Final Project: Childhood Bullying and Subsequent Drug Use

Shelley Facente, Steph Holm, Lizzy Kinnard, Veronica Pear

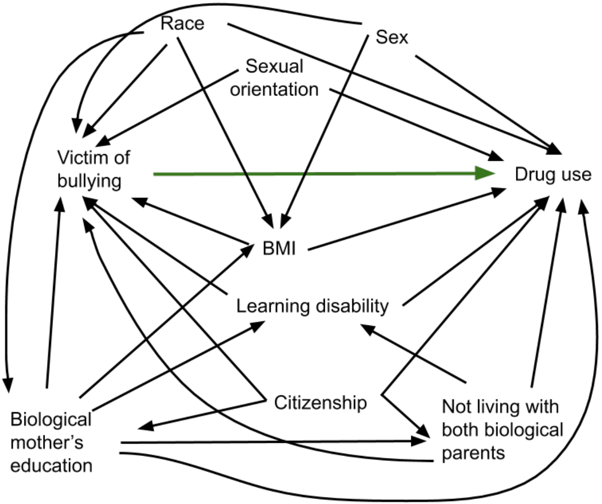
Spring 2019

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

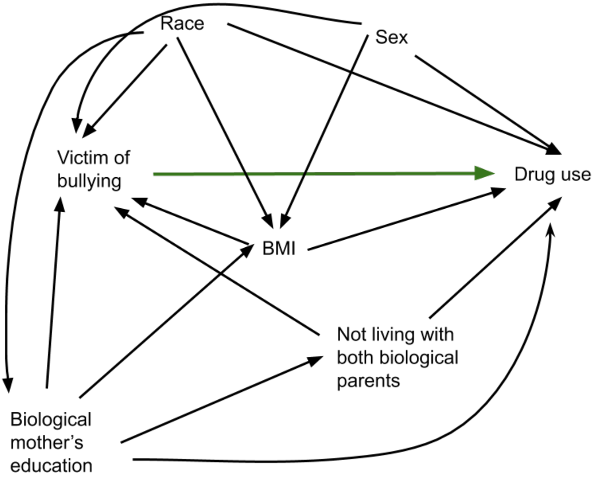
Specify a Causal Model

Our dataset: - National Longitudinal Survey of Youth - Nationally representative cohort of youth age 12-16 - Initial recruitment n = 9000 in 1997

Original DAG



Final DAG


Race
Sex
Victim of bullying
Drug use
BMI
Not living with both biological parents
Biological mother's education

Structural Equations

LIZZY TO CLEAN THIS UP Endogenous nodes: X = (W, A, Y), where W = (W1, W2, W3, W4, W5) is the set of baseline covariates, A is victim of bullying, and Y is drug use.

Background variables (Exogenous nodes): U = (UW, UA, UY) ∼ PU. We place no assumptions on the distribution PU. We have not placed any restrictions on the functional forms.

Structural equations F:

W1 = fW1 (UW1, W3) W2 = fW2 (UW2) W3 = fW3 (UW3) W4 = fW4 (UW4, W1) W5 = fW5 (UW5, W1, W2, W3) A = fA (UA, W1, W2, W3, W4, W5) Y = fY (UY, A, W1, W2, W3, W4, W5)

W1 = Mother's education; W2 = Sex; W3 = Race/ethnicity; W4 = Living with both biological parents; W5 = BMI

Target Causal Parameter:

- 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.

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, age, 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.

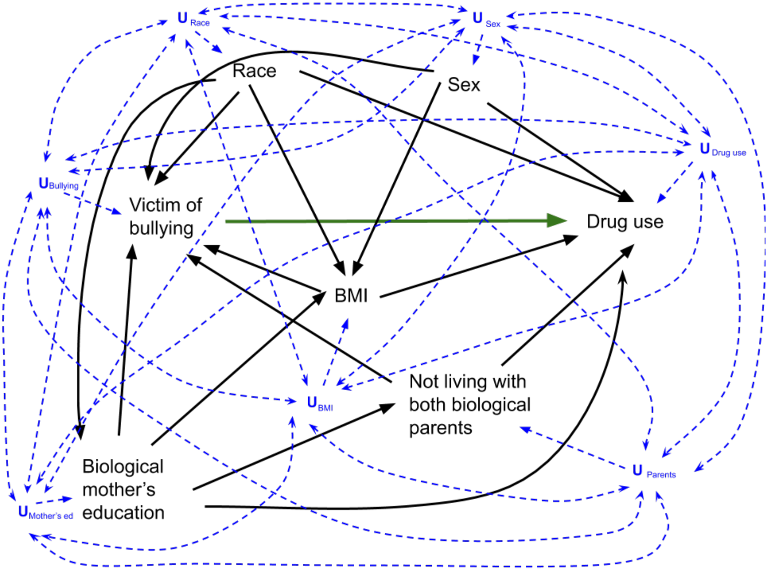
Link to the SCM

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.

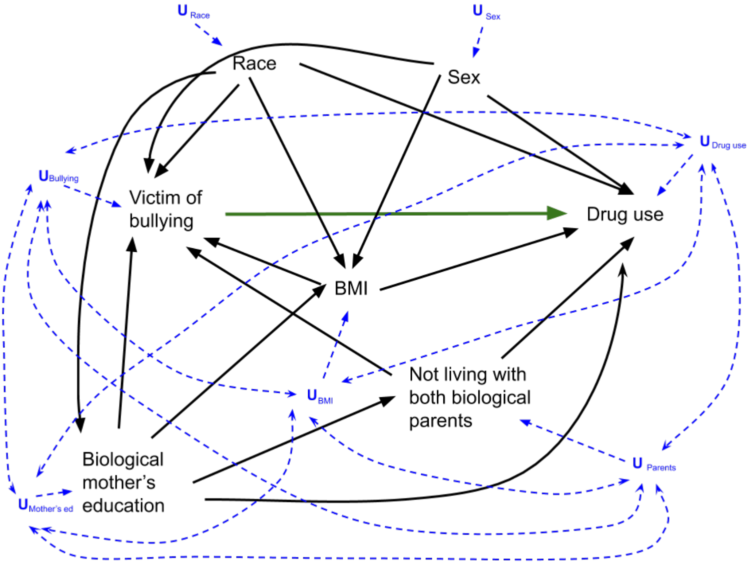
Table 1

Covariate	Drug use (%)	No drug us
Drug use (Total)	1330 (0.17266%)	6373 (0.82
Victim of bullying		
Yes	319 (3.7%)	1340 (17.4
No	1011 (2.9%)	1788 (23.2
Mother's education		
High school or less	3867 (50.2%)	732 (9.5%
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
White	815 (10.6%)	3245 (42.1
Living with both biological parents		
Lives with both biological parents	645 (8.4%)	3176 (41.2
Doesn't live with both biological parents	685 (8.9%)	3197 (41.5

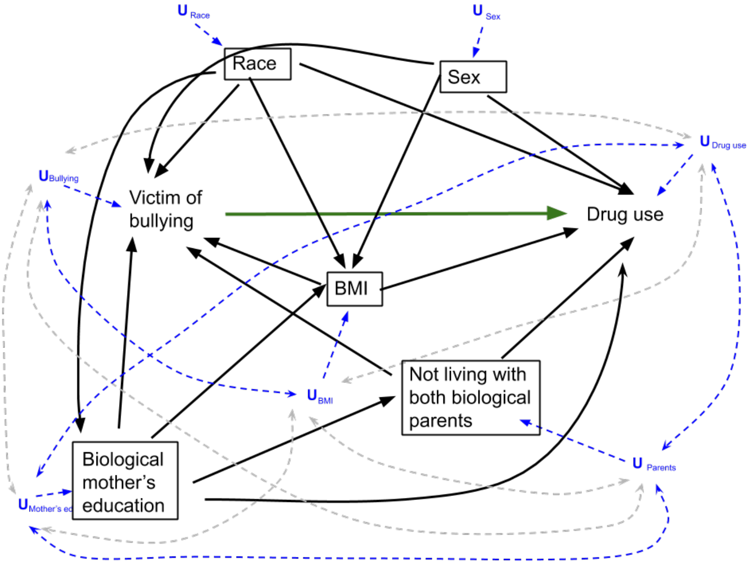
Identifiability



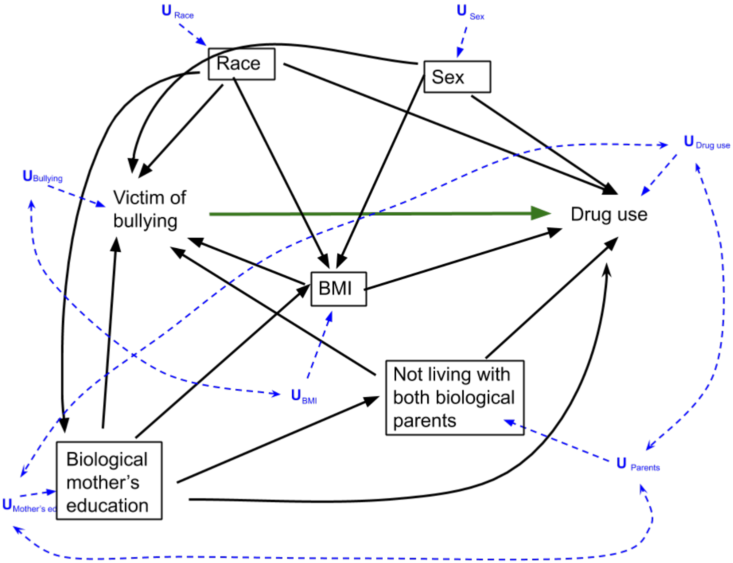
Identifiability



Identifiability



Identifiability



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:

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

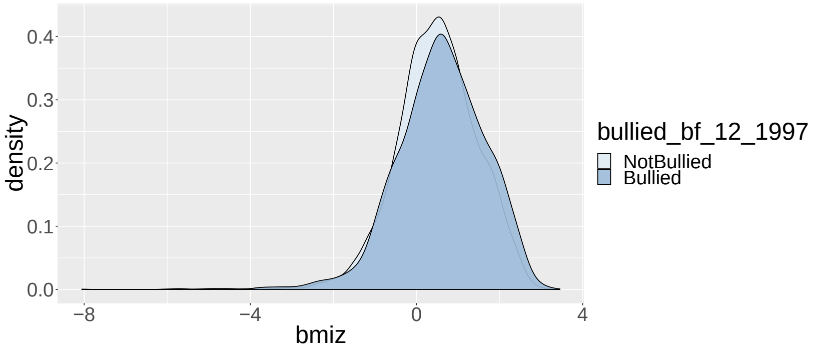
This is our statistical estimand.

Estimation: Unadjusted ATE & SuperLearner

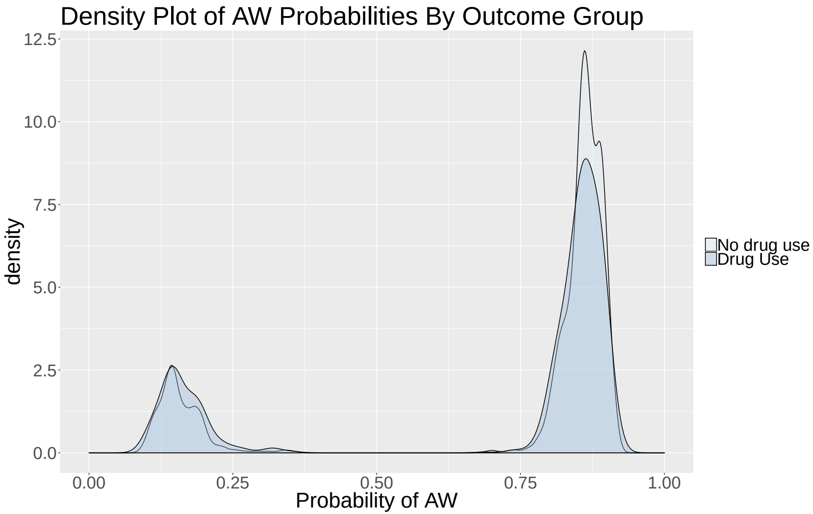
- The unadjusted ATE = mean(Y|A=1 - Y|A=0) = 0.05
- We use SuperLearner for prediction in all models.
- Library: SL.glm, SL.glm.interaction, SL.glmnet, SL.bayesglm, SL.randomForest, SL.step, SL.mean, SL.loglinear
- 5-fold cross-validation

Checking for Practical Positivity Violations

- We tabulated exposure and outcome across all possible levels of our categorical variables
- Observations exist in every possible category of our variable set
- Our 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



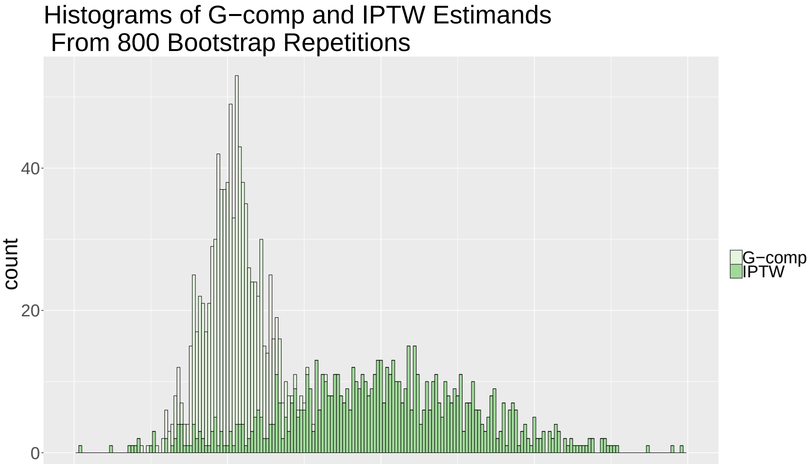
Positivity: Assessing the Model Weights



Confidence Intervals

- For TMLE, used the robust method built in to the ltmle package
- For G-comp and IPTW we performed a bootstrap

```
## Warning: Removed 4 rows containing missing values (geom_
```


Histograms of G−comp and IPTW Estimands
 From 800 Bootstrap Repetitions

Estimation: G-comp, IPTW, & TMLE

Estimator	ATE (95% CI)
G-computation	0.039 (0.017, 0.034)
Stabilized IPTW	0.045 (0.017, 0.081)
TMLE	0.044 (0.007, 0.08)

Estimation: SuperLearner convex combinations

Algorithm	A Risk	A Coefficient	Y Risk	Y Coefficient
glm	0.1549655	0	0.1408945	0.4628552
glm.interaction	0.155051	0.2086733	0.141501	0
glmnet	0.1549798	0	0.1409052	0
bayesglm	0.1549653	0	0.1408937	0
randomForest	0.1903391	0.4607251	0.1707125	0.2239811
step	0.1549472	0.267622	0.1408951	0.2476689
mean	0.1566939	0.0629796	0.1428772	0.0654947
loglinear	0.1549381	0	0.1409062	0

Estimation: SuperLearner performance

CV.SuperLearner

Algorithm	Avg Risk	SE
SuperLearner	0.1412062	0.0028693
Discrete SL	0.1407186	0.0027614
glm	0.1407131	0.0027616
glm.interaction	0.141278	0.0027675
glmnet	0.1407129	0.002763
bayesglm	0.1407127	0.0027616
randomForest	0.1678716	0.0041689
step	0.1407191	0.0027614
mean	0.1428802	0.0028201
loglinear	0.140726	0.0027623

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
- **causal interpretation**: if people are bullied they are about 4% more likely to use drugs later in life than if they are not bullied

Results

Estimator	ATE (95% CI)
G-computation	0.039 (0.017, 0.034)
Stabilized IPTW	0.045 (0.017, 0.081)
TMLE	0.044 (0.007, 0.08)

Limitations

1. Important exogenous variables
   -Parent drug use
   -Mental health
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
- Establishement 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
- Coding of SuperLearner estimation: Veronica
- Coding of Practical Positivity Checks and Bootstrapping of Confidence Intervals: Stephanie
- Interpretation of Results: Entire Group