Final Project:
Childhood Bullying and Subsequent Drug Use

Shelley Facente, Steph Holm, Lizzy Kinnard, Veronica Pear

Spring 2019

For our final project, we have worked to answer an important causal question using a real-life dataset and applying the steps of the causal roadmap:

1. **Specify causal model representing real background knowledge.**
2. **Specify causal question.**
3. **Specify observed data and link to causal model.**
4. **Identify: Knowledge + data sufficient?**
5. **Commit to the best estimand possible, and an appropriate statistical model.**
6. **Estimate.**
7. **Interpret results.**

Background

The relationship between bullying and drug use has previously been explored. This association has been examined both among youth who are perpetrators of bullying and youth who are victims of bullying. A 2016 meta-analysis found that youth who bully are at least twice as likely compared with non-involved students to use drugs later in life (OR = 2.22, 95% CI: 1.60-3.07). However, when adjusting for confounding variables, the adjusted summary effect size was markedly reduced to an OR of 1.41 (95% CI: 1.20-1.66), suggesting that much of the variation is explained by other contributing factors.[1]

According to a 2012 paper, youth involved in bullying were more likely than students not involved in bullying to use substances, with bullying victims reporting the greatest levels of substance use.[2] Longitudinal analyses have shown that youth who experience mental or physical bullying, separately or in combination, were more likely to subsequently report use of substances (alcohol, cigarettes, marijuana, and inhalants). This finding held after controlling for baseline covariates (gender, grade level, ethnicity and substance).[3]

Drug use in adolescence or adulthood has been associated with adverse health outcomes, such as substance use disorder, overdose, infectious disease acquisition, and other major medical illnesses. Preventing bullying victimization may have downstream effects by preventing substance use initiation.[4]

To our knowledge, no studies have evaluated the relationship between bullying victimization and drug use using causal inference approaches. This study fills a gap in the literature by studying this question in a causal framework.

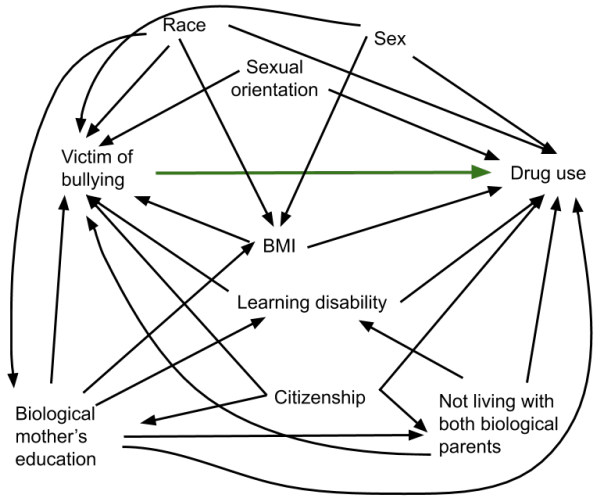
Step 1. Specify a Causal Model

Our data are from the National Longitudinal Survey of Youth 1997. This was recruited as a nationally representative cohort of youth ages 12-16 (initial n=9000) in 1997. At baseline, youth were interviewed as well as one of their parents. These youth have since been followed longitudinally. Our analytic dataset includes 7703 subjects, for reasons we will explain in greater detail in subsequent sections.

The target population is youth in the United States.

Causal graphs

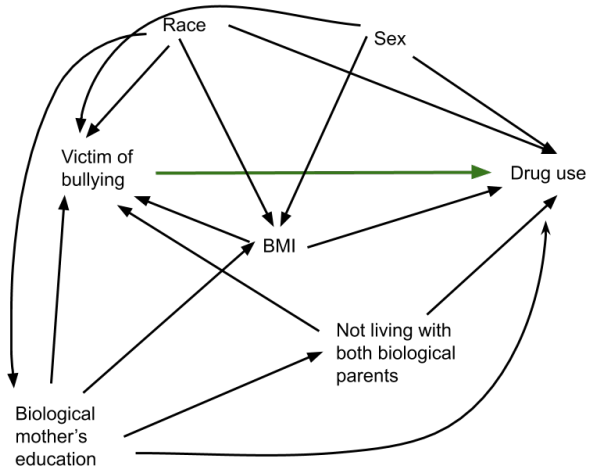
This is the original DAG we developed while examining what was available in the dataset:



While our initial DAG included sexual orientation, citizenship, and learning disability as covariates, we decided to exclude them from our final DAG for the following reasons (which will be discussed in more detail in the Positivity section).

- **Sexual orientation** was predominantly reported as same-sex partners only, and our measure of association was homogenous across strata of sexual orientation (p=0.09).
- Upon inspecting the NLSY codebook, **citizenship** was actually county of birth, which is not the same underlying construct we were trying to capture with citizenship, and very few subjects reported being born outside of the United States.
- Finally, **learning disability** actually captured a learning or emotional problem, and we decided that the latter could be influenced by our exposure (i.e. there could be temporal concerns), in which case it would be on the causal pathway and we would not want to control for it.

Therefore, below is our final DAG for this analysis:



Structural Equations

Our endogenous nodes include: $X = (W, A, Y)$, where $W = (W_1, W_2, W_3, W_4, W_5)$ is the set of baseline covariates, where

- W_1 = mother's education
- W_2 = sex
- W_3 = race/ethnicity
- W_4 = not living with both biological parents
- W_5 = BMI z-score

A = bullied before the age of 12 (asked in 1997), and
Y = incident drug use ("cocaine or other hard drugs") after 1997.

Our background variables (exogenous nodes) include: $U = (U_W, U_A, U_Y) \sim \mathbb{P}_U$.

We place no assumptions on the distribution $\mathbb{P}_U$. We have not placed any restrictions on the functional form.

Our structural equations $\mathcal{F}$ are:

$$\begin{aligned}
W_1 &= f_{W_1}(U_{W_1}, W_3) \\
W_2 &= f_{W_2}(U_{W_2}) \\
W_3 &= f_{W_3}(U_{W_3}) \\
W_4 &= f_{W_4}(U_{W_4}, W_1) \\
W_5 &= f_{W_5}(U_{W_5}, W_1, W_2, W_3) \\
A &= f_A(U_A, W_1, W_2, W_3, W_4, W_5) \\
Y &= f_Y(U_Y, A, W_1, W_2, W_3, W_4, W_5)
\end{aligned}$$

It is clear from these structural equations that we are making exclusion restrictions, most notably that race (W_3) and sex (W_2) have no endogenous parents. We will discuss our independence assumptions during step 4 of the roadmap (identification).

Step 2. Specify Causal Question

Our causal question is:

What is the effect of having been bullied prior to age 12 on incidence of drug use in adolescence or adulthood?

Target Causal Parameter

Given our causal question, our target causal parameter is the difference in the counterfactual probability of drug use if all kids were bullied prior to age 12, and the counterfactual probability of drug use if all kids were not bullied prior to age 12, represented by the following equation:

$$\psi^F(P_{U,X}) = P_{U,X}(Y_1 = 1) - P_{U,X}(Y_0 = 1) = E_{U,X}(Y_1) - E_{U,X}(Y_0)$$

where Y_a denotes the counterfactual outcome under an intervention to set bullying status $A = a$. This target causal parameter is the average treatment effect (ATE), or causal risk difference.

Step 3. Specify observed data and link to causal model

Our Observed Data

As previously described, our observed data includes a sample of 7,703 people who participated in the National Longitudinal Survey of Youth 1997. Table 1 includes descriptive frequencies of our exposure and covariates, stratified by outcome.

Table 1

Covariate	Drug use (%)	No drug use (%)
Drug use (Total)	1330 (17.3%)	6373 (82.7%)
Victim of bullying		
Yes	319 (4.1%)	1179 (15.3%)
No	1011 (13.1%)	5194 (67.4%)
Mother's education		
High school or less	732 (9.5%)	3867 (50.2%)
Some college or more	598 (7.8%)	2506 (32.5%)
Sex		
Female	591 (7.7%)	3218 (41.8%)
Male	739 (9.6%)	3155 (41%)
Race/ethnicity		
Black	227 (2.9%)	1788 (23.2%)
Hispanic	288 (3.7%)	1340 (17.4%)
Non-Black, Non-Hispanic	815 (10.6%)	3245 (42.1%)
Living with both biological parents		
Yes	645 (8.4%)	3176 (41.2%)

Covariate	Drug use (%)	No drug use (%)
No	685 (8.9%)	3197 (41.5%)
BMI z-score	0.513 (*mean*) 1.03 (*sd*)	0.505 (*mean*) 0.98 (*sd*)

In our sample, approximately 19.5% of youth had been bullied before age 12, and 17.3% had incident drug use during follow-up.

The variables with missing data were mother's educational status (with 7% of the observations missing) and BMI (with 5% of the observations missing). We used multiple imputation by change equations to impute these values, as implemented in the mice package in R. For this assignment, we only used one imputed dataset for simplicity. However, if we refine this analysis for publication, we will impute multiple datasets to account for uncertainty in the imputation procedure.

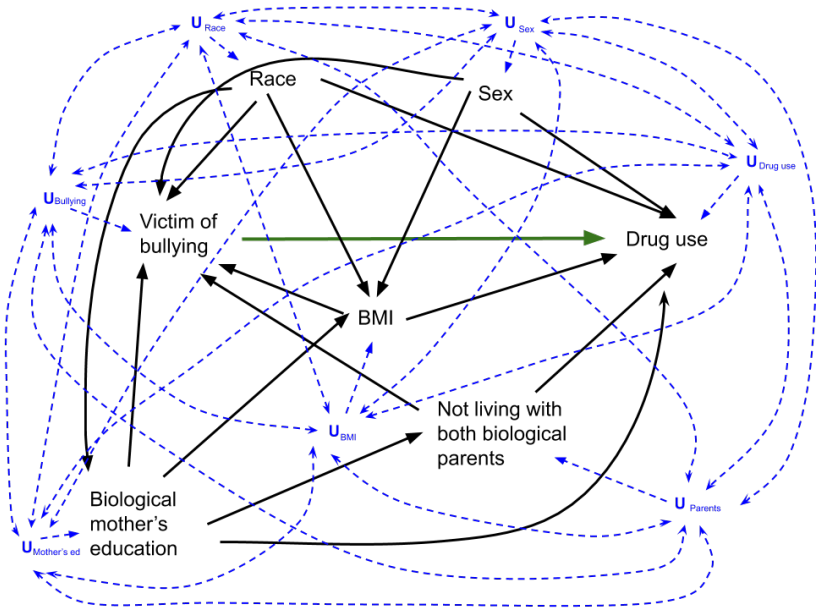
Link to our SCM

These data are part of a nationally representative sample; therefore, they have survey weights associated with them and are not truly generated from idependent, identically distributed draws from the random variable O. However, for this project we have used a simple link. Thus, we assume that the observed data $O = (W, A, Y) \sim \mathbb{P}_0$ were generated by sampling 7703 i.i.d. times from a data generating process described by the SCM.

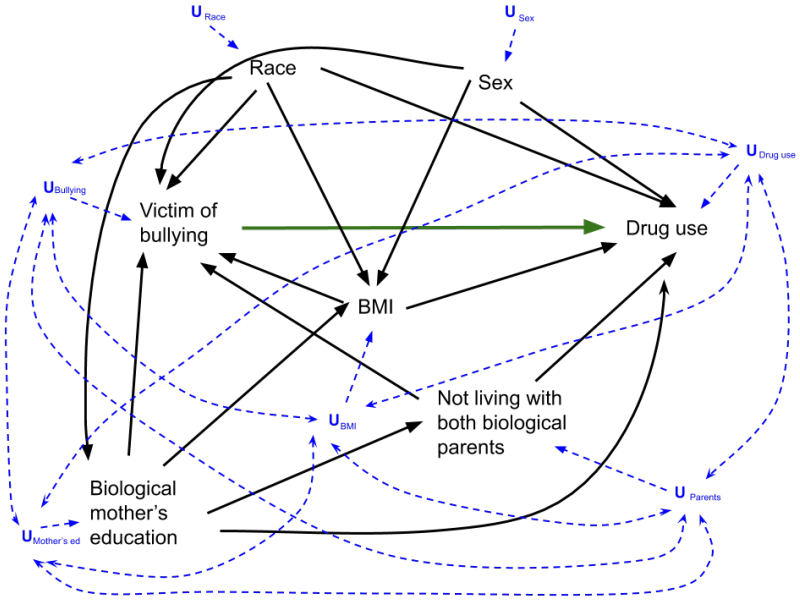
Given our knowledge of our data, we have chosen a statistical model $\mathcal{M}$ for the set of allowed distributions for the observed data that is non-parametric.

Step 4. Identify

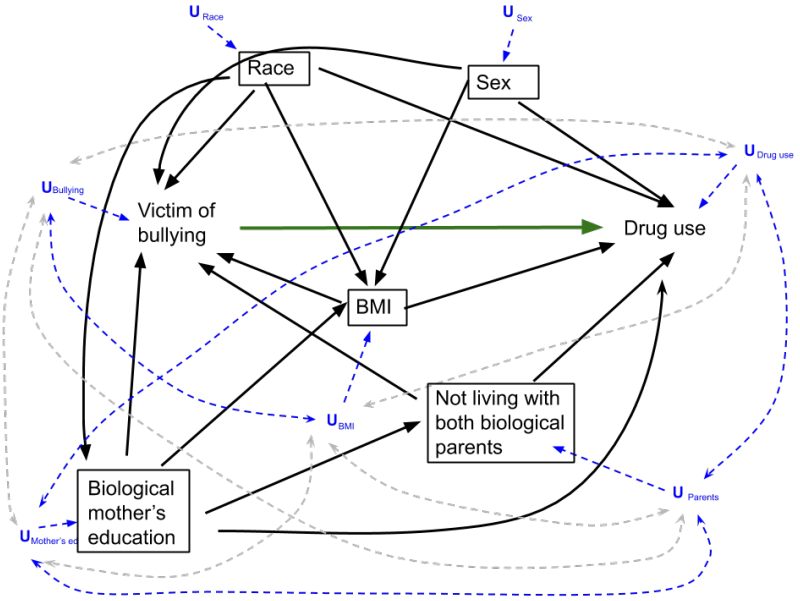
If there were no independence assumptions, this would be our DAG:



In reality, we believe there are no shared unknowns between race or sex with any other variables; therefore, this is what we think is most true for our DAG:

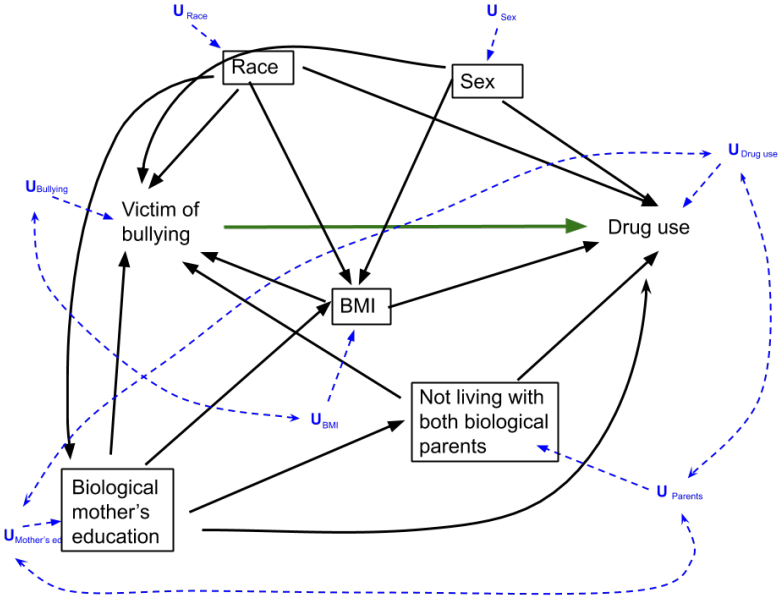


However, under this initial causal model the target parameter cannot be identified, because there are several backdoor pathways. To block the backdoor pathways and create an identified target parameter, we would need to control for all of the endogenous covariates in our model, and would also need to make a series of independence assumptions for convenience (with the previously assumed shared unknowns shown below in grey). We think the shared unknowns represented by the grey dotted lines are less plausible than in the remaining cases of shared unknowns (remaining in blue), allowing us to make progress on the causal roadmap.



We could improve the plausibility of these additional independence assumptions by identifying potentially shared unknowns and gathering data on them so we could control for them in the model. For example, income and access to fresh, healthy food are both examples of potential shared confounders of biological mother's education and BMI. If we measured those variables and controlled for them in our model, it would be more plausible to assume independence between the unknowns contributing to both of those nodes.

Once making these convenience-based independence assumptions, we have d-separation and can proceed with our analysis using an identifiable target parameter, using the final DAG below.



Step 5. Commit to an Estimand and Statistical Model

The target parameter of the observed data distribution (which equals the causal parameter in the augmented causal model $\mathcal{M}^{F\star}$) is the G-Computation formula, which is our statistical estimand:

$$\psi(\mathbb{P}_0) = \mathbb{E}_0[\mathbb{E}_0(Y|A=1,W) - \mathbb{E}_0(Y|A=0,W)] =$$
$$\sum_{w1,w2,w3,w4,w5} [\bar{Q}_0(1, W1=w1, W2=w2, W3=w3, W4=w4, W5=w5) -$$
$$\bar{Q}_0(0, W1=w1, W2=w2, W3=w3, W4=w4, W5=w5)] *$$
$$\mathbb{P}_0(W1=w1, W2=w2, W3=w3, W4=w4, W5=w5)$$

As we said previously, given our knowledge of our data, the statistical model $\mathcal{M}$ for the set of allowed distributions for the observed data is non-parametric.

Step 6. Estimate

Estimators Used

We estimated the ATE (under causal assumptions) using three estimators: simple substitution (G-computation), inverse probability of treatment weighting (IPTW) with stabilized weights, and targeted maximum likelihood estimation (TMLE). We also calculated the unadjusted ATE (i.e., the mean difference in Y between the exposed and unexposed). The formula for each estimator is below.

***Simple Substitution*:**

$$\hat{\Psi}_{SS}(P_n) = 1/n \sum_{i=1}^{n} (\hat{\overline{Q}}(1, W_i) - \hat{\overline{Q}}(0, W_i))$$

- P_n is the empirical distribution and $\hat{\overline{Q}}(A, W)$ is the estimate of the emprical mean of Y, given A and W (a vector of baseline covariates).

***IPTW with stabilized weights*:**

$$\hat{\Psi}_{St.IPTW}(P_n) = \frac{1/n \sum_{i=1}^{n} \frac{I(A_i=1)}{\hat{g}(A_i|W_i)} Y_i}{1/n \sum_{i=1}^{n} \frac{I(A_i=1)}{\hat{g}(A_i|W_i)}} - \frac{1/n \sum_{i=1}^{n} \frac{I(A_i=0)}{\hat{g}(A_i|W_i)} Y_i}{1/n \sum_{i=1}^{n} \frac{I(A_i=0)}{\hat{g}(A_i|W_i)}}$$

- $\hat{g}(1|W_i)$ is an estimate of the propensity score.

***TMLE*:**

$$\hat{\Psi}_{TMLE}(P_n) = 1/n \sum_{i=1}^{n} (\overline{Q_n^*}(1, Wi) - \overline{Q_n^*}(0, Wi))$$

- $\overline{Q_n^*}(A, W)$ is the targeted estimate of the conditional mean outcome, given A and W.

We used SuperLearner for prediction with all estimators. Drawing from a pre-specified library of candidate algorithms, SuperLearner uses V-fold cross-validation to create a convex combination of algorithms that minimizes a loss function (non-negative least squares, by default). Because we only had 319 people in our data who were bullied and who later use drugs, we set the number of folds to 5 (keeping in mind that CV.SuperLearner would break each fold into an additional 5 folds).

Consistent with our nonparametric model assumptions, we included an array of candidate algorithms in the SuperLearner library, including parametric and nonparametric appraoches. Specifically, we included glm, glm.interaction, glmnet, bayesglm, randomForest, step, and mean in the library. All of these are appropriate for binary outcomes. GLM fits a generalized linear model including all of the main terms in the model. Glm.interaction adds to this second-order polynomials and main-term interactions. Glmnet implements a peanized likelihood model with LASSO or elastic net regularization. Bayesglm uses a Bayesian approach to fit a glm, rather than a frequentist (maximum likelihood) appraoch. RandomForest is a machine learning algorithm using decision trees. Step does forward and backward model selection using AIC. Lastly, mean takes the mean of Y, which we included for reference.

Our team also wrote our own wrapper for performing a log-linear model. However, due to the computational resources needed to bootstrap confidence intervals, we determined that we needed to do so in a paralellized fashion. Our self-made wrapper produced unusual errors when paralellized resulting in a failure to bootstrap and ultimately the wrapper was dropped. Regardless, it generally received a low weight in the superlearner algorithms and therefore removing this wrapper likely did not adversely affect our analysis.

Assessing Positivity Assumptions

Given that young children can sometimes be cruel to their peers, there are not theoretical positivity violations expected; children with all combinations of the covariates could be bullied. For assessment of practical positivity violations, we initially tabulated exposure and outcome across all possible levels of our categorical variables. After doing this initial tabulation, we noted 3 variables that contributed to positivity violations.

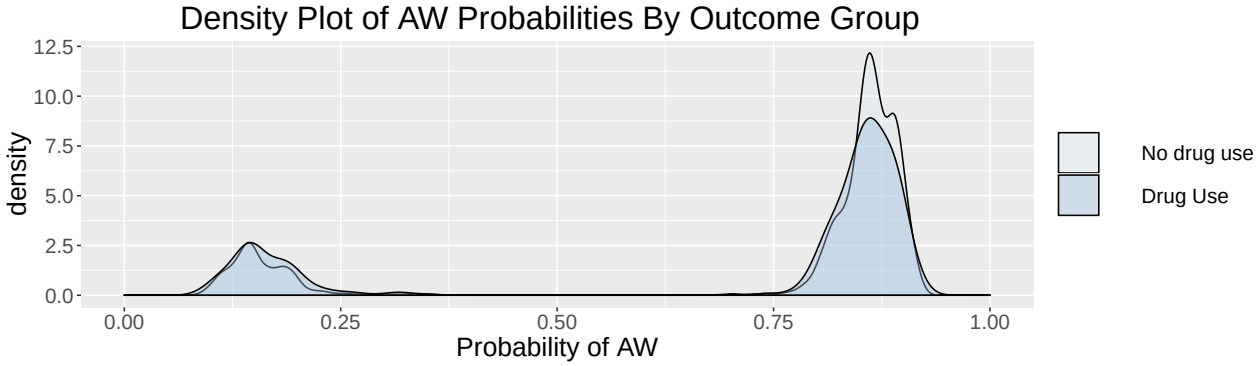
In the race variable, fewer than 1% of the observations were in the 'mixed race' group (n= 69), leading to multiple practical positivity violations. The decision was made to remove this subgroup from our analyses, such that the race variable is still part of the model, but the mixed race individuals are not considered in the analysis. This does limit the generalizability of the findings (to non-mixed race individuals), but as it allows the race variable to continue to be used in the model this helps ensure the validity of the estimate within the groups for which there is sufficient data. As has been described,[5] all solutions to practical positivity violations require a trade-off between improving the positivity issue and potentially introducing bias into the target of causal inference. This solution was chosen as the least likely to bias the inference.

Whether or not a participant had any same-sex romantic partners was also initially considered as a potential covariate in the model. There were very few participants with one or more same sex partners (n=269 or <4% of our sample). Moreover, when two by two tables relating bullying and drug use were stratified by whether or not a participant had a same sex partner, there was no significant difference in the measure of association between the strata (test of homogeneity χ^2 value 1.49, p=0.223). This suggests that the covariate set could likely be restricted to exclude this variable with minimal effect on our target of inference, and thus this variable was not included in the analyses.

Third, we initially hoped to include citizenship status. However, we realized that this measure was actually assessing place of birth, not citizenship, and therefore was likely not estimating the construct that we felt was important. Moreover, there were very few participants who were born outside the United States, and many with no data on their birthplace, so this could have also created positivity violations.

For the final variable set, observations exist in every possible category of our variable set. For our only continuous variable, BMI z-score (for age and sex), we looked at the distribution of BMI z-scores in the two exposure categories. These were very similar across the entire distribution suggesting good support for our analyses in the data (see below).

We also observed the distribution of propensity scores of each covariate-exposure combination and created density plots of this, differentiating between those with each outcome. These show a bimodal distribution (see below; as expected given that many of the covariates are bivariate). Importantly, these distributions have roughly equal support between outcome groups.



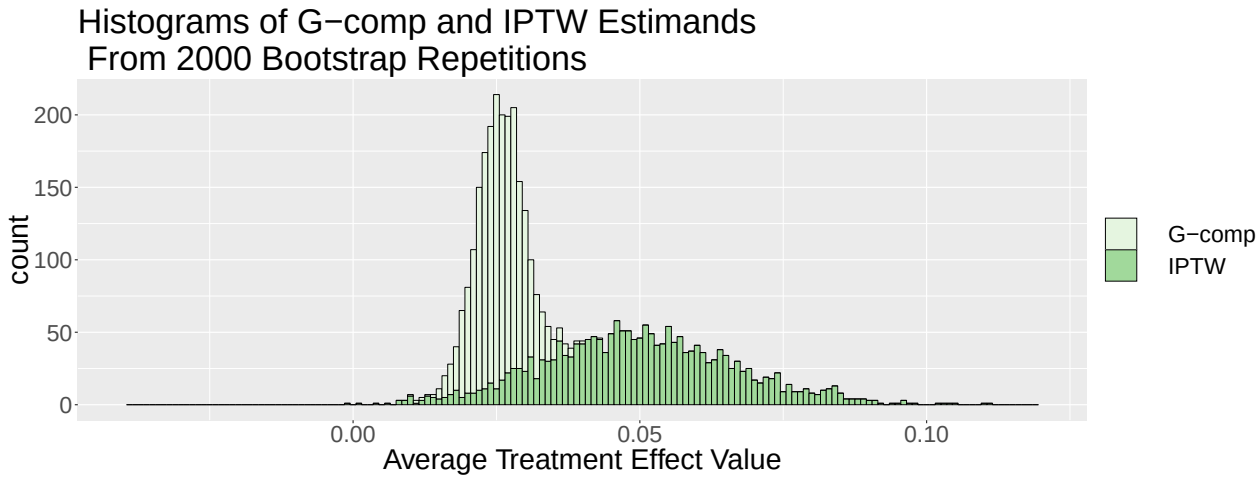
In summary, in our final set of variables, we are confident that an adequate volume of data exists across all levels of the covariate and exposure combinations. We chose not to look at truncation of the weights as we had good overlap between weights and chose to avoid introducing bias in this manner. Thus, we may proceed with estimation without concerns over practical positivity violations.

Confidence Intervals

As non-parametric modeling techniques were used for estimation, standard parametric confidence intervals would be inappropriate. For the TMLE estimator, we used the robust method built in to the ltmle package to obtain a confidence interval. This package implements two separate methods: an influence curve-based method as well as a robust method which uses TMLE to estimate the variance. The method which produced the more conservative interval is then used.

For the G-computation (SS) and IPTW estimators we performed a non-parametric bootstrap. As Super-Learner was used to arrive at both estimates, there is no guarantee that the data are asymptotically linear so these intervals may not be reliable. These are shown in the histogram below. This theoretical concern regarding the bootstrap for these estimators may explain why the point estimate from the simple substitution model is outside the calculated confidence interval, and why these two bootstrapped approximations of the sampling distributions have substantively different means and standard errors. We chose to proceed with bootstrapping as it was the tool we have learned in this class, but would likely consider another approach in the future.

The decision to use the package's built in method for assessing the confidence interval for TMLE was based on confidence that this would also provide a robust method, as well as time constraints making it difficult to run bootstraps on 3 separate SuperLearner runs (those for G-computation, IPTW and TMLE).



Results

The unadjusted ATE (i.e., the mean difference in Y between the exposed and unexposed) was 0.05. The adjusted estimates, which marginalize over the distribution of the covariates, are shown in the table below. Under the causal assumptions of randomization and positivity, these estimate the ATE of childhood bullying on incident drug use.

Table 2: Average Treatment Effect estimates

Estimator	ATE (95% CI)
G-computation	0.039 (0.017, 0.034)
Stabilized IPTW	0.045 (0.018, 0.084)
TMLE	0.044 (0.007, 0.08)

The point estimate for the ATE from the G-computation model is outside the range of the bootstrapped 95% CI. This could be due to potential overfitting in the predictive model (as is shown in the table below, randomForest was given a large weight in the convex combination and is known to cause these kinds of problems). It is also illustative of the fact that bootstrapping CIs for G-comp is not grounded in statistical theory.

Table 3: SuperLearner predictive model details from TMLE

Algorithm	A Risk	A Coefficient	Y Risk	Y Coefficient
glm	0.15497	0	0.14089	0.463
glm.interaction	0.15505	0.209	0.1415	0
glmnet	0.15498	0	0.14091	0
bayesglm	0.15497	0	0.14089	0
randomForest	0.19034	0.461	0.17071	0.224
step	0.15495	0.268	0.1409	0.248
mean	0.15669	0.063	0.14288	0.065

In this case, SuperLearner did not have the lowest risk of all the possible algorithms we considered in our library. However, the risks were very similar to one another across the board. This might suggest that our library could be larger and more diverse. We would like to include more machine learning algorithms in the future, but need to learn more about what options are available and appropriate for our data.

Table 4: SuperLearner performance

Algorithm	Avg Risk	SE
SuperLearner	0.14121	0.00287
Discrete SL	0.14072	0.00276
glm	0.14071	0.00276
glm.interaction	0.14128	0.00277
glmnet	0.14071	0.00276
bayesglm	0.14071	0.00276
randomForest	0.16787	0.00417
step	0.14072	0.00276
mean	0.14288	0.00282

Step 7. Interpret Results

According to our analysis, the difference between the average counterfactual risk of drug use if everyone was bullied versus if no one was bullied is 0.04; this is the causal interpretation. We do not think our results should be interpreted causally, however, becuase of the unrealistic convenience assumptions this interpretaion requires. The statistical interpretation of our findings is that being bulled before at 12 is associated with an increase of 0.04 in the probability of subsequent drug use, compared to not being bullied. This is a plausible finding, though we expected it to be even higher. It makes sense that we would see this in our data, however, because drug use is typically underreported due to social desirability bias, biasing our risk difference toward the null.

Our G-computation estimator had the lowest ATE estimate but greatest precision of our three estimators at 0.039 (95% CI: 0.017 - 0.034). Our stabilized IPTW estimator had the highest ATE estimate and moderate precision, at 0.045 (95% CI: 0.018 - 0.084). Our TMLE estimator produced an estimate of 0.044 (95% CI 0.007 - 0.08), with lowest precision. Ultimately, these estimators appear to have performed similarly, with much overlap between the point estimate and confidence intervals of each estimator's ATE. Moreover, the G-comp, IPTW, and TMLE estimates are all very close to the unadjusted estimate, suggesting that there was not a lot of confounding by the Ws we included in the model.

Limitations

Our analysis includes a number of limitations. First, there were a number of covariates that were exogenous, which we would have measured and included had we been prospectively collecting data instead of using an existing dataset. As one important example of this, we did not have any information on parent drug use, which we would have wanted to control for as an important confounder in our model (for identifiability we also needed to assume no shared unknowns between bullying and drug use, where this would obviously have also fit). It is likely incorporated into our model as one of the shared unknowns of not living with both biological parents, and our outcome (drug use).

Second, while the data were collected, we did not include any mental health variables as endogenous variables in our model, because the questions were extremely vague, and the temporality was unclear (i.e., we couldn't determine that mental health wasn't caused by exposure). Mental health is likely an important covariate to include in this analysis, potentially as a confounder or a mediator of the exposure and outcome.

Third, the independence assumptions that we created out of necessity to identify the target causal parameter were likely not accurate, introducing confounder-based bias into our results.

And finally, this was not a random sample, yet our link was simplified for this project, as described in step 4, above. In subsequent analyses we would plan to use the sample weights such that these analyses would be reflective of the US population, as intended when the study was designed.

Impacts

The results of this analysis allow policymakers and school administrators to better identify youth who are at risk for starting to use drugs as a result of bullying, and provide them with additional services and social supports. It also supports the use of anti-bullying interventions in schools, as prevention of incident substance use later in life will also lead to a reduction in many of the adverse health outcomes associated with drug use.

Contributions of the Team Members

- Suggestion of a dataset and potential issues for exploration: Veronica
- Project management: Shelley
- Background and literature review: Lizzy
- Delineation of the causal model, causal question and estimand choice: Entire group
- Development of the structural equations: Lizzy
- Identifiability considerations: Shelley
- Data cleaning and multiple imputation of missing covariate data: Veronica
- Practical Positivity Checks: Stephanie
- Creation of SuperLearner library: Veronica
- Coding of ATE point estimates: Veronica
- Bootstrapping of Confidence Intervals: Stephanie
- Interpretation of Results: Entire Group

References

1. Ttofi MM, Farrington DP, Losel F, Crago RV, Theodorakis N. (2016) School bullying and drug use later in life: A meta-analytic investigation. School Psychology Quarterly. 31(1): 8-27.
2. Radliff KM, Wheaton JE, Robinson K, Morris J. (2012) Illuminating the relationship between bullying and substance use among middle and high school youth. Addictive Behaviors. 37(4): 569-572.
3. Tharp-Taylor S, Haviland A, D'Amico EJ. (2009) Victimization from mental and physical bullying and substance use in early adolescence. Addictive Behaviors. 34(6-7): 561-567.
4. Sarlin E. (2017) Substance Use Disorders Are Associated With Major Medical Illnesses and Mortality Risk in a Large Integrated Health Care System. Bethesda, MD: National Institute on Drug Abuse, October 24.
5. Petersen M, Porter KE, Gruber S, Wang Y, van der Laan MJ. (2012) Diagnosing and responding to violations in the positivity assumption. Stats Methods Med Res. 21(1):31-54.