The causal effect of smoking on gene expression in the lungs

Hadas Biran

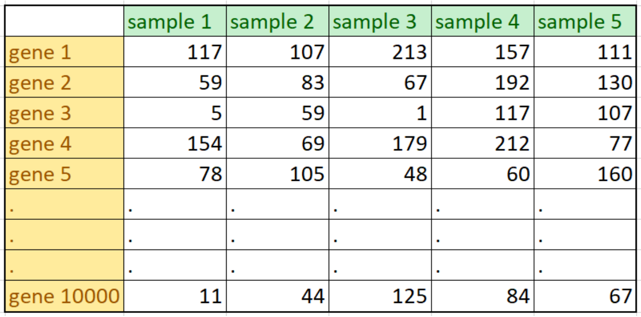
March 2021

1 Introduction

1.1 A short explanation about gene expression

The central dogma of molecular biology states that accurate plans of all proteins are encoded in the DNA. When a cell is in need of a certain protein, it copies the plan from the appropriate gene in the DNA to a newly created messenger RNA (mRNA) strand, and then cell ribosomes use this mRNA to make the protein [Brainard, 2012 (accessed December 3, 2020].

Gene expression profiling methods, such as RNA-sequencing (RNA-seq), generally work by detecting RNA strands, mapping them to genes and quantifying them. So, their output per sample is the number of captured RNA strands of each gene (see Figure 1). These methods have opened the window to a better understanding of active biological processes, gene functions and regulation mechanisms in different tissues and stages of development.



	sample 1	sample 2	sample 3	sample 4	sample 5
gene 1	117	107	213	157	111
gene 2	59	83	67	192	130
gene 3	5	59	1	117	107
gene 4	154	69	179	212	77
gene 5	78	105	48	60	160
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
gene 10000	11	44	125	84	67

Figure 1: A example of a gene expression table. Each column is the output of RNA-seq for a single sample of tissue. For example: 67 RNA strands of gene 2 were detected in sample 3.

1.2 The causal question

The causal question is: what is the short-term effect of smoking on gene expression in normal lung tissues?

Since gene expression may be affected by age [De Magalhães et al., 2009, Glass et al., 2013, Ogueta et al., 1999], gender [Ogueta et al., 1999, Heidecker et al., 2010], race [Patel et al., 2010, Hicks et al., 2013] and BMI [de Souza Batista et al., 2007], and since the likelihood of being a smoker is correlated with each of these characteristics, it is important to answer this question using causal inference methods, in order to overcome confounding biases.

2 Methods

2.1 data collection

I used the Genomic Data Commons (GDC) data portal of the National Cancer Institute (NCI) [Grossman et al., 2016] to select 209 lung cancer patients which fit the following criteria:

- Patients should have a documented evaluation of their smoking habits. This information is kept using a standard measure of the NCI's cancer Data Standards Registry and Repository (caDSR) (data element ID 2181650), which has the following possible values:

  1. Lifelong non-smoker
  2. Current smoker
  3. Current reformed smoker for > 15 years
  4. Current reformed smoker for ≤ 15 years
  5. Current reformed smoker, duration not specified
  6. Smoker at diagnosis
  7. Smoking history not documented

  I only selected patients with values $1 - 6$.

- Patients should have at least one gene expression profile of the lung tissue in the database.

I then downloaded 395 lung gene expression profiles of the selected patients (each patient has $1 - 3$ profiles): 210 of them were samples taken from a tumor and 185 were taken from a normal tissue. The gene expression profiles were produced by RNA-seq.

However, including more than one sample per patient would be a critical violation of the SUTVA (Stable Unit Treatment Value Assumption). Also, gene expression profiles of normal tissues differ greatly from tumorous tissues (see Figure 2), so it seems more appropriate to consider the effect of smoking on each of these populations separately. Since I wanted to focus on the immediate effect of smoking (rather than the long run cumulative effects), I chose to work with the normal samples only. To conclude, my final data set was constructed of gene expression profiles of 185 normal lung tissue samples (one sample per patient). Each gene expression profile consists of the expression values of 57760 genes.

I also downloaded the available clinical information of the patients:

- Age (average age is 64.8)
- Gender (53 women, 132 men)
- Race
- Body mass index (BMI) (average BMI is 25.1)
- Treatment type
- Age at diagnosis

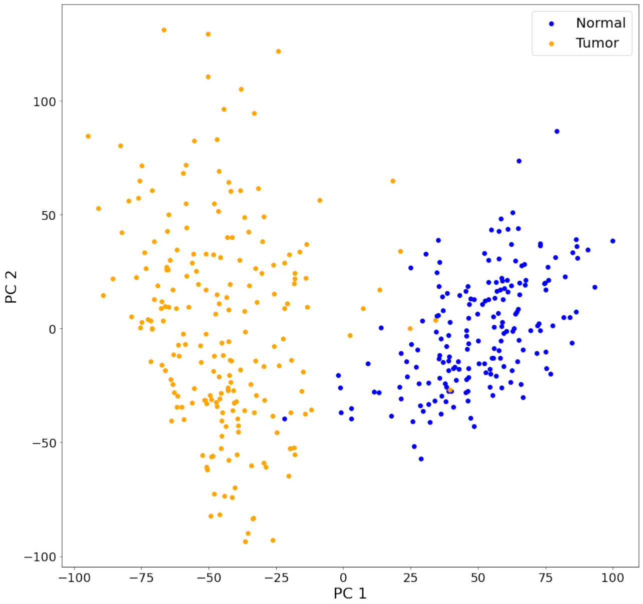


Figure 2: PCA plot of the gene expression profiles.

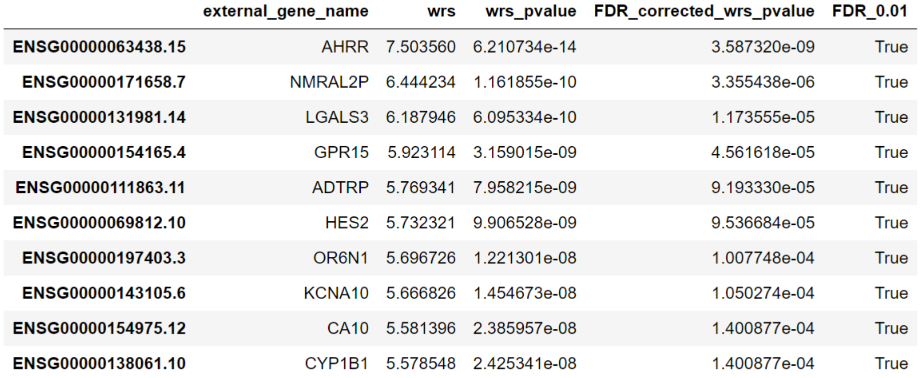
2.2 data preprocessing and preparation

To focus on the immediate effect of smoking, I considered all current smokers (labeled 2) as treated ($T = 1$) and the rest as untreated ($T = 0$). Some data was missing: ”treatment type” was only available for 3 patients and ”race” was only available for 63 patients, so both of these features were excluded. ”Age at diagnosis” was also excluded since it was identical to the ”age” for all patients.

The gene expression data was normalized to ”counts per million” as commonly done in the field.

2.3 Evaluating correlation

I used the Wilcoxon rank-sum test to evaluate the difference in expression values between the treated group (currently smoking) and the untreated group, for each gene separately. Then, I corrected the Wilcoxon p-values by using the Benjamini-Hochberg FDR correction procedure. 435 genes were found to be differentially expressed between the two groups for $FDR = 0.01$ (see the top 10 differentially expressed genes in the table in Figure 3).



	external_gene_name	wrs	wrs_pvalue	FDR_corrected_wrs_pvalue	FDR_0.01
ENSG00000063438.15	AHRR	7.503560	6.210734e-14	3.587320e-09	True
ENSG00000171658.7	NMRAL2P	6.444234	1.161855e-10	3.355438e-06	True
ENSG00000131981.14	LGALS3	6.187946	6.095334e-10	1.173555e-05	True
ENSG00000154165.4	GPR15	5.923114	3.159015e-09	4.561618e-05	True
ENSG00000111863.11	ADTRP	5.769341	7.958215e-09	9.193330e-05	True
ENSG00000069812.10	HES2	5.732321	9.906528e-09	9.536684e-05	True
ENSG00000197403.3	OR6N1	5.696726	1.221301e-08	1.007748e-04	True
ENSG00000143105.6	KCNA10	5.666826	1.454673e-08	1.050274e-04	True
ENSG00000154975.12	CA10	5.581396	2.385957e-08	1.400877e-04	True
ENSG00000138061.10	CYP1B1	5.578548	2.425341e-08	1.400877e-04	True

Figure 3: Top 10 differentially expressed genes based on the Wilcoxon rank-sum test and FDR correction of p-values.

2.4 Evaluating causality

2.4.1 Causal graph of the problem

A directed acyclic graph (DAG) is displayed in Figure 4. The observed features are age, gender and BMI. The unobserved features are race, age at diagnosis, socioeconomic status, treatment type and exposure to other factors.

The backdoor criterion does not apply here, since there are 3 unobserved confounders: race, socioeconomic status and treatment type. As to treatment type, while it may have some effect on a person’s decision to continue smoking, and also some effect on gene expression in adjacent normal tissue (and not just in the tumor site), I assume that both of these effects are minor. As to race and socioeconomic status, I perform sensitivity analysis in Section 2.4.4.

It is worth noting that BMI is a collider which might cause M-bias (with the unobserved confounders race and socioeconomic status). Nevertheless, following Ding and Miratrix [2015]’s advice, I choose to adjust for BMI.

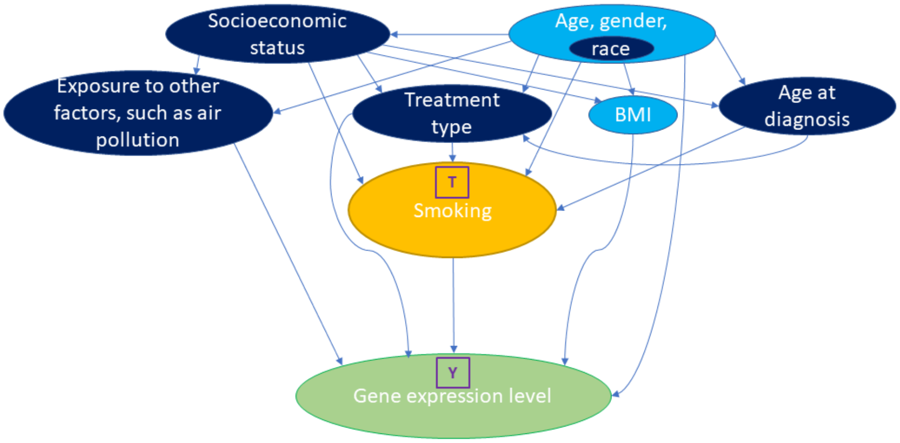


Figure 4: Directed acyclic graph (DAG) of the causal effect of smoking (T=treatment) on gene expression (Y=outcome). Data was available for the features in light blue. Data for the features in dark blue was not available.

I now consider the causal inference assumptions:

1. SUTVA: There's no interference between samples (or patients), so the assumption applies here.
2. Consistency: Relying on RNA-seq's sufficient accuracy and the reliability of the data set, we can say that the consistency assumption also applies here.
3. Ignorability: Does not apply (as discussed above).
4. Common support: As to known covariates, is appears from Figure 5 that the treated and untreated groups are pretty well mixed in the male group, and to a lesser degree in the female group. An evaluation of the common support of propensity scores is done in the following section.

2.4.2 Computing propensity scores

I used logistic regression to compute propensity scores. I then found that age, gender and BMI are (observed) confounders, since they are correlated with the propensity score (see Figure 6).

I evaluated the propensity score common support by observing its distribution (see Figure 7(a)). Since the overlap in the graph is not perfect, I also divided the propensity scores into quintiles and calculated the average propensity score in each quintile, for the smoking and the non-smoking groups separately (as suggested in [Garrido et al., 2014], see Figure 7(b)). Since the average values in each quintile were similar, I concluded that the overlap is satisfactory.

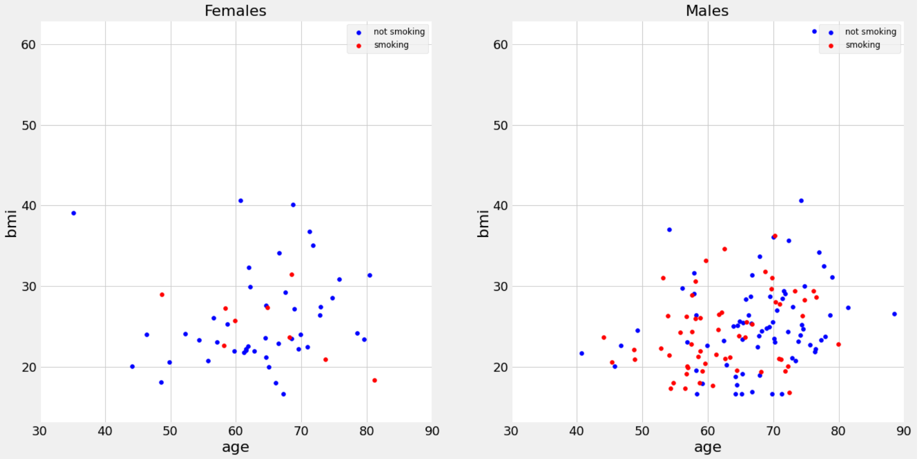


Figure 5: Patients' clinical data: gender, age and BMI.

2.4.3 ATE computation methods

I used the following approaches to compute ATE for each of the 435 differentially expressed genes:

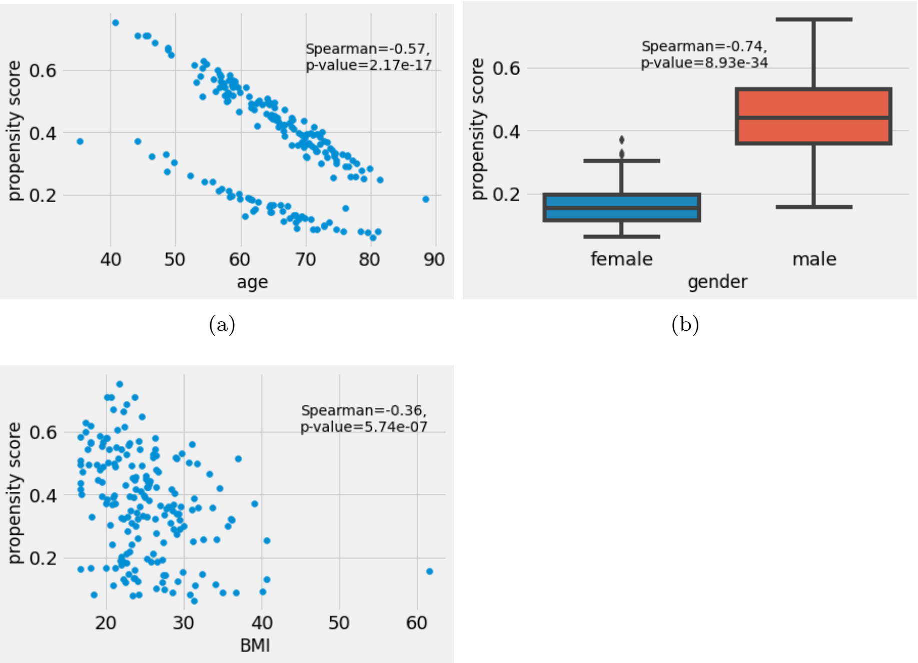
1. Inverse propensity score weighting (IPW)
2. Propensity score matching, with replacement, 1 : 1
3. Propensity score matching, with replacement, 1 : N
4. Covariate matching, with replacement, 1 : 1
5. Covariate matching, with replacement, 1 : N
6. Genetic optimal matching, without replacement, 1 : 1

Following are brief explanations on the methods.

Inverse propensity score weighting (IPW) I computed the average treatment effect (ATE) using the following formula:

$$ATE = \frac{1}{n}\sum_{t_i=1}\frac{y_i}{p(t_i=1|x_i)} - \frac{1}{n}\sum_{t_i=0}\frac{y_i}{p(t_i=0|x_i)} \tag{1}$$

In this case, the ATE is the supposed additional ”counts per million” for a single gene, due to smoking.

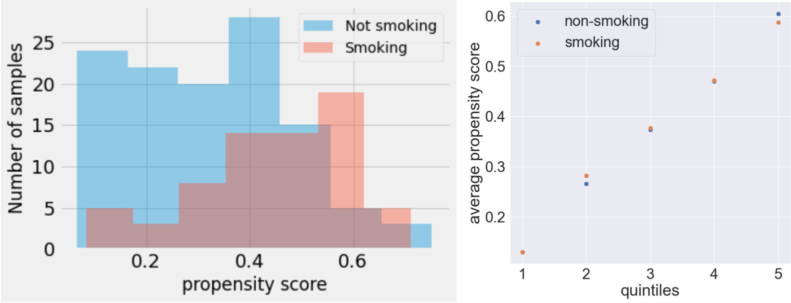


(a)

(b)

(c)

Figure 6: Confounding evidence for the 3 observed features: **(a)** age, **(b)** gender, and **(c)** BMI.



(a)

(b)

Figure 7: **(a)** Propensity score distribution. **(b)** Average propensity score for the smoking and non-smoking in each propensity score quintile.

Propensity score matching, with replacement Following the advice of Austin [2011], I performed matching on the logit of propensity scores, with a caliper width that is equal to 0.2 of the standard deviation of the logit of the propensity score. I used k-nearest neighbors (KNN) with 10 neighbors and euclidean distance to perform both $1:1$ matching and $1:N$ matching.

I denote the group of samples that have at least one counterfactual neighbor within caliper by I_c, and the closest counterfactual neighbor of a sample i by i_j. So, for the $1:1$ matching, the ATE formula is:

$$ATE = \frac{1}{n} \sum_{i \in I_c} (2 \cdot t_i - 1) \cdot (y_i - y_{i_j}) \tag{2}$$

I denote the set of counterfactual neighbors of a sample i that are inside the caliper and are a subset of the 10 closest neighbors of i (counterfactual or not) by $J(i)$. Then, for the $1:N$ matching, the ATE formula is:

$$ATE = \frac{1}{n} \sum_{i \in I_c} (2 \cdot t_i - 1) \cdot \left(y_i - \frac{1}{|J(i)|} \sum_{j \in J(i)} y_j \right) \tag{3}$$

Covariate matching, with replacement I normalized the age and the BMI covariates to the range $0-1$. Then, similarly to the former section (but without a caliper), I performed $1:1$ and $1:N$ matching and computed the ATE for each of these methods.

Genetic optimal matching I used the "rbounds" R package to compute optimal genetic matching ($1:1$, without replacement) of the covariates.

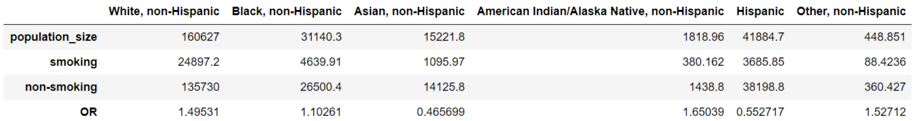
2.4.4 ATE confidence interval computation techniques

I used both (1) bootstrapping and (2) sensitivity analysis to produce confidence intervals to the formerly computed ATEs. Following are brief explanations about these techniques.

Bootstrapping It is possible to use bootstrapping to compute a confidence interval for the ATE (as in [Alves, 2020]). Practically, I sampled 1000 times with replacement from each of the groups (smoking and non-smoking), and computed the ATE as in the original method. I then computed the 95% confidence interval.

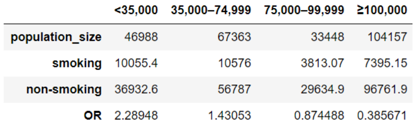
I used bootstrapping to compute confidence intervals for the ATEs computed by methods $1-5$ in Section 2.4.3, for genes with $|ATE| \geq 5$.

Sensitivity analysis I used the "rbounds" R package to perform sensitivity analysis. More specifically, I computed the Rosenbaum Bounds for Hodges-Lehmann Point Estimate [Rosenbaum, 2002] using the "hlsens" function. This algorithm requires the parameter Γ, which is the maximal odds ratio of the relationship between the exposure (smoking) and the unobserved confounder



	White, non-Hispanic	Black, non-Hispanic	Asian, non-Hispanic	American Indian/Alaska Native, non-Hispanic	Hispanic	Other, non-Hispanic
population_size	160627	31140.3	15221.8	1818.96	41884.7	448.851
smoking	24897.2	4639.91	1095.97	380.162	3685.85	88.4236
non-smoking	135730	26500.4	14125.8	1438.8	38198.8	360.427
OR	1.49531	1.10261	0.465699	1.65039	0.552717	1.52712

(a)



	<35,000	35,000–74,999	75,000–99,999	≥100,000
population_size	46988	67363	33448	104157
smoking	10055.4	10576	3813.07	7395.15
non-smoking	36932.6	56787	29634.9	96761.9
OR	2.28948	1.43053	0.874488	0.385671

(b)

Figure 8: Odds ratio (OR) tables for the relationships between: **(a)** race and smoking, and **(b)** household income and smoking (United States, 2019). Numbers for ”population size”, ”smoking”, and ”non-smoking” are in thousands.

(in this case, race and socioeconomic status) [Liu et al., 2013]. To compute it, I downloaded:

- The ”2019 Monthly National Population Estimates by Age, Sex, Race, Hispanic Origin, and Population Universe for the United States” table from United States Census Bureau [b].
- Data regarding smoking habits by race from Center for Disease Control and Prevention (CDC) (also from 2019).
- Number of people in 4 household income ranges in the US in 2019 from United States Census Bureau [a] (I subtracted the number of children to get only the number of adults per household range).
- Data regarding smoking habits by household income from Center for Disease Control and Prevention (CDC) (also from 2019).

Then, for each race vs. all others, I computed the odds ratio of the relationship between the race and smoking. I did a similar protocol for each range of household income (vs. all others). The maximal odds ratio was ~ 2.29, so I ran ”hlsens” with $\Gamma = 2.3$ to get the minimal lower bound and the maximal upper bound of the ATEs (see Figure 8).

I applied sensitivity analysis to compute confidence intervals for ATEs computed by the $1:1$ methods (i.e. methods $2, 4, 6$ in Section 2.4.3).

3 Results

3.0.1 Comparison between ATE computation methods

I computed the ATE using each of the 6 methods, for all 435 genes that were found to be differentially expressed between the smoking and the non-smoking groups, based on a Wilcoxon rank sum test. I then computed the Pearson correlation between methods (see Figure 9). I chose to present here the Pearson correlation (rather than Spearman), since the values themselves have biological meaning (not just their order). While all the methods seems to be fairly correlated (Pearson coefficient above 0.95 for all method-pairs), the two methods that differ the most were optimal genetic matching and IPW.

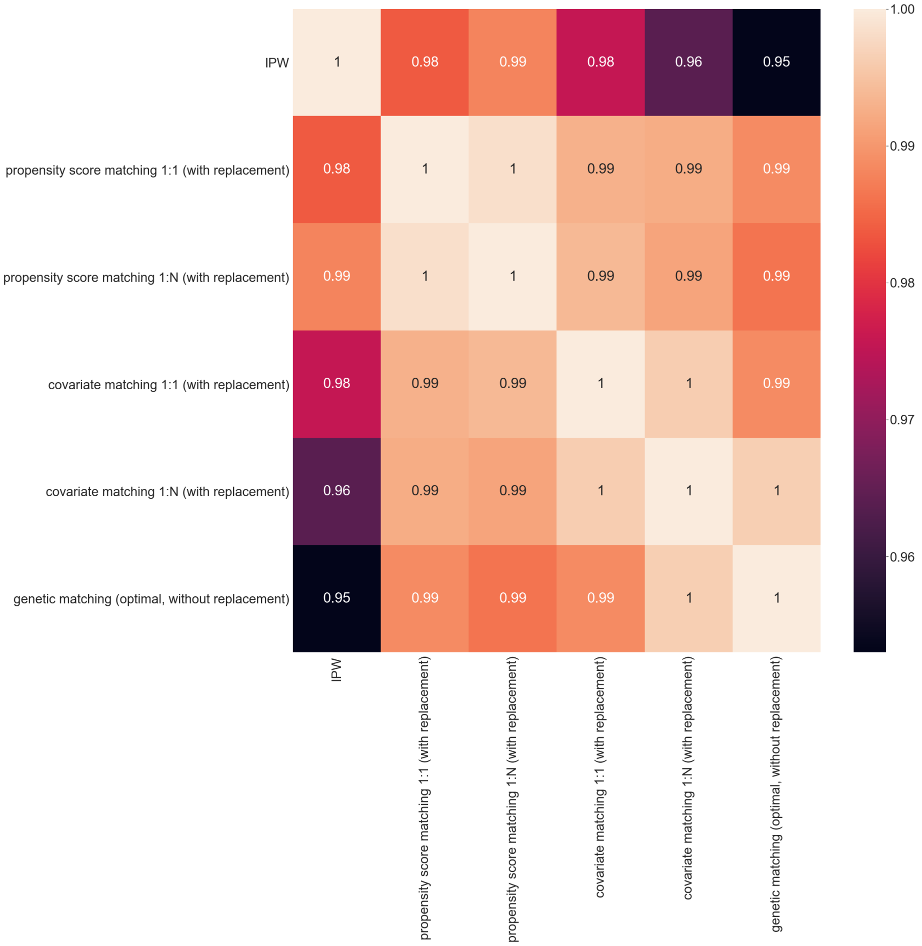


Figure 9: Pearson correlation scores between the different methods of ATE computation, over 435 (differentially expressed) genes.

3.0.2 The relationship between ATE and confidence interval size

In general, the confidence interval tend to be wider for larger ATEs. This is true both for the confidence intervals produced by bootstrapping and for the confidence intervals produced by sensitivity analysis (see Figure 10).

We can use the bootstrapping procedure to determine the methods robustness to the exact composition of the data set. Methods with narrower confidence intervals and a low rate of change in respect to the ATE should be considered to be robust. As appears from Figure 10(a), IPW seems to be the least robust, followed by covariate matching (with replacement, $1:1$). Propensity score matching (with replacement, $1:1$) is the most robust method here.

3.0.3 Comparison between bootstrapping and sensitivity analysis

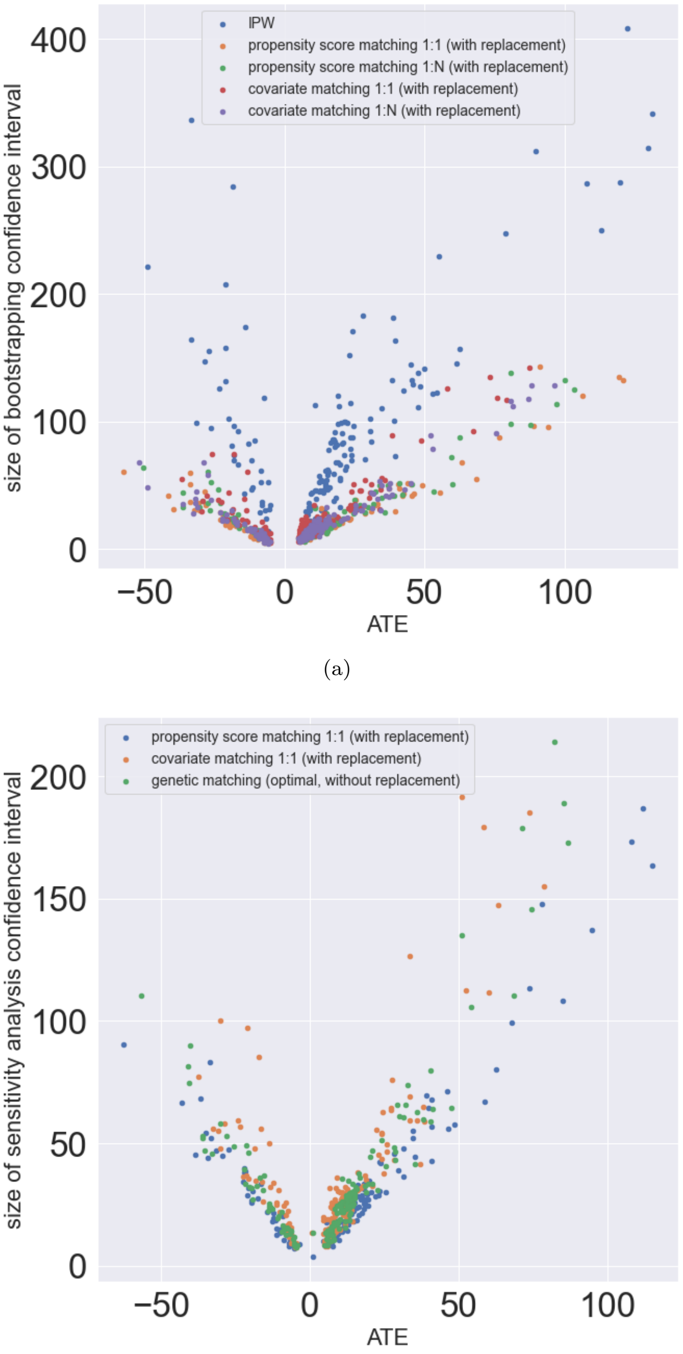
For propensity score matching and covariate matching (both with replacement, $1:1$) I computed confidence intervals both by bootstrapping and by sensitivity analysis. Figure 11 displays both of these confidence intervals for genes with ATE in the range $[5, 250]$ (for a clearer visualization). It generally appears that the sensitivity confidence interval is more conservative (wider) in this case, probably due to the large Γ (see Section 2.4.4). We can conclude that in cases with obvious unobserved confounders such as this one, a simple bootstrapping procedure is not enough to get a reliable confidence interval, and it is necessary to perform sensitivity analysis.

3.0.4 Comparison between final gene sets

For each each of the three $1:1$ ATE computation methods, I created a set of genes with $|ATE| \geq 5$. I then excluded from this set genes with ATE confidence interval (produced by sensitivity analysis) that included 0 (that is, a negative lower bound and a positive upper bound), since the expression levels of these genes might not be affected at all by smoking. The gene sets included 95 genes for genetic matching (optimal, without replacement), 177 genes for propensity score matching (with replacement) and 62 genes for covariate matching (with replacement). I used a Venn diagram to evaluate the level of overlap of the final gene sets (see Figure 12).

4 Limitations

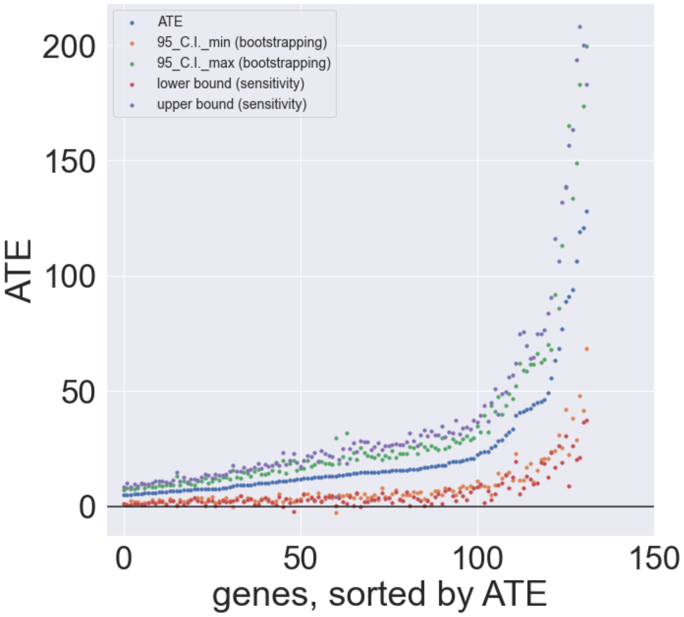
In this work I examined the effect of smoking on gene expression in normal lung tissues. The most obvious caveat here is that all samples were taken from cancer patients. Even though I only considered samples from healthy tissues, it is certainly possible that small metastatic processes had already began in some of the samples, or that normal tissues are somehow affected by cancer treatment. To get a more reliable answer to the causal question (which focused on the short-term effect of smoking), we must add to the analysis samples from healthy patients.



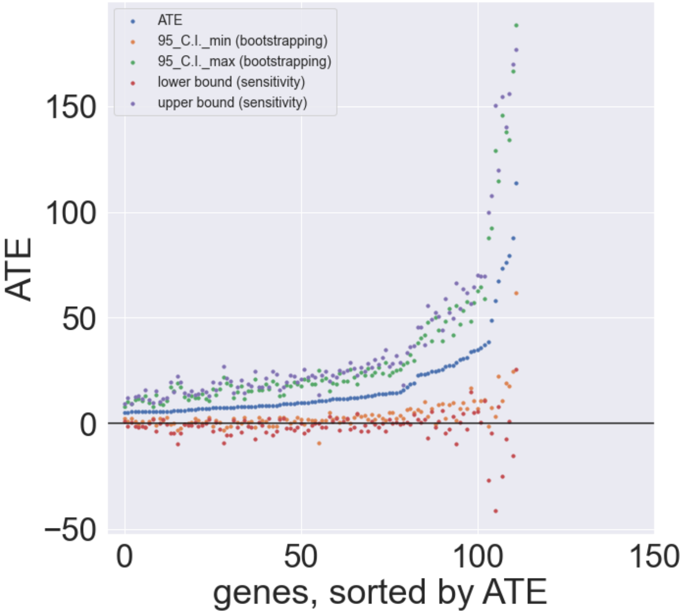
(a)

(b)

Figure 10: Width of confidence interval vs. ATE, for genes with ATE $\in [5, 100]$. Confidence intervals were produced by: **(a)** bootstrapping and **(b)** sensitivity analysis.



(a)



(b)

Figure 11: ATE confidence intervals produced by both bootstrapping and sensitivity analysis, for genes with ATE $\in [5, 250]$. **(a)** Propensity score matching with replacement, $1:1$. **(b)** Covariate matching with replacement, $1:1$.

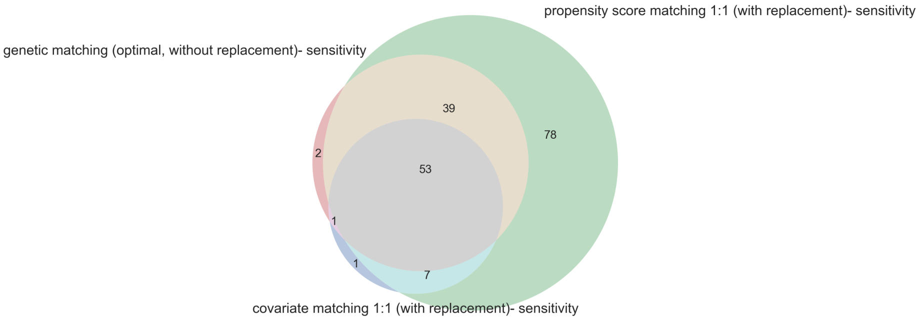

propensity score matching 1:1 (with replacement)- sensitivity
genetic matching (optimal, without replacement)- sensitivity
39
78
2
53
1
1
7
covariate matching 1:1 (with replacement)- sensitivity

Figure 12: Final gene sets of three 1 : 1 methods, with confidence interval produced by sensitivity analysis.

Also, I only tested causation for the 435 genes that were found to be differentially expressed by Wilcoxon. However, since causation without correlation is possible [Frakt], it makes sense to test the causal effect of smoking on the expression levels of all other genes as well.

5 Discussion

In this work I used 6 ATE computation methods to evaluate the short-term change in expression level of genes in normal lung tissue due to smoking. I found that the ATE estimations of the 6 methods were overall similar to each other. I also examined 2 techniques which estimate the ATE confidence interval: bootstrapping and sensitivity analysis. While bootstrapping simply uses random sampling with replacement, the sensitivity analysis procedure was tailored to the specific unobserved confounders. I found that in this case, sensitivity analysis usually provided a wider confidence interval. Using the first technique, bootstrapping, I was also able to infer that IPW is far more sensitive to the exact composition of the data set than propensity score matching. Finally, I found that smoking has a causal effect on the expression levels of at least 53 genes, based on their ATEs (computed by genetic matching, propensity score matching and covariate matching) and their confidence intervals (computed by sensitivity analysis).

6 Code repository

The project code wad uploaded to https://github.com/hadasbi/SmokingGeneExpression.

References

Matheus Facure Alves. *Causal Inference for the Brave and True*. 2020. URL https://matheusfacure.github.io/python-causality-handbook/11-Propensity-Score.html. Accessed on March 2021.

Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424, 2011.

Jean Brainard. *CK-12 Biology*. FlexBook, 2012 (accessed December 3, 2020). URL https://bio.libretexts.org/Bookshelves/Introductory_and_General_Biology/Book%3A_Introductory_Biology_(CK-12)/04%3A_Molecular_Biology/4.01%3A_Central_Dogma_of_Molecular_Biology#:~:text=The%20central%20dogma%20of%20molecular%20biology%20states%20that%20DNA%20contains,DNA%20to%20RNA%20to%20Protein.

Center for Disease Control and Prevention (CDC). Burden of cigarette use in the u.s. URL https://www.cdc.gov/tobacco/campaign/tips/resources/data/cigarette-smoking-in-united-states.html#by_race. Accessed on March 2021.

João Pedro De Magalhães, João Curado, and George M Church. Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics*, 25(7):875–881, 2009.

Celia M de Souza Batista, Rong-Ze Yang, Mi-Jeong Lee, Nicole M Glynn, Dao-Zhan Yu, Jessica Pray, Kelechi Ndubuizu, Susheel Patil, Alan Schwartz, Mark Kligman, et al. Omentin plasma levels and gene expression are decreased in obesity. *Diabetes*, 56(6):1655–1661, 2007.

Peng Ding and Luke W Miratrix. To adjust or not to adjust? sensitivity analysis of m-bias and butterfly-bias. *Journal of Causal Inference*, 3(1):41–57, 2015.

Austin Frakt. Causation without correlation is possible. URL https://theincidentaleconomist.com/wordpress/causation-without-correlation-is-possible/. Accessed on March 2021.

Melissa M Garrido, Amy S Kelley, Julia Paris, Katherine Roza, Diane E Meier, R Sean Morrison, and Melissa D Aldridge. Methods for constructing and assessing propensity scores. *Health services research*, 49(5):1701–1720, 2014.

Daniel Glass, Ana Viñuela, Matthew N Davies, Adaikalavan Ramasamy, Leopold Parts, David Knowles, Andrew A Brown, Åsa K Hedman, Kerrin S Small, Alfonso Buil, et al. Gene expression changes with age in skin, adipose tissue, blood and brain. *Genome biology*, 14(7):1–12, 2013.

Robert L Grossman, Allison P Heath, Vincent Ferretti, Harold E Varmus, Douglas R Lowy, Warren A Kibbe, and Louis M Staudt. Toward a shared vision for cancer genomic data. *New England Journal of Medicine*, 375(12):1109–1112, 2016.

Bettina Heidecker, Guillaume Lamirault, Edward K Kasper, Ilan S Wittstein, Hunter C Champion, Elayne Breton, Stuart D Russell, Jennifer Hall, Michelle M Kittleson, Kenneth L Baughman, et al. The gene expression profile of patients with new-onset heart failure reveals important gender-specific differences. *European heart journal*, 31(10):1188–1196, 2010.

Chindo Hicks, Lucio Miele, Tejaswi Koganti, LaFarra Young-Gaylor, Deidre Rogers, Vani Vijayakumar, and Gail Megason. Analysis of patterns of gene expression variation within and between ethnic populations in pediatric b-all. *Cancer informatics*, 12:CIN–S11831, 2013.

Weiwei Liu, S Janet Kuramoto, and Elizabeth A Stuart. An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention science*, 14(6):570–580, 2013.

Sandra B Ogueta, Steven D Schwartz, Clyde K Yamashita, and Debora B Farber. Estrogen receptor in the human eye: influence of gender and age on gene expression. *Investigative ophthalmology & visual science*, 40(9):1906–1911, 1999.

Tejal A Patel, Gerardo Colon-Otero, Celyne Bueno Hume, John A Copland III, and Edith A Perez. Breast cancer in latinas: gene expression, differential response to treatments, and differential toxicities in latinas compared with other population groups. *The Oncologist*, 15(5):466, 2010.

Paul R. Rosenbaum. Springer New York, New York, NY, 2002.

United States Census Bureau. Hinc-03. people in households-households, by total money income, age, race and hispanic origin of householder., a. URL https://www.census.gov/data/tables/time-series/demo/income-poverty/cps-hinc/hinc-03.html. Accessed on March 2021.

United States Census Bureau. National population by characteristics: 2010-2019, b. URL https://www.census.gov/data/tables/time-series/demo/popest/2010s-national-detail.html. Accessed on March 2021.