Causal Inference Final Project: Effect of Smoking on 10-year Development of Coronary Heart Disease

Bianca Doone, Michael Attah, Graham Casey Gibson, Daniel Saunders, Nutcha Wattanachit

November 25, 2018

Background Story

Coronary heart disease (CHD) is the leading cause of death and serious illness in the United States. The Framingham Heart Study's objective was to identify the common factors or characteristics that contribute to CHD by following its development over time in a large group of participants who had not yet developed overt symptoms of CHD or suffered a heart attack or stroke.

The researchers recruited 5,209 men and women between the ages of 30 and 70 from Framingham, Massachusetts, and began the first round of extensive physical examinations and lifestyle interviews that they would later analyze for common patterns related to CHD development. Over the years, careful monitoring of the Framingham Study population has led to the identification of the major CHD risk factors – high blood pressure, high blood cholesterol, smoking, obesity, diabetes, and physical inactivity. We are interested how the extent of smoking affects the development of CHD, specifically, it is not immediately obvious whether smoking 5 cigarettes per day affects the development of CHD differently than smoking 15 cigarettes per day does.

Causal Roadmap

Step 0: Specify the Scientific Question

What is the effect of smoking on the ten-year development of Coronary Heart Disease?

Target population

The target population is white middle-class men and women aged 30 to 70 in the US.

The sameple in this study is white middle-class men and women aged 30 to 70 (at baseline) in Framingham, Massachusetts. The importance of the major CHD risk factors identified in this group have been shown in other studies to apply almost universally, even though the patterns of distribution may vary. Thus, we are willing to generalize to the target population.

Step 1: Specify a Causal Model

- Endogenous nodes: $X = (W, Z, A, Y)$, where
- W is age, gender, education
- Z is blood pressure (systolic and diastolic), total Cholesterol, prevalence of hypertension, prevalence of stroke, heart rate, BMI, Diabetes prevalence
- A is the number of cigarettes smoked per days
- Y is the ten-year development of coronary heart disease (CHD).

- Exogenous nodes: $U = (U_W, U_Z, U_A, U_Y) \sim \mathbb{P}_U$. We make no assumptions about the distribution $\mathbb{P}_U$.
- Structural equations F:

$$\begin{aligned} W &\leftarrow f_W(U_W) \\ Z &\leftarrow f_Z(W, A, U_Z) \\ A &\leftarrow f_A(W, U_A) \\ Y &\leftarrow f_Y(W, Z, A, U_Y) \end{aligned}$$

There are no exclusion restrictions or assumptions about functional form.

Causal Graph

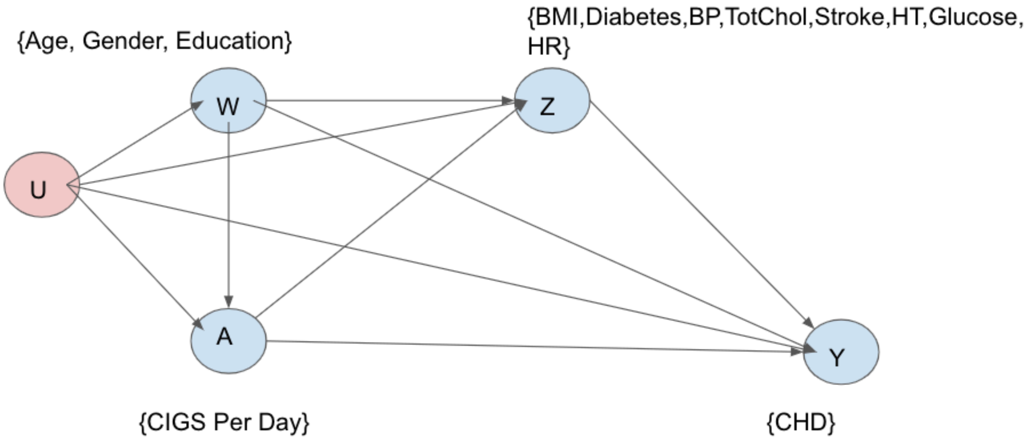

Figure 1: Causal Graph for the SCM

Step 2: Counterfactuals & Causal Parameter

Causal Parameter

$$\Psi^{*i}(\mathbb{P}^*) = \mathbb{E}^*[Y_i] \quad i \in \{1, 2, 3, 4\}$$

where i represent the bin of cigarettes smoked per day. Y_i denotes the counterfactual outcome (the ten-year development of cardiovascular disease), if possibly contrary to fact, a person's number of cigarettes smoked per day is within i^{th} bin.

Step 3. Specify your observed data and its link to the causal model

The dataset is adapted from Framingham Heart Study. We assume that Gender, Age, Education, and Number of Cigarettes per Day (A) were collected in a questionnaire at baseline. Then, BMI, Diabetes Status, Prevalence of Stroke, Prevalence of Hypertension, Indication of Blood Pressure Medication, Total Cholesterol Level, Blood Pressure, and Heart Rate were all collected after the questionnaire at a doctor's office. Our outcome, Coronary Heart Disease, is collected at a 10-year follow up. Note that this is unlike the original study. We assume our observed data were generated by sampling n from a system described by our structural causal model, so we have $n = 4211$ copies of $O \overset{i.i.d}{\sim} \mathbb{P}_O$. We place no restrictions on the statistical model $\mathcal{M}$, which is thereby non-parametric. BMI was binned using guidelines from the World

Health Organization. Total Cholesterol was binned using guidelines from the National Heart, Lung and Blood Institute (NHLBI). Table 1 below shows the counts for each variable in each bin of the exposure, as well as a χ^2-test of independence.

Table 1: Number of Observations in Each Bin

	Level	[0,1)	[1,10)	[10,19)	[20,70]	p
n		2089	471	380	1168	
diabetes (%)	0	2022 (96.8)	463 (98.3)	372 (97.9)	1145 (98.0)	0.079
	1	67 (3.2)	8 (1.7)	8 (2.1)	23 (2.0)	
prevalentStroke (%)	0	2071 (99.1)	469 (99.6)	377 (99.2)	1166 (99.8)	0.095
	1	18 (0.9)	2 (0.4)	3 (0.8)	2 (0.2)	
prevalentHyp (%)	0	1337 (64.0)	338 (71.8)	293 (77.1)	860 (73.6)	<0.001
	1	752 (36.0)	133 (28.2)	87 (22.9)	308 (26.4)	
age (%)	[32, 42)	363 (17.4)	117 (24.8)	110 (28.9)	309 (26.5)	<0.001
	[42, 49)	464 (22.2)	137 (29.1)	125 (32.9)	401 (34.3)	
	[49, 56)	527 (25.2)	97 (20.6)	71 (18.7)	254 (21.7)	
	[56, 70]	735 (35.2)	120 (25.5)	74 (19.5)	204 (17.5)	
education (%)	1	915 (43.8)	188 (39.9)	137 (36.1)	470 (40.2)	0.003
	2	574 (27.5)	150 (31.8)	121 (31.8)	399 (34.2)	
	3	367 (17.6)	76 (16.1)	71 (18.7)	170 (14.6)	
	4	233 (11.2)	57 (12.1)	51 (13.4)	129 (11.0)	
BP (%)	0	1139 (54.5)	267 (56.7)	211 (55.5)	701 (60.0)	0.025
	1	950 (45.5)	204 (43.3)	169 (44.5)	467 (40.0)	
totChol (%)	[0, 200)	386 (18.6)	108 (23.4)	84 (22.3)	227 (19.7)	0.004
	[200, 240)	723 (34.9)	151 (32.7)	157 (41.8)	401 (34.9)	
	[240, 600]	962 (46.5)	203 (43.9)	135 (35.9)	522 (45.4)	
gender (%)	0	1400 (67.0)	346 (73.5)	222 (58.4)	386 (33.0)	<0.001
	1	689 (33.0)	125 (26.5)	158 (41.6)	782 (67.0)	
bmi (%)	[0, 18.5)	19 (0.9)	13 (2.8)	7 (1.8)	17 (1.5)	<0.001
	[18.5, 25)	785 (37.8)	245 (52.2)	225 (59.2)	568 (48.8)	
	[25, 30)	940 (45.3)	170 (36.2)	117 (30.8)	467 (40.1)	
	[30, 56.8]	333 (16.0)	41 (8.7)	31 (8.2)	112 (9.6)	
heartRate (%)	[0, 60)	122 (5.8)	27 (5.7)	19 (5.0)	28 (2.4)	<0.001
	[60, 143]	1967 (94.2)	444 (94.3)	360 (95.0)	1140 (97.6)	
CHD (%)	0	1784 (85.4)	420 (89.2)	320 (84.2)	958 (82.0)	0.002
	1	305 (14.6)	51 (10.8)	60 (15.8)	210 (18.0)	

Step 4. Identifiability

Since we made no independence assumptions on our exogenous background factors, we will need to make additional independence assumptions for identifiability. For the target causal parameter in the SCM $\mathcal{M}^*$ to be identified from the observed data distribution, we need to make a randomization and a positivity assumption.

1) Randomization Assumption

We could assume that all unmeasured background factors in our SCM are independent, which is sufficient, but not minimally sufficient. In the augmented/working SCM ($\mathcal{M}^{**}$) that we selected, the unmeasured background factor of A (cigarettes smoked per day) is independent of the unmeasured background factor of Y (10yr CHD), the unmeasured background factor of W (baseline age, gender, education, diabetes, BMI), and the unmeasured background factor of Z (prevalence of stroke, hypertension, blood pressure, blood pressure medication, heart rate). Thus, conditional on W, the counterfactual outcome is independent of the observed treatment: $Y \perp A|W$.

Since W,Z, and Y include SES and biological factors that affect human health, we aviod assuming independence between their unmeasured background factors. Thus, we consider it is more plausible to make the indepedence assumptions listed. We do not adjust the mediator Z to avoid opening a backdoor path. Under M^{**}, the backdoor criterion holds conditional on W. Additional data on factors that affect heath status and determinants could help with identifiability, but those factors are not well-understood.

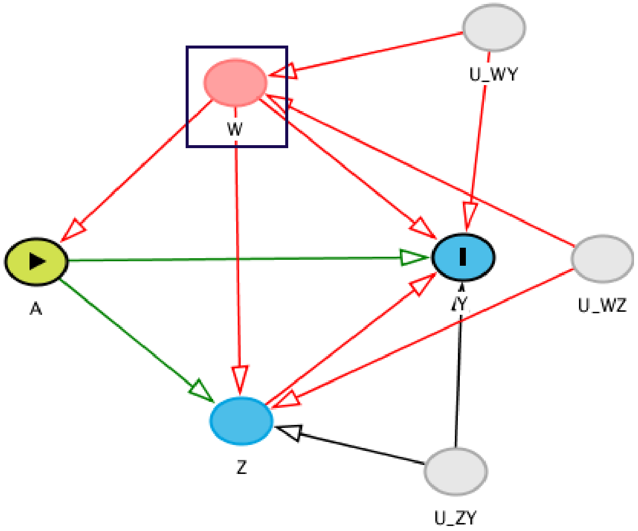

Figure 2: Causal Graph for an augmented SCM

$$U_A \perp U_Y, U_A \perp U_W, U_A \perp U_Z$$

2) Positivity Assumption

There must be a positive probability of each treatment condition within each possible strata of W. We need a positive probability of cigarettes smoked per day in bin i for each strata of W.

$$min_{i \in A} \mathbb{P}_0(A = i|W = w) > 0$$
$$\text{for all } w \text{ for which } \mathbb{P}_0(W = w) \geq 0$$

where i denote the index of a bin of A.

We are concerned about a positivity assumption violation since binning covatiates could make particular stratas of W have low probabilities of smoking certain numbers of cigarettes per day, especially a bin with high numbers of cigarettes smoked per day. We can informally check for a positivity assumption violation from tables of A given a strata of W.

From the table from one of the stratas of W below, we can see that we have no female who smokes 10 to less than 20 cigarettes per day with missing education level in the age ranges [42, 49) and [56, 70] respectively.

Thus, we have some sparsity issue which could affect the estimator performance, particularly the inverse probability of treatment weighting (IPTW).

Table 2: Table for Gender = 0 (Female), cigsPerDay (A) = [10,20)

Strata	Education: 1	Education: 2	Education: 3	Education: 4	Education: NA
Age: [32, 42)	16	26	16	11	1
Age: [42, 49)	23	33	17	10	0
Age: [49, 56)	16	12	8	2	3
Age: [56, 70]	15	11	5	1	0

We also investigated the treatment mechanism by calculating the predicted probability of each bin of A (the number of cigarettes smoked per day) given a strata of W:

Table 3: Predicted Probabilities of A for each Strata of W

	Min	Mean	Max
0 cigarettes per day	0.2853671	0.5085200	0.7440546
[1,10) cigarettes per day	0.0612951	0.1146543	0.1733320
[10,19) cigarettes per day	0.0592084	0.0925024	0.1428044
[20,70] cigarettes per day	0.0765399	0.2843233	0.5630990

From the table's minimum probability column, we can see that some stratas of W have relatively low probabilities for the number of cigarettes smoked per day in the 2nd, 3rd, and 4th bin, which would results in those stratas having high weights in the IPTW estimator. However, the probabilities are not close to zero. Therefore, we do not have a practical violation of positivity assumption. Theoretically, randomizing the number of cigarettes smoked per day could rid of positivity assumption violation concerns, but it is not feasible.

Step 5. Statistical Model and Estimand

The target parameter of $\mathbb{P}_0$, which equals the causal parameter in the augmented causal model $\mathcal{M}^{**}$ is given by the G-Computation formula:

$$\begin{aligned}\Psi_0(\mathbb{P}_0^i) &= \mathbb{E}_0[\mathbb{E}_0[Y|A=i, W=w]] \\ &= \sum_w \mathbb{E}_0[Y|A=i, W=w] * \mathbb{P}_0(W=w)\end{aligned}$$

where i represent the bin of cigarettes smoked per day.

Step 6. Estimation

In order to estimate the causal effect of smoking on risk of CHD we evaluate 6 separate estimators. The first is a classical estimate made by logistic regression. The second is a simple subsitution estimator. The next 3 are variations of the IPTW estimator and finally we evaluate the TMLE estimator. All confidence intervals presented below are based on $1,000$ bootstrap samples, expect in the case of IPTW-IC where the influence curve confidence intervals were obtained and in the case of the classical GLM where theoretical CI were obtained. This was done mostly to evaluate the performance of the bootstrap confidence agains the theoretical confidence derived from the influence curve.

Classical Model Based on Chi-Squared

In order to comapre the causal based estimators to traditional statistics, we first fit a generalized linear model as follows.

$$\begin{aligned} logit(CHD) \sim \beta_0 + \beta_1 * \text{education} + \beta_2 * \text{age} + \beta_3 * \text{total cholesterol} \\ + \beta_4 * \text{prevelant Hyp} + \beta_5 * \text{BP} + \beta_6 * \text{diabetes} + \epsilon \end{aligned}$$

Here we have included all variables that were considered signficiantly correlated with the outcome under the χ^2 test of independence. This includes conditioning on mediator variables such as a diabetes, a distinct diference from the causal models presented below. In order to obtain an estimate of $E(Y|A = a_i)$ we simply call the predict method on our existing data (with no specific intervention) and average the results.

Simple Substitution

In order to evaluate our causal estimate we implemented a non-parametric simple substiution estimator for the conditional mean outcome. We used a saturated logistic regression model that included all possible interaction terms to model $P(Y|A = a, W)$. The resulting estimator is given for the i^{th} bin by

$$SS^i = E_0[E_0[Y|A = a_i, W]]$$

Confidence intervals were obtained via the bootstrap.

Inverse Probability of Treatment Weighting (IPTW)

In order to evaluate our causal estimate we examined three variations on the traditional IPTW estimator as follows.

For the i^{th} bin:

$$IPTW^i = \frac{1}{n_i} \sum_{j}^{n_i} Y \frac{\mathbb{I}[A = i]}{P(A = i|W)}$$

We evaluated 3 IPTW estimators:

- the simple IPTW estimator
- the Horvitz-Thompson weighted IPTW estimator
- the IPTW estimator with variance derived from influence curve (IC) calculations

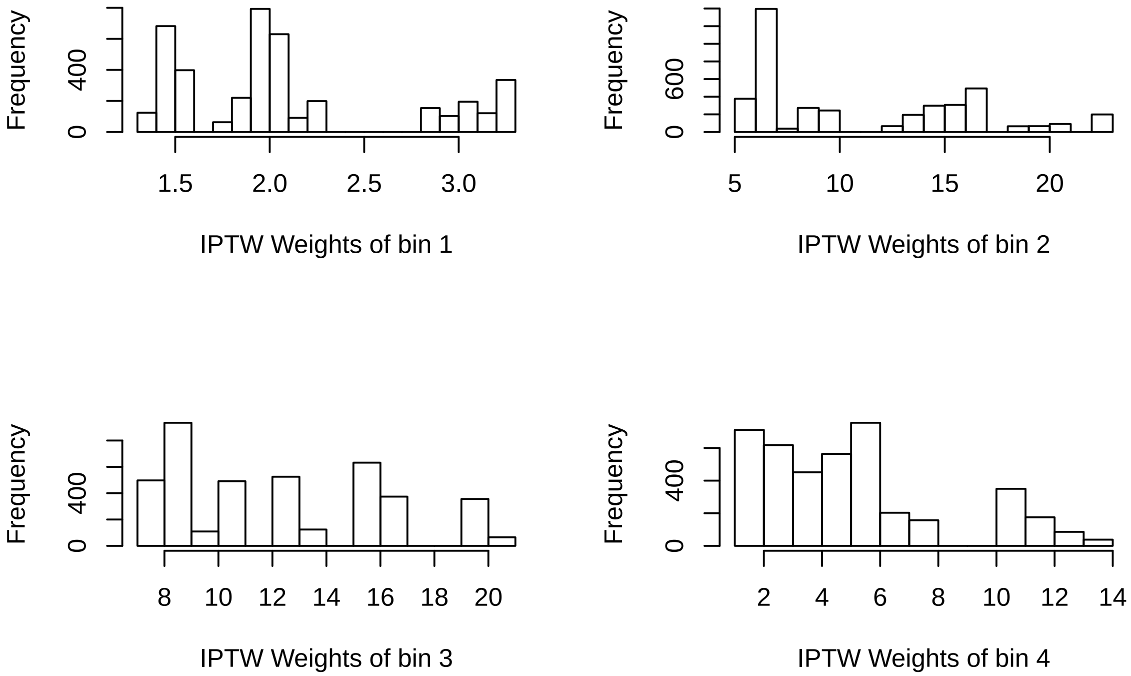

Figure 3: Distribution of Weights

We can see from the weight histograms that the second bin has the largest discrepency in the weighting. This may help explain why we see a protective effect of smoking in the second bin.

Targeted Maximum Likelihood Estimator (TMLE)

Finally, we examined our causal estimate under targeted maximum likelihood. In our superlearner library we evaluated 4 candidate algorithms:

- Penalized regression using elastic net (SL.glmnet)
- Random forest (SL.randomForest)
- Neurel network (SL.nnet)
- Multivariate adaptive regression splines (SL.earth)

Both our estimate for the conditional mean outcome and the treatment probability were obtained from superlearner.

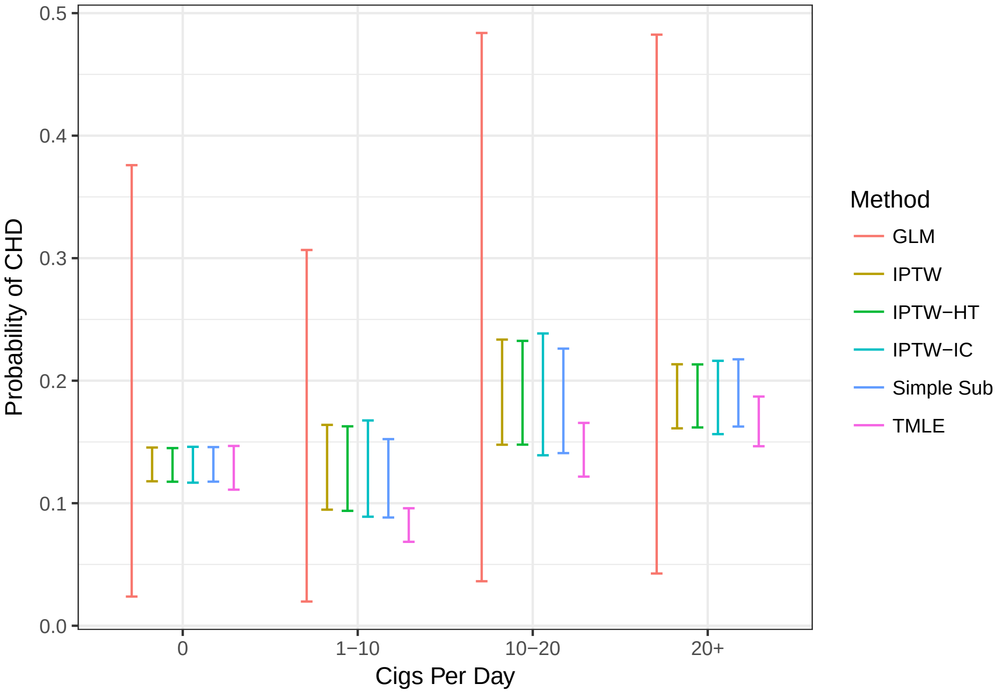

Figure 4: Confidence Intervals for Estimators

As we can see from Fig. 4 the glm estimate is highly variable compared with the causal estimator. In addition we see that estimates given by the Horvitz-Thompson IPTW and simple IPTW estimators are quite comaprable with only slightly smaller confidence intervals under horvitz thompson, as to be expected. The theoertical intervals obtained from the influence curve for the IPTW estimator were slightly larger than the bootstrap intervals, which might be an issue with the bootstrap sample size. Finally, the estimates from TMLE were conservative with respect to the other estimators

Step 7. Result Interpretation

Statistical

From the estimation of our statistical model $\mathcal{M}$, we found that the probability of developing coronary heart disease increases as the binned exposure (`cigsPerDay`) increases from, standardadized to the distribution of baseline covariates W. However, in bin #2, corresponding to `cigsPerDay` $\in [1, 10)$, we found what seems to be a "protective" effect, in which smoking less than a pack of cigarettes as day is associated with a lower risk of CHD than not smoking at all. The difference risk in this bin is estimated at around 0.10-0.11, compared with . We argue that this may be due to *non-differential misclassification*, since participants may have not self-reported smoking due to the negative stigma associated with it. For `cigsPerDay` ≥ 20, the probability of CHD plateaus near 0.2 (for the conditional mean outcome and IPTW estimators) and 0.16 (for TMLE).

Counterfactual

The estimated effect of the exposure is the counterfactual probability of developiong coronary heart disease in 10 years under the exposure category `cigsPerDay` $(A_i, i \in \{1, 2, 3, 4\})$, with the assumptions that conditioning on the baseline covariates W (age, sex, & education) and assumpting certain independence assumptions satisfies the backdoor criterion, and said independence assumptions and positivity holds. The observed general increase in estimated risk of CHD as the (binned) exposure increases is certainly plausible; however, the estimated reduction in risk of CHD for `cigsPerDay` $\in [1, 10)$ is less plausible, again perhaps due to biased data from non-differential misclassification on the exposure, or could in fact be evidence for a protective effect of low degrees of smoking on risk of CHD.

Limitations

It is not likely that our target causal parameter is identifiable due to violations of the chosen independence assumptions, and thus our counterfactual estimation should not be accepted without some healthy skepticism. Many unmeasured confounders are possible, including pre-existing medical conditions not considered by the Framingham Heart Study patient surveys, diet and exercise habits, family history, genetic factors, and more. Future studies might expand the set of measured variables to include such information. As mentioned, we are also concerned about potential issues with reporting bias in the data. Finally, we did not have access to time-varying aspects of the data; e.g., smoking habits and biomarkers before or after baseline measurements were taken, many of which could contain information relevant to prediction of CHD outcomes.

We also evaluated the sensitivity to bin-sizing to evaluate the seemingly protective effect smoking $[1, 10)$ cigarettes has. To investigate, we first removed all binning and treated each observed smoking value as a unique factor. We then plotted the resulting estimates of each $E(Y|A = a)$ for all a levels of cigarettes per day, along with the corresponding sample sizes used to estimate the counterfactual.

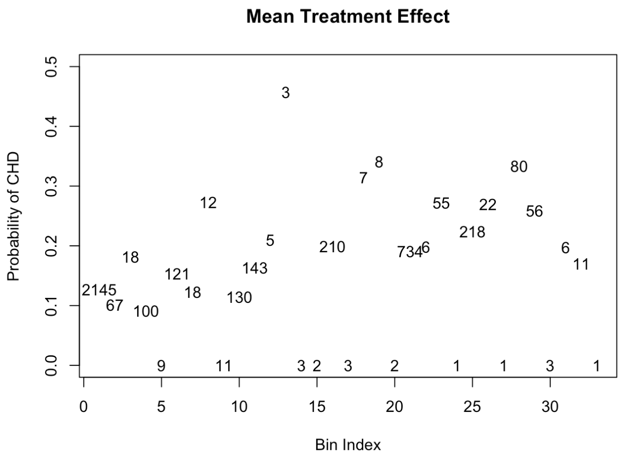

Figure 5: Expected Outcome

As we can see, there are quite a few levels of cigarettes per day that have a low sample size (<10). In order to correct for this, we first removed all estimates of $E(Y|A = a)$ for all levels of cigarettes with samples size < 10. We next fit a loess curve to the resulting estimates to get a sense of the overall trend.

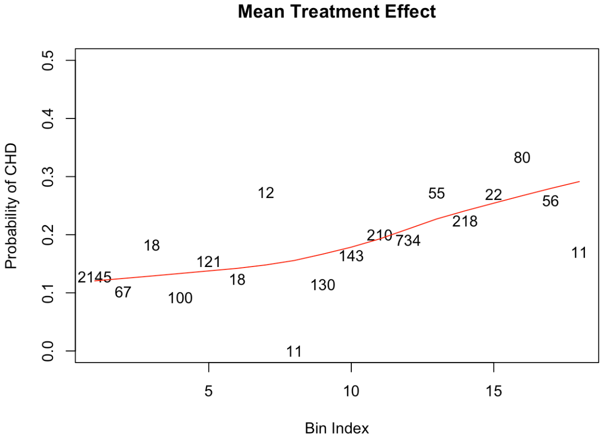

Figure 6: Expected Outcome

We can see from Fig 6 that by removing the bins with low sample size we see a general positive trend as cigarettes per day increase. Although the protective effect requires further exploration, bin size does seem to play an important role.

Team Member Contribution

Member	Contributions
Michael Attah	- group discussion of causal question - design of causal graph - design of descriptive table - categorize covariates: Z - interpretation: causal - references
Bianca Doone	- group discussion of causal question - design of causal graph - time-varying assumptions - chi-squared test of indepedence - categorize covariates: W (age, education, gender) - bootstrap confidence intervals: IPTW - interpretation: statistical - limitations
Casey Graham	- group discussion of causal question - design of causal graph - categorize covariates: cigsPerDay - computation of simple substitution - computation of IPTW - computation of Superlearner/TMLE - bootstrap confidence intervals: TMLE, GLM
Daniel Saunders	- group discussion of causal question - design of causal graph - design of descriptive table - statistical estimand: G-computation - bootstrap confidence intervals: TMLE, simple substitution - statistical estimand: G-computation
Nutcha Wattanachit	- group causal question - design of causal graph - background story and target population - specifying causal parameter - design of working SCM ($\mathcal{M}^{**}$) - specify observed data and link to SCM - identifiability: randomization assumption - identifiability: positivity assumption

References

Boston University & the National Heart, Lung, & Blood Institute, Framingham Heart Study. 5 Dec. 2018: https://www.framinghamheartstudy.org

World Health Organization. (2018). Body mass index - BMI. 29 Nov. 2018: http://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi

The National Heart, Lung, & Blood Institute, ATP III Guidelines At-A-Glance. 29 Nov. 2018: https://www.nhlbi.nih.gov/files/docs/guidelines/atglance.pdf