Final Project: Childhood Bullying and Subsequent Drug Use

Shelley Facente, Steph Holm, Lizzy Kinnard, Veronica Pear

Spring 2019

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

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]

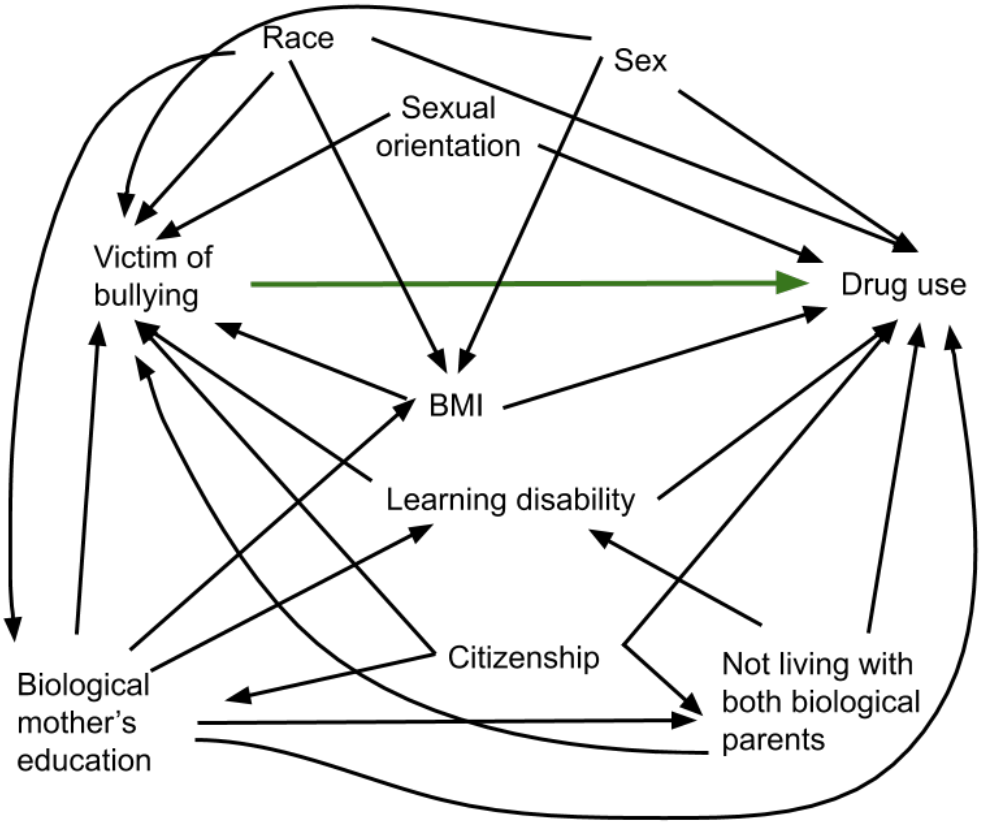
To our knowledge, no studies have evaluated the relationship between bullying victimization and drug use using causal inference approaches. This study fills a gap in the literature by studying this question in a causal framework.

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 age 12-16 (initial n=9000) in 1997. These youth have since been followed longitudinally. The target population is youth in the United States. Our particular analyses have 7703 subjects in the final dataset.

Original DAG

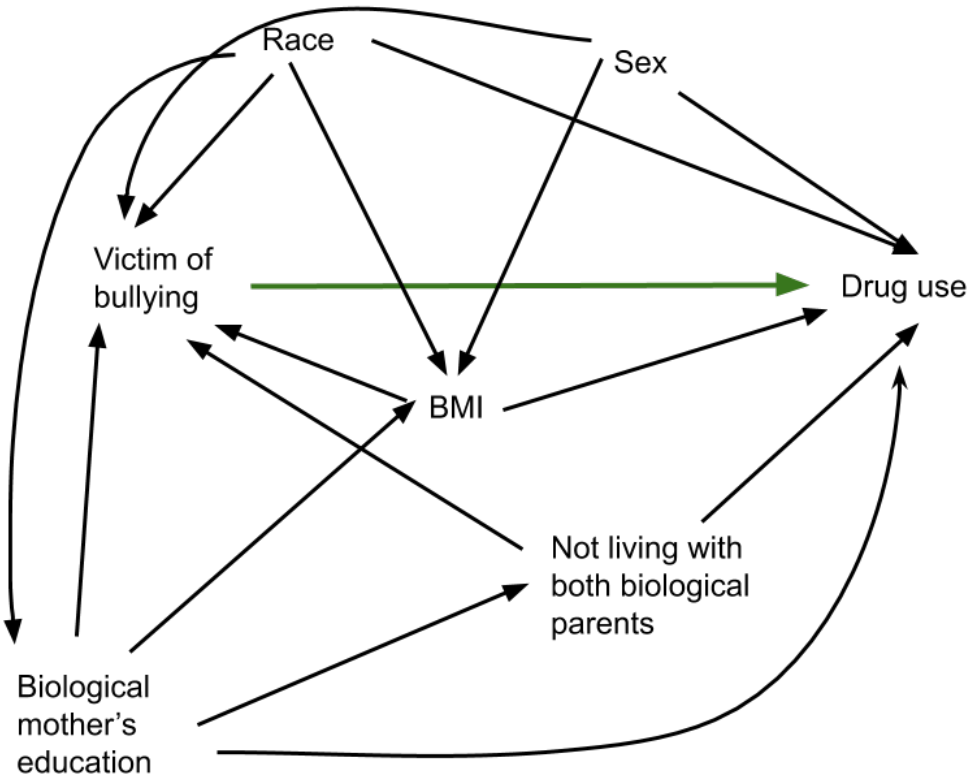
This is the original DAG we developed while looking at what was available in the dataset:



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. Sexual orientation wasn't associated with our exposure/outcome, and strata were homogenous by 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. Finally, learning disability actually captured a learning or emotional problem, and we decided that the latter could be influenced by our exposure, being bullied, and thus excluded it.

Therefore, below is our final DAG for this analysis:

Final DAG



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, A is victim of bullying, and Y is incident drug use.

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.

Structural Equations

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

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

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); Y = incident drug use ("cocaine or other hard drugs") after 1997.

Target Causal Parameter

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:

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

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.

Our Observed Data

A: Bullying before the age of 12 (asked in 1997)

Y: Incident drug use ("cocaine or other hard drugs") after 1997

Ws: Race/ethnicity, sex, BMI, not living with both biological parents, mother's educational status (all Ws were measured at baseline)

Sample size: 7,703

The target population is youth in the United States.

Link to the SCM

We assume that the observed data O = (W, A, Y) $\sim \mathbb{P}_0$ were generated by sampling n times from a data generating process described by the SCM. The statistical model $\mathcal{M}$ for the set of allowed distributions for the observed data is non-parametric.

Table 1

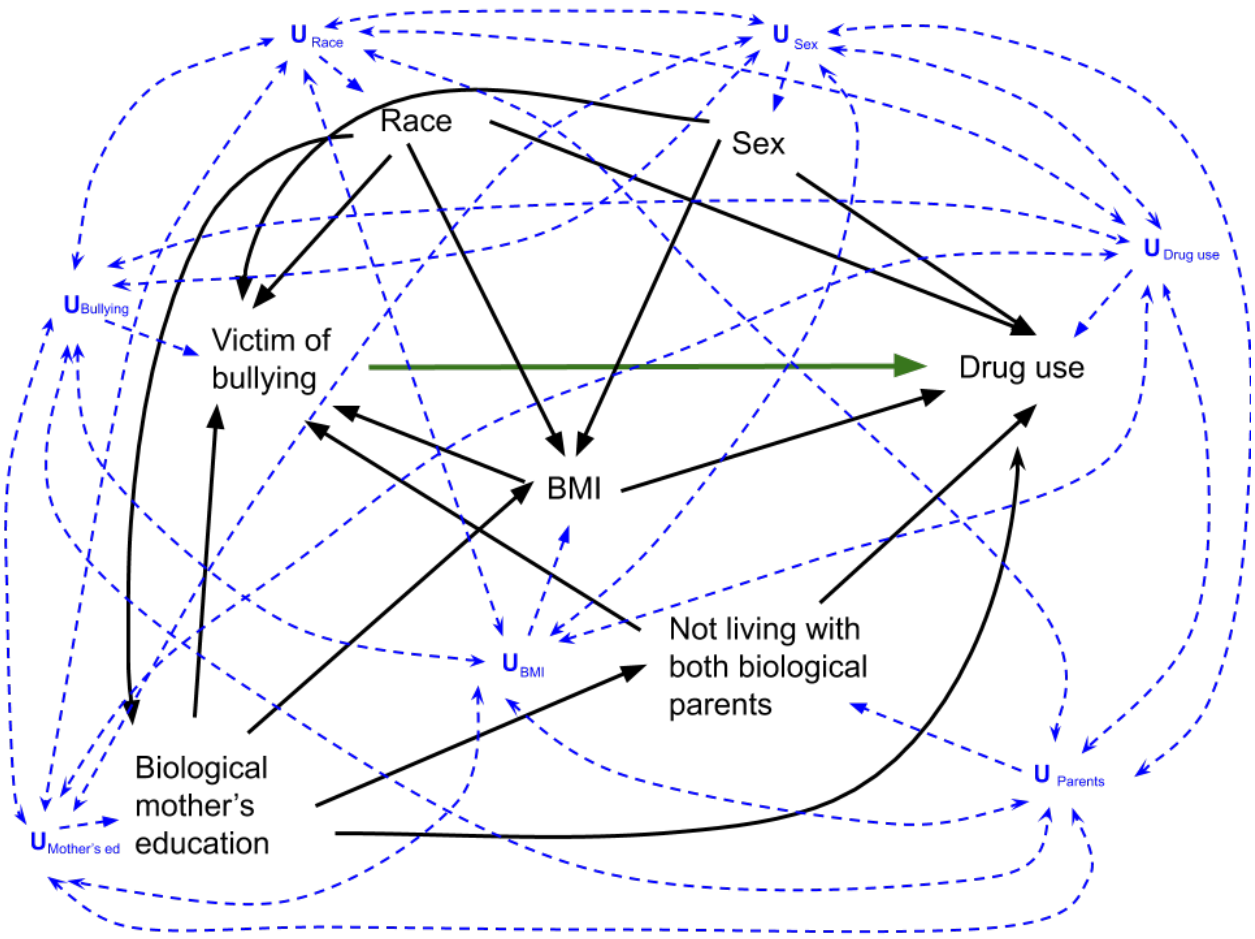
Covariate	Drug use (%)	No drug use (%)
Drug use (Total)	1330 (17.3%)	6373 (82.7%)
Victim of bullying		
Yes	319 (4.1%)	1179 (15.3%)
No	1011 (13.1%)	5194 (67.4%)
Mother's education		
High school or less	732 (9.5%)	3867 (50.2%)
Some college or more	598 (7.8%)	2506 (32.5%)
Sex		
Female	591 (7.7%)	3218 (41.8%)
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%)
No	685 (8.9%)	3197 (41.5%)
BMI z-score	0.513 (*mean*) 1.03 (*sd*)	0.505 (*mean*) 0.98 (*sd*)

Marginal distribution of exposure and outcome:

Variable	No	Yes
Bullied < 12	6205	1498
Incident drug use	6373	1330

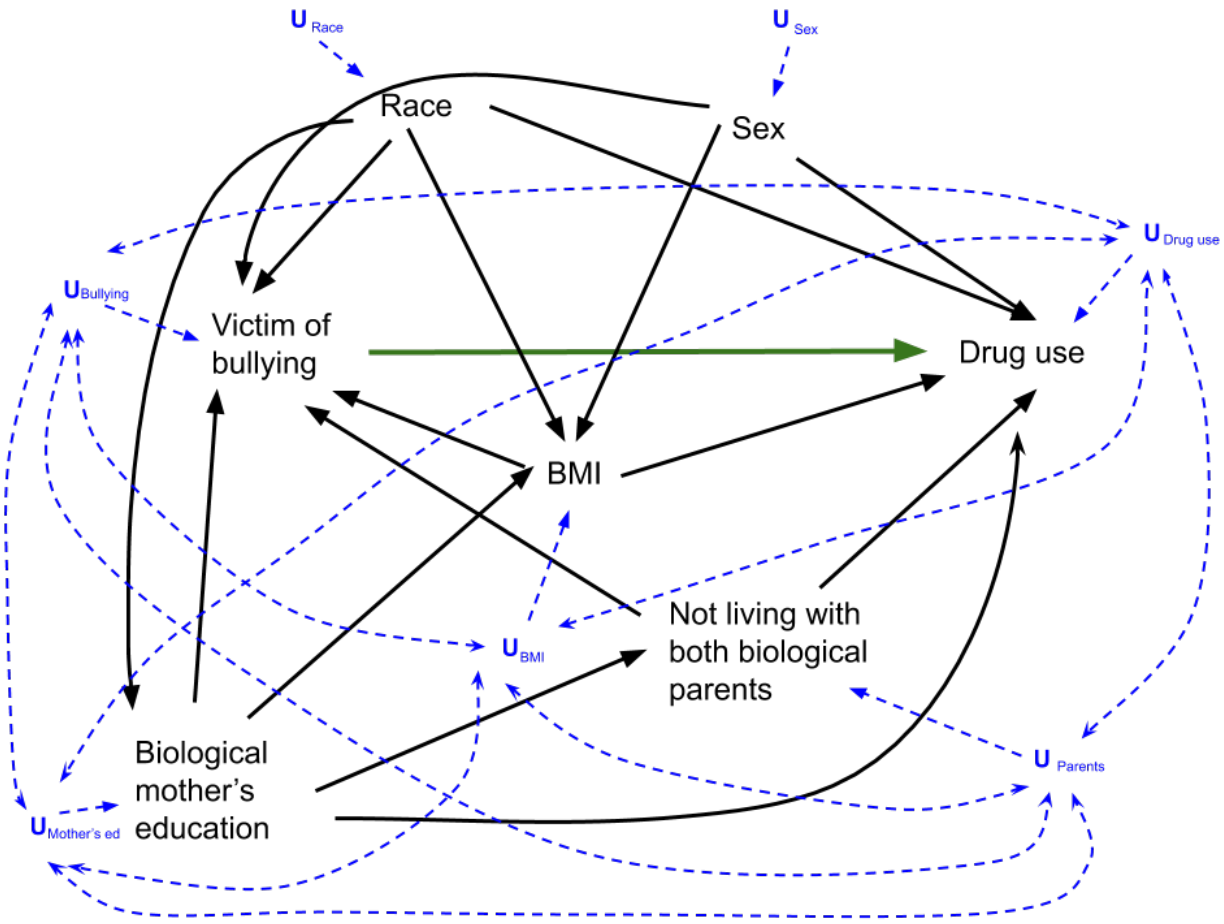
Identifiability

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



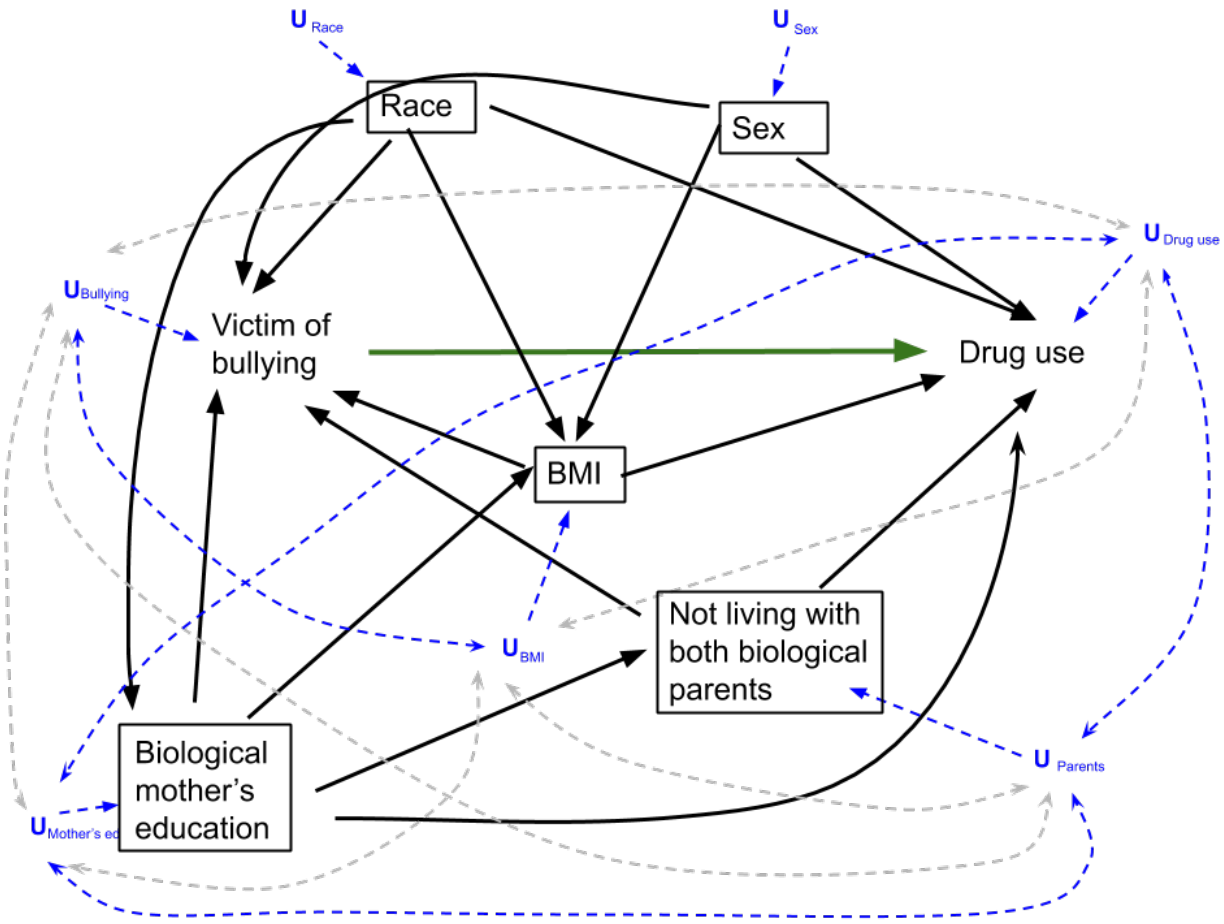
Identifiability

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:



Identifiability

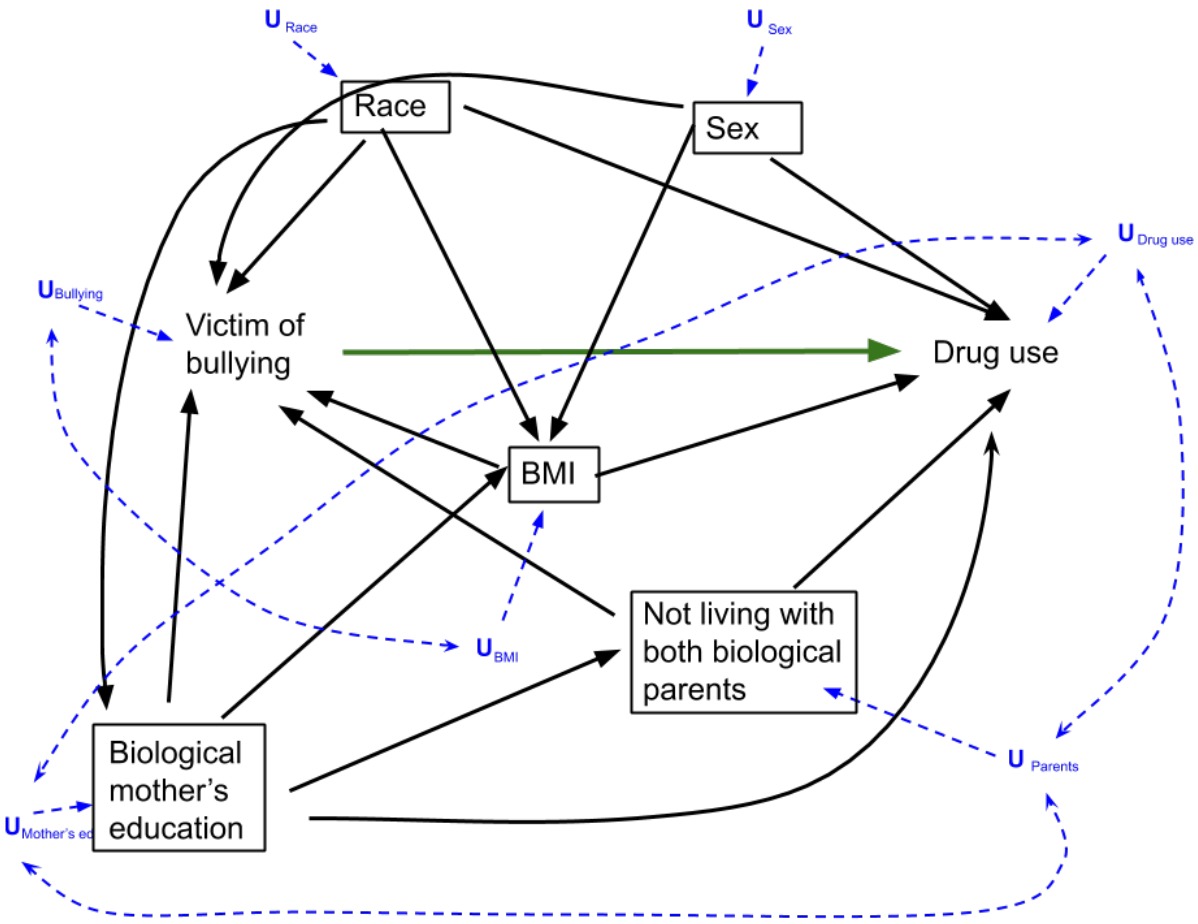
However, under this initial causal model the target parameter cannot be identified, because there are a large number of backdoor pathways. To block the backdoor pathways and create an identified target parameter, we would need to control for all of the endogenous covariates in our model, and would need to also add 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), and allow us to make progress on the causal roadmap.



Identifiability

We could improve the plausibility of these additional independence assumptions by identifying potentially shared unknowns and gathering data on them so they can be controlled for in the model. For example, income; social status; and access to fresh, healthy food are all 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 these independence assumptions have been agreed to for convenience, we have d-separation and can proceed with our analysis using an identifiable target parameter, using the final DAG below.



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

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

the G-Computation formula:

This is our statistical estimand.

Positivity

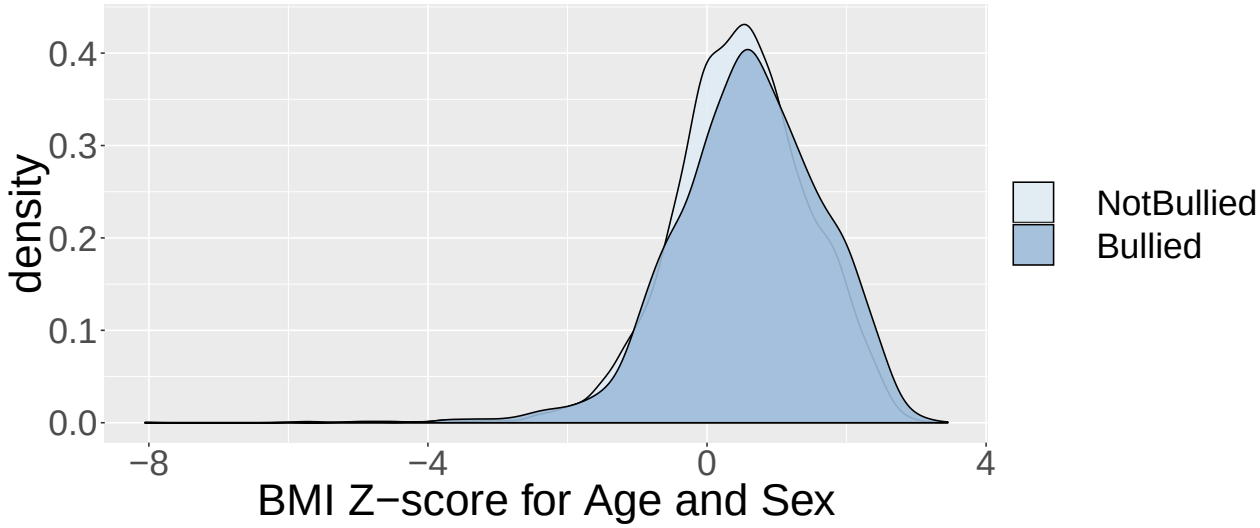
Given that young children can sometimes be cruel to their peers, there are not theoretical positivity violations expected; children with all combinations of the covariates might be bullied. For assessment of practical positivity violations, we initially tabulated exposure and outcome across all possible 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.

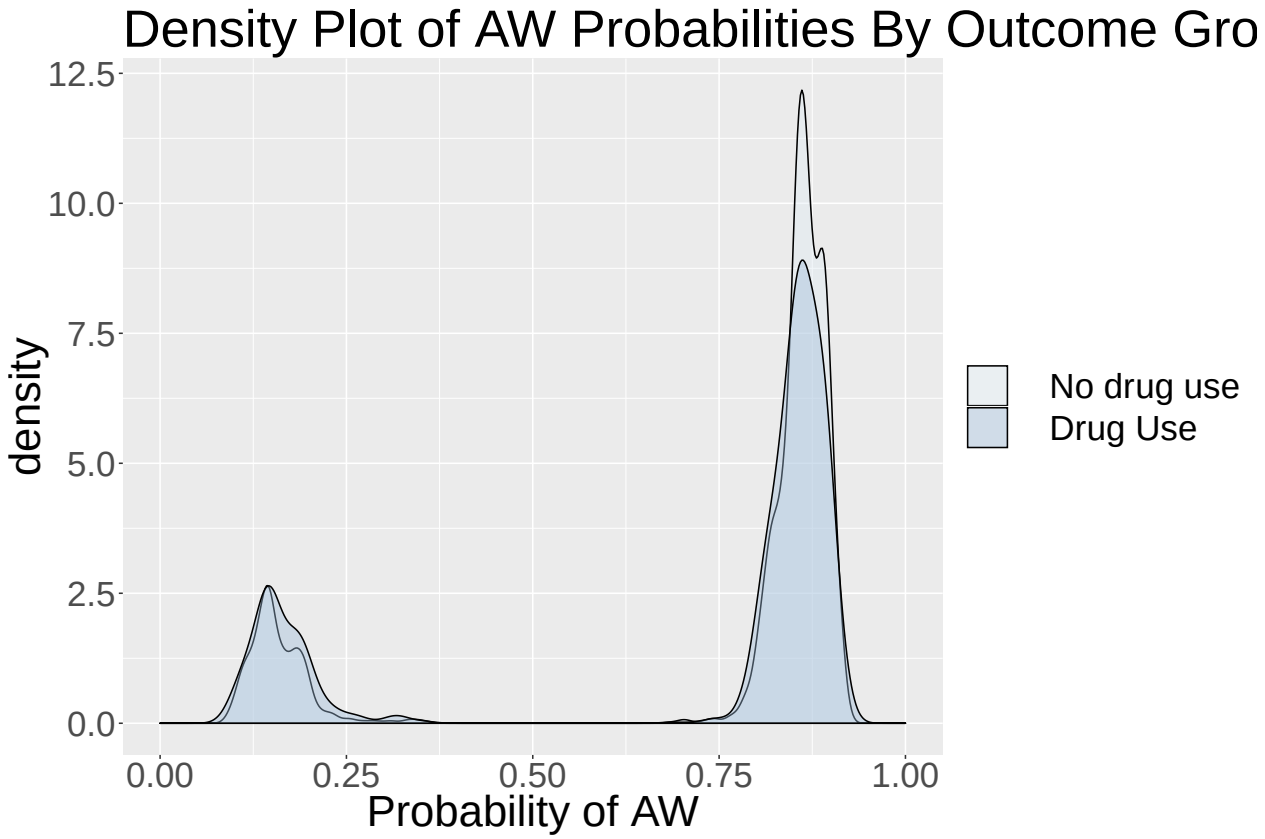
Whether or not a participant had any same-sex romantic partners was also initially considered as a potential covariate in the model. There were very few participants with one or more same sex partners (n=269 or <4% of our sample). Moreover, when two by two tables relating bullying and drug use were stratified 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 hace 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 differentiationg 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 with 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. Thus, we may proceed with estimation without concerns over practical positivity violations.

Estimation

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 used SuperLearner for prediction with all estimators. SuperLearner uses v-fold cross-validation to create a convex combination of algorithms, drawing from a pre-specified library, that minimizes a loss function (non-negative least squares, by default).

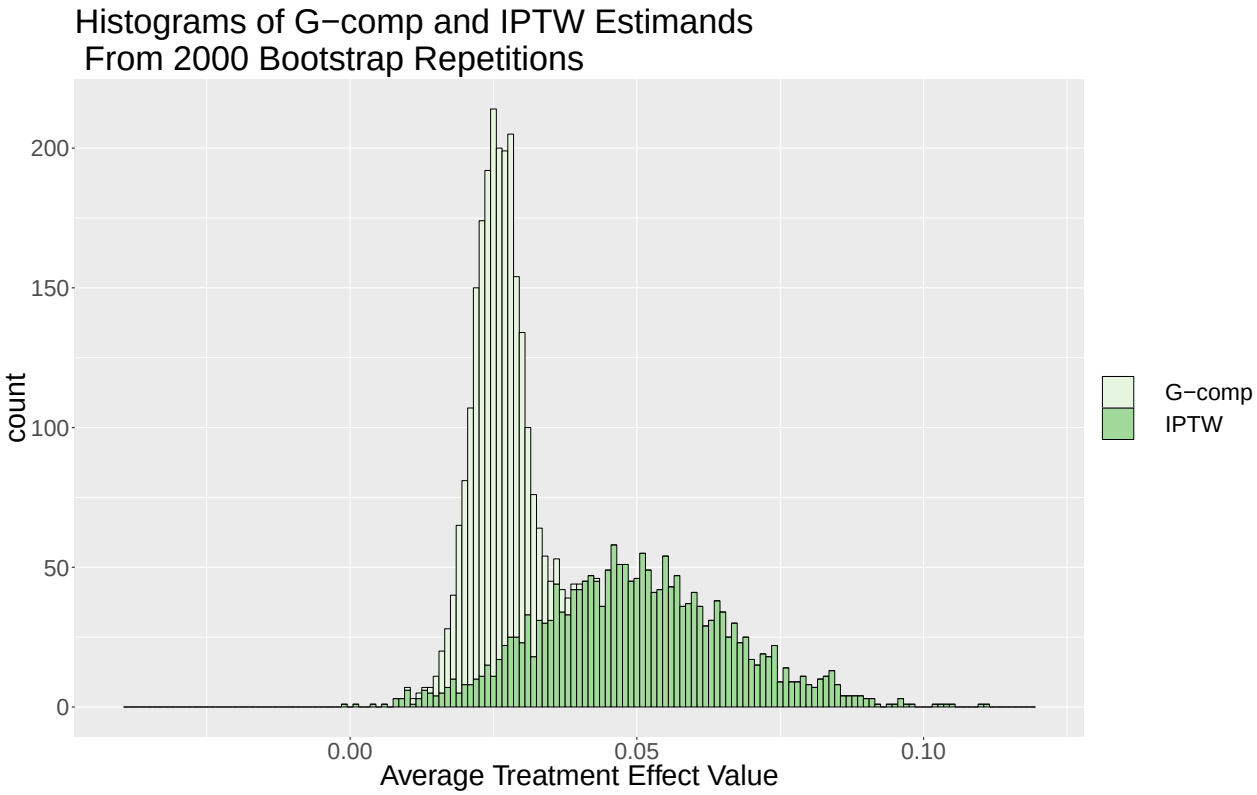
- We use SuperLearner for prediction in all models.
- Library: SL.glm, SL.glm.interaction, SL.glmnet, SL.bayesglm, SL.randomForest, SL.step, SL.mean
- 5-fold cross-validation
- We use the following estimators: G-computation (simple substituion estimator), stabilized IPTW, and TMLE

Confidence Intervals

As non-parametric modeling techniques were used for estimation, standard parametric confidence intervals would be inappropriate. For the TMLE estimatnd, we used the robust method built in to the ltmle package to obtain a conference interval. This package implements two separate methods: an influence cure 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 (simple substitution) and IPTW estimands we performed a non-parametric bootstrap. As superlearner 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 estimatnds may explain why the point estimate for the simple substitution estimand is outside the calculated confidence interval, and why these two bootstrapped approximations of the sampling distributions have different mean values. 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).



Estimation: Results

- The unadjusted ATE = mean(Y|A=1 - Y|A=0) = 0.05

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)

Estimation: SuperLearner convex combinations

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

Estimation: SuperLearner performance

CV.SuperLearner

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

Algorithm	Avg Risk	SE
mean	0.14288	0.00282

Interpretation

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, causally interpreted to mean that if people are bullied they are about 4% more likely to use drugs later in life than if they are not 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.

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. And third, the independence assumptions that we created out of necessity identify the target causal parameter were likely not correct, introducing confounder-based biased into our results.

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 use of drugs.–>

According to our analysis:

- **Statistical interpretation**: being bullied before the age of 12 is associated with a 4% absolute increase in drug use compared with not being bullied
- **Causal interpretation**: the difference between the average counterfactual risk of drug use if everyone was bullied versus if no one was bullied is 0.04

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)

Limitations

1. Important exogenous variables
    - Parent drug use
    - Mental health
    - And more...
2. Necessary independence assumptions

Impacts

- Identify youth who are at risk for starting to use drugs as a result of bullying
- Supports use of anti-bullying interventions in schools

Contributions of the Team Members

- Suggestion of a dataset and potential issues for exploration: Veronica
- Particular expertise that we each contributed:
    1. Shelley - Project management
    2. Stephanie - Pediatrics
    3. Veronica - Social and Substance Use Epi
    4. Lizzy - Social and Substance Use Epi

- Establishment of the causal model, delineation of the causal question and estimand of choice: Entire group
- Identifiability considerations: Entire group, with Lizzy and Shelley working on the DAG
- Creation of slides for causal question, SCM, background on our dataset: Lizzy and Shelley
- Creation of SuperLearner library: Veronica
- Coding of ATE point estimates: Veronica
- Coding of Practical Positivity Checks and 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