental factors produce models that are more likely. For example, had the model with fine particles as a substitute for elevation been more likely, the title of this study could have been "lung cancer incidence increases with the level of fine particles in the air." Even though no explicit test was run on the eight variables (seven substitutes and elevation) with the goal of deciding which hypothesis test to run, the fact that the results of these tests could have influenced the way in which the study was shaped, this procedure gives rise to issues of implicit multiple testing, adding all the potential ways in which the analysis could have been done differently to S, the set of tests that is mapped to the decision to reject the null hypothesis involving lung cancer and elevation.

Another instance of issues of implicit multiple testing arises in tests that were run on the four different subgroups, two age groups and two genders, and the authors test the significance of the same set of covariates they used on the entire population on each of the subgroups. While the results of these four tests do not contribute significantly to the conclusion of the study, the act of running the tests on the subgroups incurs issues of implicit multiple testing. Had there been a model run on a subgroup that turned out to be more significant than the models run on other subgroups and on the entire population, the topic of the study might have been different. As an example, the focus of the study could have been changed to "lung cancer incidence among those older than 65 decreases with the level of elevation." Since the decision on which hypothesis to test, or question to ask, could have been a function of the results of running these tests on the subgroups, the act of running such tests increases the size of the set $S$ consisting of all possible ways in which the question could have been formulated differently, giving room for issues of implicit multiple testing, just as is the case in the example discussed above.

In order to address potential confounding between smoking and elevation, the authors investigate the possibility of interaction between the two variables, by running the model with the interaction term smoking x elevation, the process which gives room for implicit issues of multiple testing. Although it is likely that the authors only intended to study the effect of interaction between smoking and elevation based on some prior knowledge, it is also possible that they decided on this particular interaction term after looking at the data. If this is the case, since the interaction term they chose would have come from a pool of many different interaction terms which had equally valid reasons for being tested prior to looking at the data, testing the significance of the interaction term between smoking and elevation, and not adjusting for the fact that such pool existed, could be considered to be an example of an implicit issue that rises from multiple comparisons. However, in the case that the authors decided on this interaction term prior to looking at the data, this test is valid.

# 3 Other Statistical Issues

*The paper also contains statistical issues that are not related to multiple comparisons, such as issues revolving confounders that are not accounted for, data-dependent choices, missing data, as well as the representation of the population in the data that is not accurate.*

## 3.1 Confounding Variables

There could be confounding effects from variables that were not addressed in the study, which could cause problems in the model. There could potentially be other variables that have a high correlation with both elevation and lung cancer. One example of such variables would be psychological factors, such as stress and depression; if those who live at a higher altitude tend to have a more positive outlook on life, then lung cancer might have also been due to this emotional factor, along with elevation. If these other factors are confounders in the true model, then the model that the study uses to test the relationship between lung cancer and elevation would be false, in which case, the validity of the conclusion would be undermined. However, it is likely the case that the authors of the paper decided on the potential confounders based on their expert domain knowledge, in which case the covariates that they considered would be enough.

## 3.2 Data-Dependent choices

The study implements the best subset and Lasso regression methods in many different instances in their statistical analysis. The act of running the Best Subset regression and Lasso regression to choose the model and using the result from the model to run inference itself is an instance of having a model that depends on the data, which incurs issues of data-dependency, violating the assumptions of their null model; looking at the data of the response variable (cancer) causes the distribution of the response variable to change, due to the dependence of the choices on the data, which is not accounted for by the multivariate linear model that they are using.

Another instance of data-dependency occurs in the way in which the authors decided which test to run. It is implied in the study that the authors shaped their hypothesis based on the results of previous research in the same domain of study. Since the topics are similar, it is likely that this study shares some of its data set with some of the past studies, and if it is actually the case that some of the data set is shared among the studies across different research groups, then the way in which the topic of this paper was chosen depends on the data, although indirectly, to an extent. As is the case with the example discussed above, this complicated dependency of the test on the data is not accounted by the linear model used in the study, or in the case of most traditional models in that matter, causing this instance of data dependence to be a potential statistical issue.

### 3.3 Missing Data

The authors state the ways in which they dealt with missing data, which is to delete them. They mention that they "created cancer-specific datasets by removing counties with any missing data for included variables." Depending on the nature of the missing data, whether they are Missing at Random, Missing Completely at Random, or Missing Not at Random, the choice to discard any observation with missing data might incur problems. Due to the nature of the variables, which are related to cancer, smoking, and obesity, it's likely that the data are not MCAR, in which case discarding the data could create problems in the new design matrix. Moreover, if the missingness depends on the response variable (cancer), which is also likely given the nature of some information, there could be a problem where the distribution of the response variable is different from the distribution assumed by the model.

### 3.4 Representativeness of Population

One of the choices the authors made in processing the data was to remove all counties with Native American population exceeding 25%, five-year immigration rates exceeding 40%, and counties with less than 10000 people. By choosing to remove parts of the data based on certain criteria whose validity is not rigorously justified, it is likely that the population that is represented in the data set is different from the population that the authors intended to represent.

## 4 Conclusion

The study contains several instances of issues of multiple testing as well as other statistical issues. Issues of multiple testing is that the actual, reported false positive rate of the study is smaller than the actual false positive rate, increasing the possibility that the reported discovery is a false positive. The authors conclude that they rejected the null hypothesis of no relationship between elevation and lung cancer, and report the p-value of elevation of less than $10^{-16}$. For this test to be also rejected at the type I error rate of $\alpha = 0.05$ with Bonferroni correction, denoting the p-value of elevation as $p$ and the number of tests in the set $S$ of tests in a multiple testing procedure as $n$, we need $p < \frac{\alpha}{n} = 10^{-16} \implies n \leqslant \frac{\alpha}{p} = \frac{0.05}{10^{-16}} = 5 \times 10^{14}$. Therefore, the number of tests in the set S, the set that is mapped to the decision to reject the aforementioned null hypothesis by the multiple testing procedure in this study, should be less than or equal to $5 \times 10^{14}$ for the null to be rejected. Given that the authors seem to have used a lot of domain-specific prior knowledge in the various choices they make in their analysis, the set of tests, which includes both the tests that were run and were not run, is likely to be smaller than $5 \times 10^{14}$. To elaborate on this point, their largest model contains 12 covariates, and they used the best subset regression with exhaustive search to find a subset of covariates, which incurred an explicit issue of multiple testing, where the number of tests is $2^{12}$, which is still much smaller than $5 \times 10^{14}$. Therefore, due to the very small p-value with which the test was rejected, even if every instance of multiple testing were accounted for in the decision to reject the test, it is likely that the null hypothesis would still have been rejected.

Even though there are other potential statistical issues stemming from confounding variables and data dependency, we believe that the authors' domain knowledge significantly reduced the chances of failing to account for major confounding variables, as well as the degree of issues of data-dependent choices. Moreover, as discussed in class, issues of data-dependency is, in most cases, unavoidable, making it inevitable that empirical statistical analysis diverges from the theoretical statistical method.

Therefore, we conclude that the results of the study, which states the correlation between elevation and instances of lung cancer is high, is valid, and that the authors seem to have carried out the statistical analysis in a rigorous way, addressing statistical issues when possible, such as by using Bonferroni correction for multiple tests they run on the four different types of cancer.

# 5 Appendix

*In our simulation, we closely follow the authors' procedure that they lay out in their script called create-models.R in their GitHub repository https://github.com/dhimmel/elevcan in order to replicate their implementation of the best subset regression as closely as possible in our simulation.*

```
## Simulation for Best Subset Regression in Explicit Issues of Multiple Testing
##--------------------------------------------------------------------------------
# Codes modified from http://www.sthda.com/english/articles/
# 37-model-selection-essentials-in-r/
# 155-best-subsets-regression-essentials-in-r/

n = 259
library(leaps)
library("ggplot2")

# best subset function with x being the number of covariates
best_subset = function(x){
  success_12 = 0
  for (i in 1:100){
    set.seed(i)
    p = x

    # generates data
    X_12 = matrix(rnorm(n * p,mean = 0, sd = 1), n, p)
    X_12 = scale(X_12, center = FALSE, scale = sqrt(colSums(X_12 ^ 2)))
    beta = rep(0, p); beta[1:2] = 5 # two true signals and other betas set to 0
    Y_12 = X_12 %*% beta + rnorm(n)
    data_12 = as.data.frame(cbind(Y_12, X_12))

    # exhaustive search using regsubsets from the leaps package
    model_12 = regsubsets(V1 ~ .,data=data_12, nvmax=x, method='exhaustive')
    res.sum_12 = summary(model_12)
    BIC_12 = which.min(res.sum_12$bic)
    predictors = names(coef(model_12, which.min(res.sum_12$bic)))[-1]
    if (BIC_12 == 2){
      if (predictors[1] == "V2" & predictors[2] == "V3"){
        success_12 = success_12 + 1
      }
    }
  }
  return(success_12)
}

# plots from p=2 to p=20
success_lst = c()
for (i in 2:20){
  success_lst = c(success_lst, best_subset(i))
}
x = 2:20
success_lst = success_lst / 100
plot_sim = ggplot(NULL, aes(x = x)) +
  geom_point(aes(y = success_lst), color = "#325C74") +
  labs(title = "Change in Success Rate as Number of Covariates Increases",
       x = "Number of Covariates", y = "Success Rate") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(text = element_text(size = 11, family = "serif")) +
  theme(axis.line = element_line(colour = "black")) +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())
plot_sim
##---------------------------------------------------------------------------------
```