Final Project: Childhood Bullying and Subsequent Drug Use

Shelley Facente, Steph Holm, Lizzy Kinnard, Veronica Pear

Spring 2019

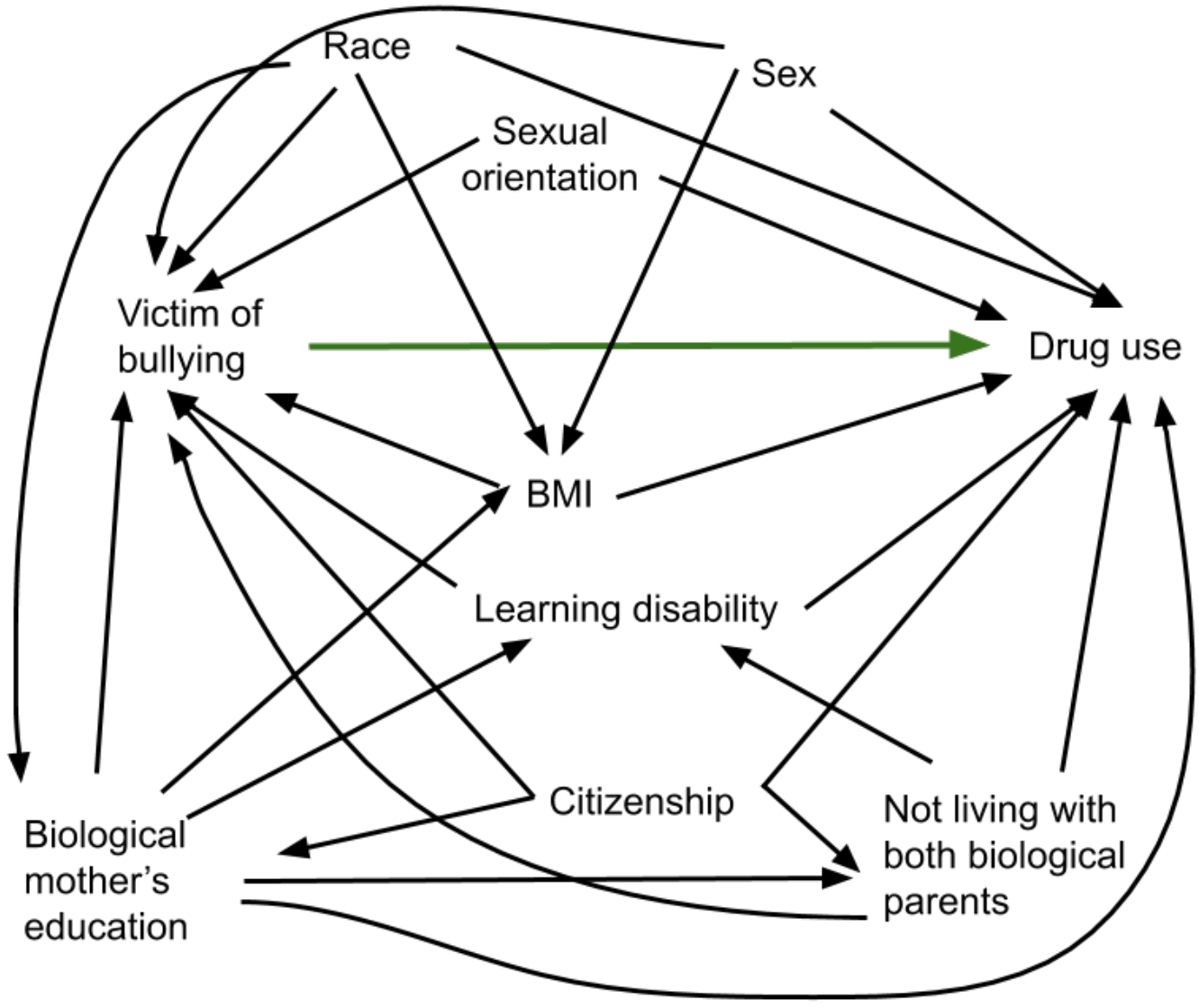
Causal Question: What is the effect of having been bullied prior to age 12 on incidence of drug use in adolescence or adulthood?

Specify a Causal Model

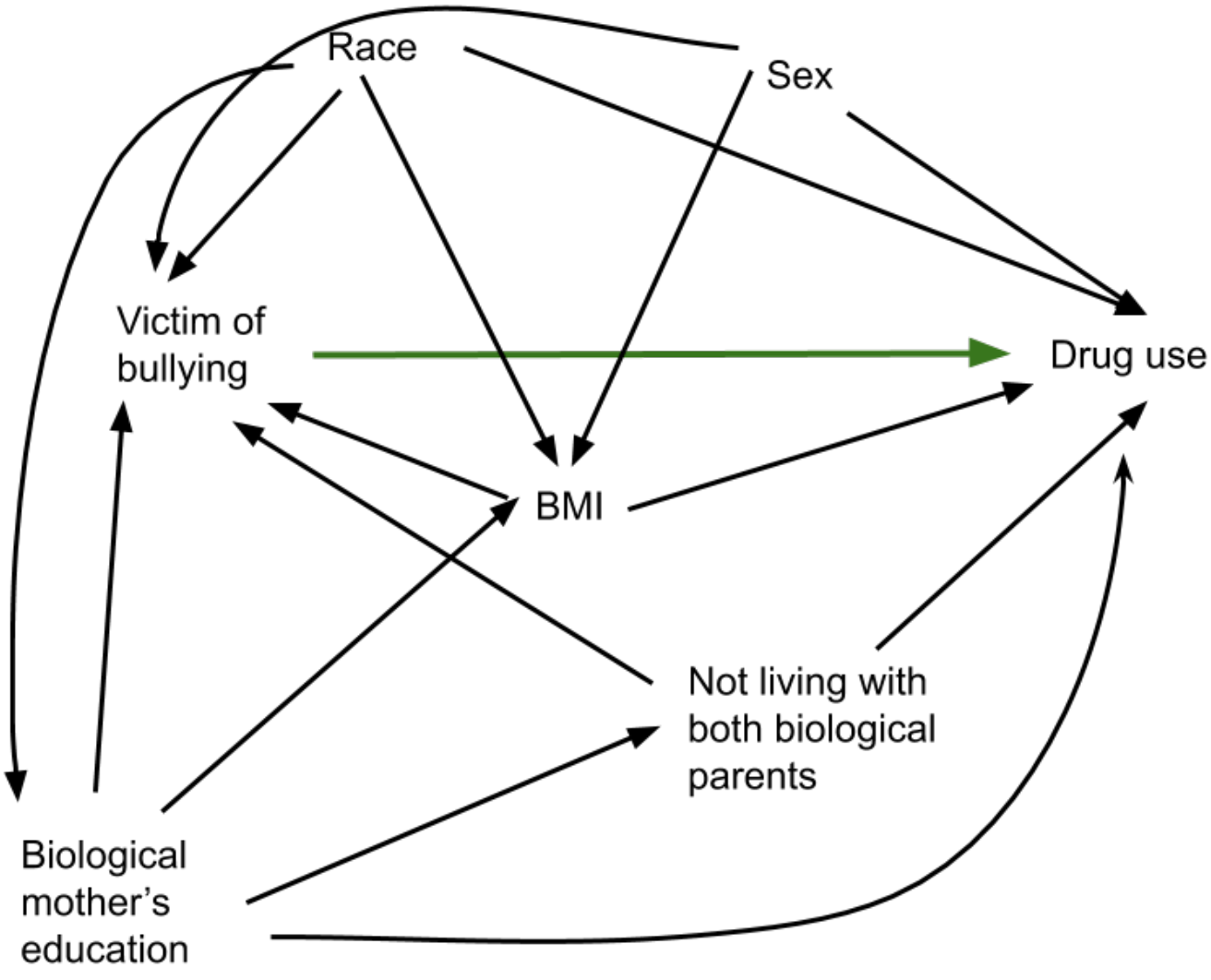
Our dataset:

- National Longitudinal Survey of Youth
- Nationally representative cohort of youth age 12-16
- Initial recruitment n = 9000 in 1997
- In our final dataset n = 7703

Original DAG



Final DAG


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

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.

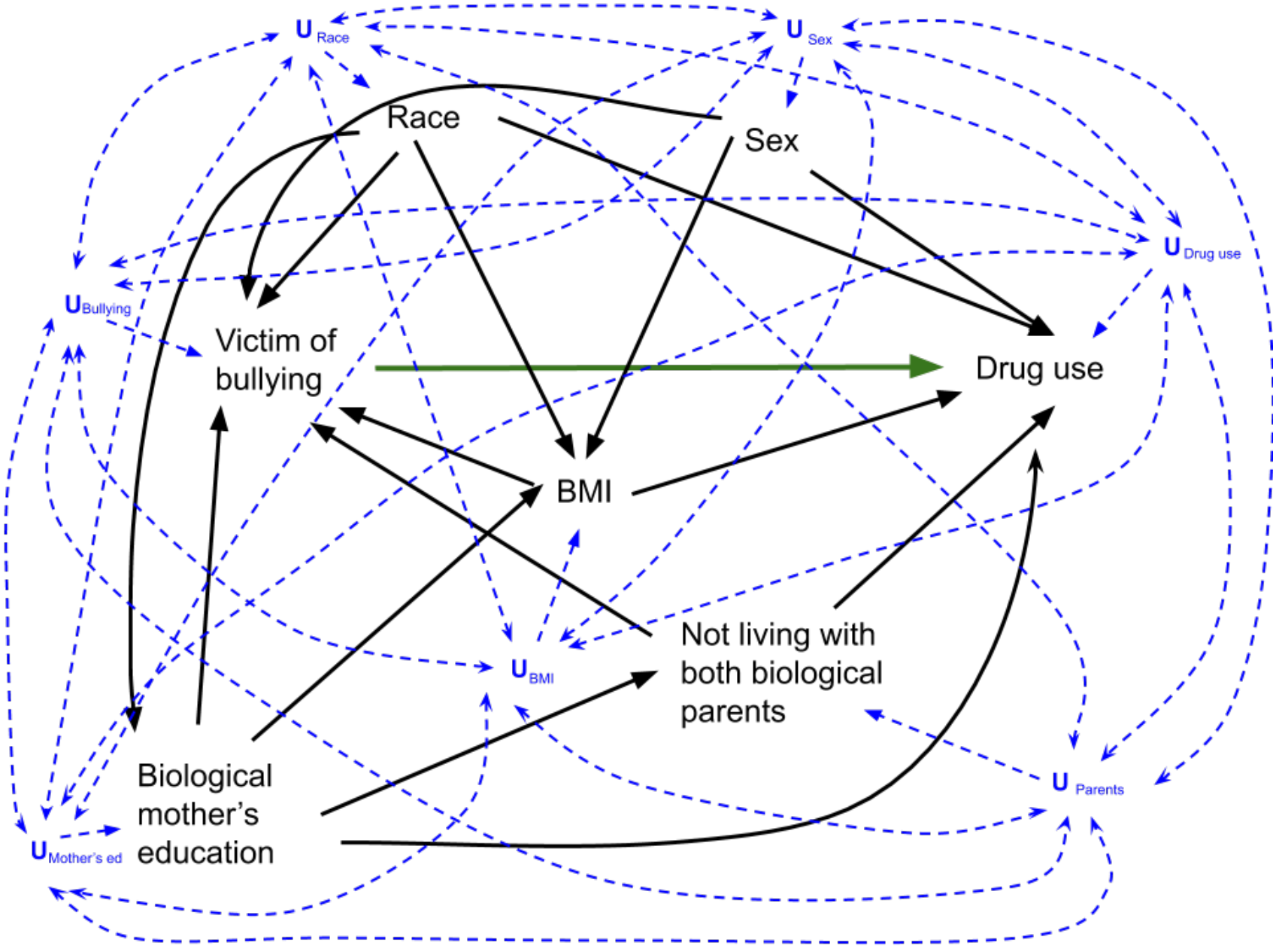
Link to the SCM

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.

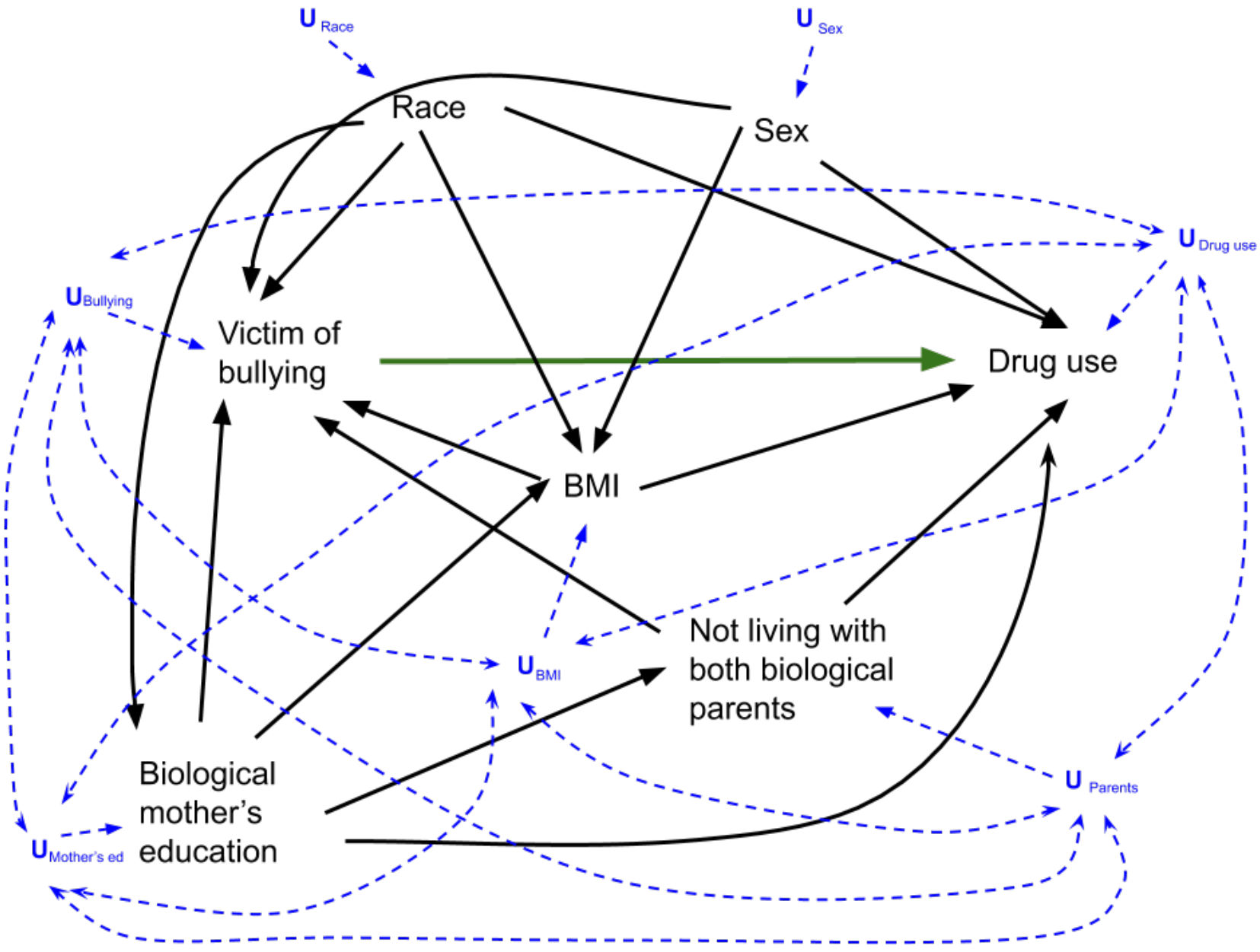
Table 1

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*)	0.505 (*mean*)
	1.03 (*sd*)	0.98 (*sd*)

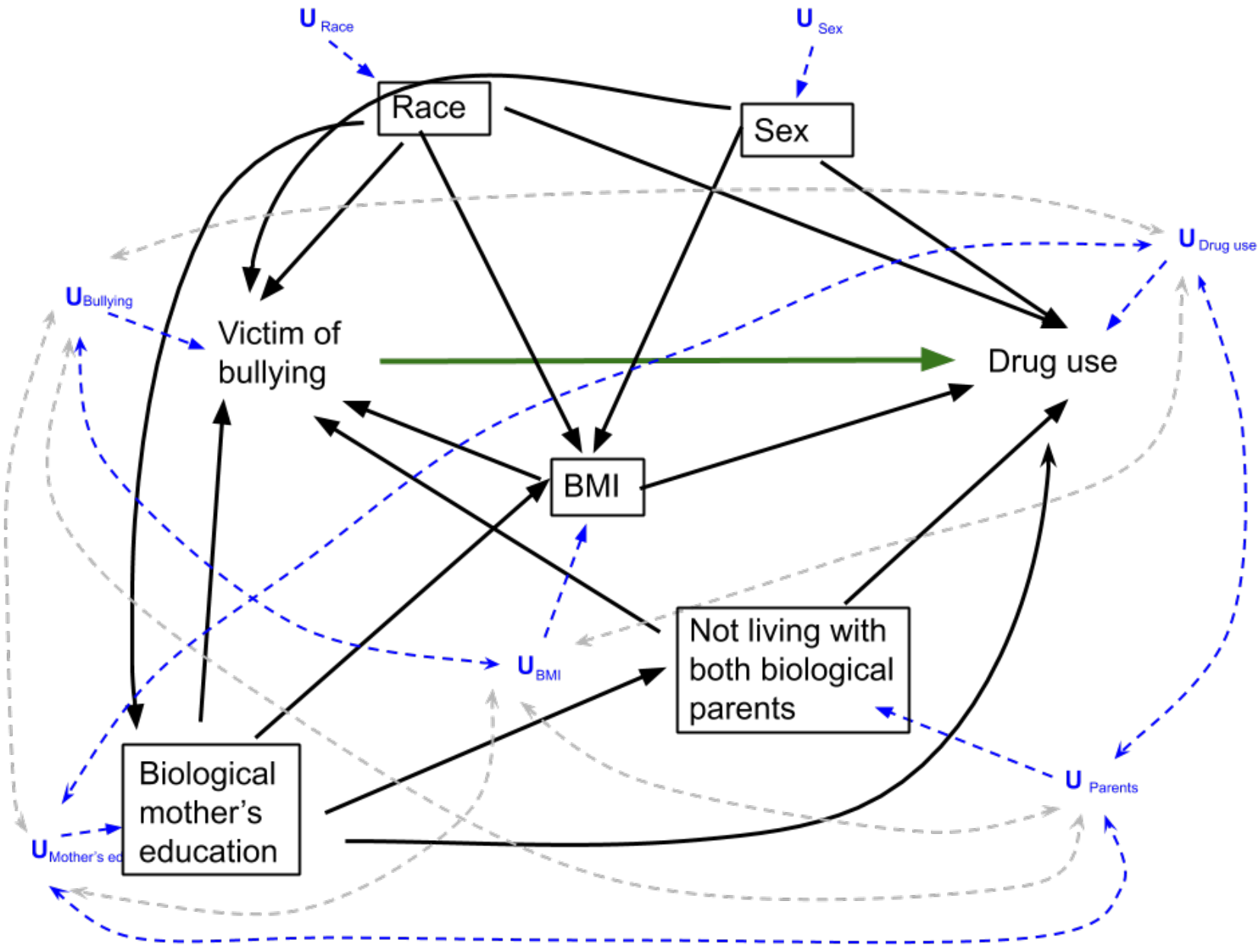
Identifiability



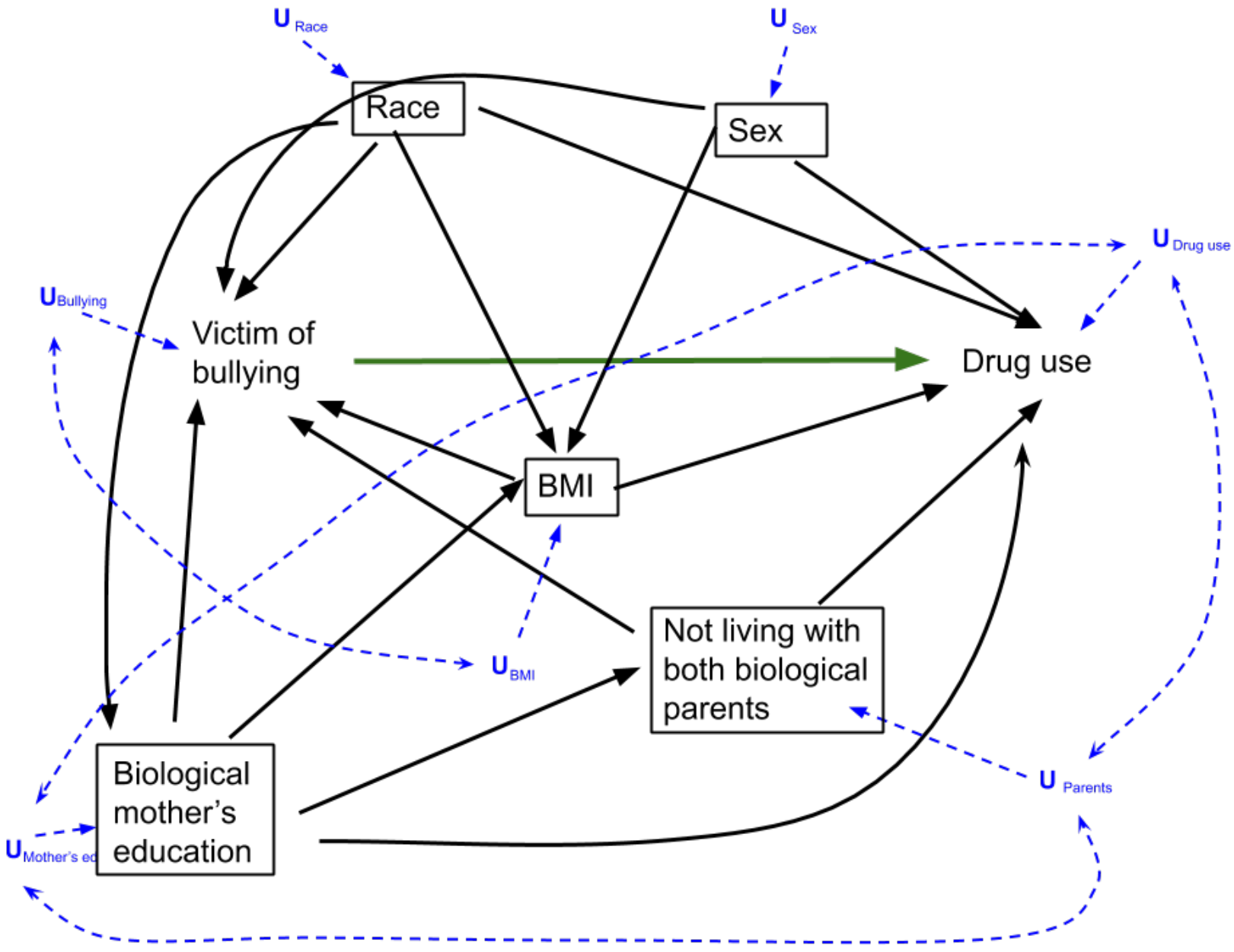
Identifiability



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:

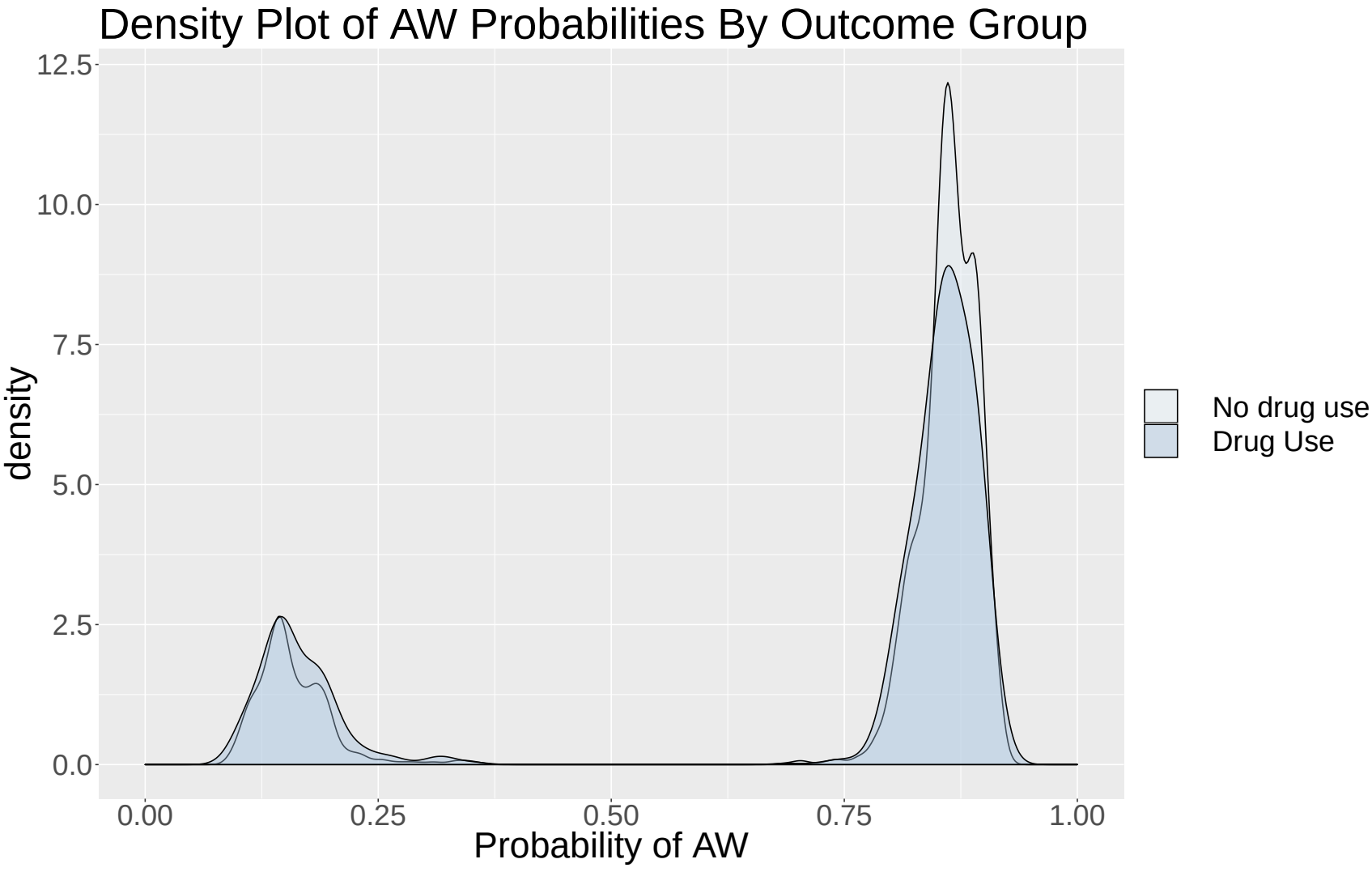
$$\psi(\mathbb{P}_0) = \mathbb{E}_0[\mathbb{E}_0(Y|A=1,W) - \mathbb{E}_0(Y|A=0,W)] =$$
$$\sum_{w1,w2,w3,w4,w5} [\bar{Q}_0(1, W1=w1, W2=w2, W3=w3, W4=w4, W5=w5) -$$
$$\bar{Q}_0(0, W1=w1, W2=w2, W3=w3, W4=w4, W5=w5)] *$$
$$\mathbb{P}_0(W1=w1, W2=w2, W3=w3, W4=w4, W5=w5)$$

This is our statistical estimand.

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
- 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

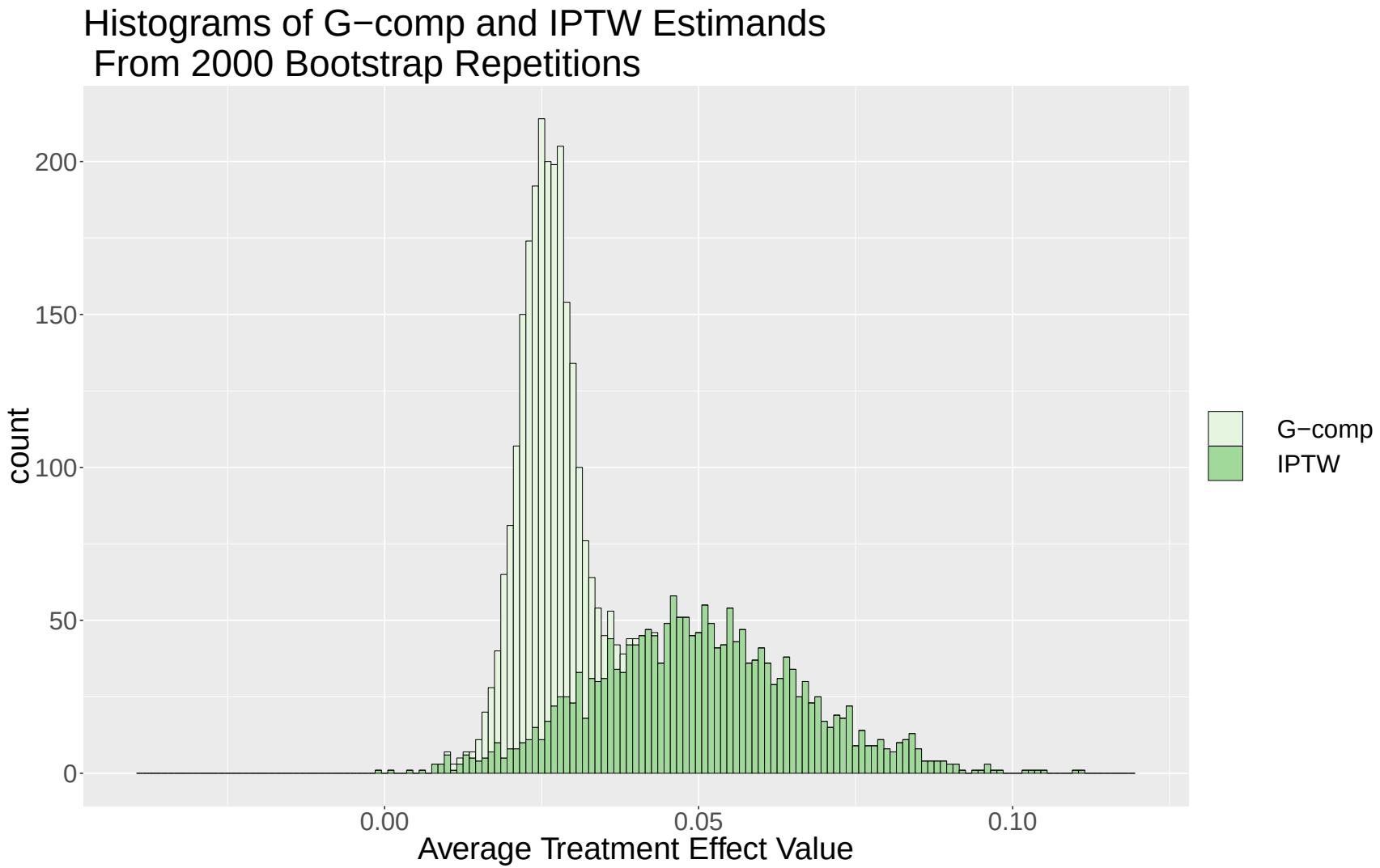


Estimation: Set-up

- 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

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



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
mean	0.14288	0.00282

Interpretation

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