

**Reliable and Fair Causal Machine Learning for Sparse  
Subpopulations in Survey Data: Substance Use and Major  
Depressive Episode among Reproductive-Age Women  
(NSDUH 2021–2023)**

Yifan Xu

This is an unpublished working draft prepared as part of an independent research project.  
Please do not circulate without permission.

## Abstract

### *Background:*

Major Depressive Episode (MDE) imposes a substantial health and social burden on reproductive-age women in the United States, with disproportionate consequences for pregnant individuals, low-income groups, and racial/ethnic minorities. Substance use—including illicit drugs, pain reliever misuse, stimulant misuse, and cannabis use—has been implicated as a potential risk factor for depressive symptoms, yet population-level causal estimates remain uncertain. Survey-based studies are further challenged by complex sampling designs, sparse subpopulations, and potential disparities in predictive performance across demographic strata.

### *Methods:*

Using the 2021–2023 National Survey on Drug Use and Health (NSDUH), this study constructed an analytic cohort of women aged 18–49. Four binary exposures (*ILL\_YR*, *PNRNMYR*, *STMNMYR*, *ANY\_CANNA\_EVER*) were evaluated against past-year MDE. Then developed a modular causal inference pipeline consisting of (A) *survey-weighted generalized linear models* (GLM) for baseline associations, (B) *Double Machine Learning* (DML) for average treatment effects (ATE), (C) *Causal Forest models for subgroup conditional average treatment effects* (CATE), and (D) *risk-model fairness assessments across pregnancy, income, and race subgroups*. All analyses incorporated survey weights and extensive covariate adjustment.

### *Results:*

Across exposures, weighted GLM models consistently produced positive odds ratios for MDE, though several estimates showed attenuation or wide uncertainty upon closer inspection. DML-based ATEs were generally modest in magnitude and frequently compatible with small or near-zero risk differences, with some exposures yielding broad confidence intervals indicative of limited subgroup information. CATE analyses revealed meaningful heterogeneity: pregnancy status, income terciles, and racial groups displayed distinct patterns of vulnerability, with certain strata showing elevated or reduced estimated effects relative to the overall average. Fairness assessments further indicated measurable disparities in predictive performance; in several demographic groups, metrics such as TPR, FPR, or PPV deviated from the population reference values, suggesting potential inequities in risk modeling.

### *Conclusions:*

This work demonstrates the feasibility and value of combining survey methodology with modern

causal machine learning to investigate sparse, high-risk subpopulations in national surveillance data. Although point estimates were often modest and uncertain, the consistent subgroup heterogeneity and fairness disparities underscore the need for caution when applying predictive tools in reproductive-age women, particularly pregnant individuals and underrepresented racial/ethnic groups. The pipeline provides a transparent, reproducible framework for future public health research on causal mechanisms and algorithmic equity in complex survey datasets.

**Keywords:** causal machine learning; Double Machine Learning; CATE; fairness; survey data; NSDUH; reproductive-age women; substance use; major depressive episode

## Introduction

Major Depressive Episode (MDE) is a leading contributor to morbidity among reproductive-age women in the United States. Beyond its immediate mental health implications, depression during reproductive years—including pregnancy—is associated with adverse outcomes such as substance use escalation, impaired maternal functioning, and intergenerational impacts. Understanding upstream modifiable factors is therefore a public health priority.

Substance use has long been examined as a correlate of depressive symptoms, yet national estimates of causal effects remain limited. Traditional epidemiologic analyses of survey data rely on regression adjustments that may not adequately address confounding, nonlinearities, or subgroup heterogeneity. Furthermore, reproductive-age women are not a homogeneous population: pregnant individuals, marginalized racial/ethnic groups, and lower-income women may represent sparse subpopulations whose experiences are poorly captured by standard models.

Recent developments in causal machine learning provide tools for addressing these gaps. Methods such as Double Machine Learning (DML) and Causal Forests allow flexible estimation of average and conditional treatment effects while accounting for rich covariates and complex outcome structures. However, the integration of these methods into survey-based public health research remains uncommon, partly due to challenges related to weighting, reproducibility, and the fairness evaluation.

This study aims to address these challenges by developing and applying a fully reproducible analytical pipeline to NSDUH 2021–2023 data. The objectives are:

1. to estimate the causal effect of four categories of substance use on past-year MDE among women aged 18–49;
2. to quantify subgroup heterogeneity—particularly for pregnancy status, income strata, and racial/ethnic groups;
3. to evaluate the fairness properties of predictive models across demographic subpopulations.

The work is conducted from the perspective of an independent researcher, with all analyses, code, and derivations publicly available.

## Methods

Below, the analytic pipeline is presented in four modular components (A+B+C+D) to mirror the accompanying source code and demo notebook.

### Data Source and Study Population

Data were obtained from the 2021–2023 National Survey on Drug Use and Health (NSDUH), which employs a multistage probability sample of the U.S. civilian, noninstitutionalized population. Following established practice, this study pooled three consecutive years and used the corresponding analysis weights (ANALWT2\_C1/C2/C3), rescaled by dividing by 3.

The analytic cohort consisted of women aged 18–49. Pregnancy status (IRPREG) was retained as a covariate and later used as a subgroup variable for heterogeneity analyses.

### Outcome

The primary outcome was past-year major depressive episode (MDE), defined using NSDUH's structured diagnostic algorithm (IRAMDEYR).

### Exposures

Four binary exposure variables were examined:

1. **ILLYR** – Illicit drug use
2. **PNRNMYR** – Pain reliever misuse
3. **STMNMYR** – Stimulant misuse
4. **ANY\_CANNA\_EVER** – Lifetime cannabis use

Each exposure was coded as 0/1. Analysis was conducted separately for each exposure.

### Covariates and Survey Weights

Covariates included: *AGE3*, *IRPREG*, *NEWRACE2*, *EDUHIGHCAT*, *INCOME*, *IRINSUR4*, *ANY\_NIC\_EVER*, *ALCMON\_bin*, and *YEAR*. Survey weights were normalized (*W\_NORM*) and applied in all analyses.

### Survey-Weighted GLM (Baseline Associations)

Weighted logistic regression models were used to estimate baseline associations between each exposure and MDE. Models incorporated W\_NORM and all covariates listed above. Odds ratios and 95% confidence intervals were extracted.

### Double Machine Learning (Avg. Treatment Effect)

For each exposure, this study estimated the **average treatment effect (ATE)** using the Double Machine Learning framework. Gradient boosting models were used for both outcome and treatment nuisance components. ATEs and confidence intervals were computed following cross-fitting procedures.

To maintain methodological transparency, no aggressive hyperparameter tuning was applied; the focus was on demonstrating consistency of the framework.

### Causal Forest (Subgroup Effect Heterogeneity)

To capture variation across demographic strata, this study applied Causal Forests separately for each exposure. The model used:

- Gradient boosted regressors (outcome)
- Gradient boosted classifiers (treatment)
- Preprocessing pipelines for categorical and numeric covariates
- Subgroup definitions based on
  - pregnancy status (IRPREG),
  - income terciles,
  - race (NEWRACE2)

CATEs were summarized using subgroup mean estimates.

### Fairness Evaluation of Predictive Models

Finally, this study assessed the fairness properties of exposure-based predictive models. Metrics derived from model predictions were stratified by pregnancy, income, and race groups. Disparities were evaluated via contrasts in TPR, FPR, and PPV across strata. Bootstrap resampling was used to quantify uncertainty.

## Results

### Baseline GLM Estimates (A)

Across exposures, weighted logistic models yielded positive odds ratios for the association between substance use and MDE. However, several estimates were attenuated after covariate adjustment, and some exposures displayed confidence intervals spanning values consistent with weaker or uncertain associations. These findings highlight the sensitivity of survey-weighted regression to sparse subgroup information.

### Double Machine Learning ATEs (B)

DML estimates produced modest average treatment effects across exposures. For several exposures, the 95% confidence intervals were wide and included values near zero, indicating limited precision. This pattern was particularly evident in exposures with small effective sample sizes within certain subgroups. The general magnitude of ATEs suggested that, at the population level, substance use was associated with small positive shifts in MDE risk, though estimates remained compatible with minimal effects.

### Subgroup Heterogeneity (C)

Causal Forest analyses revealed substantive effect heterogeneity:

- **Pregnancy:** Some exposures showed elevated estimated effects among pregnant women relative to non-pregnant women, though uncertainty remained considerable in several strata.
- **Income:** Income terciles displayed distinct patterns, with certain exposures showing stronger effects in the lowest-income tercile, and others showing minimal differences across strata.
- **Race:** Estimates varied across NEWRACE2 categories, with some racial/ethnic groups exhibiting attenuated or near-zero effects, while others showed comparatively larger estimated CATEs.

The consistent presence of heterogeneity—even when ATEs were small—underscores the value of modeling subgroup-specific dynamics.

### 3.4 Fairness Disparities (D)

Fairness evaluations demonstrated measurable disparities in model performance. Several metrics (e.g., TPR, FPR, PPV) differed across pregnancy, income, and race groups.

In particular:

- Some subgroups exhibited lower sensitivity, suggesting potential under-identification of elevated MDE risk.
- Other subgroups displayed higher false-positive rates, indicating possible over-identification.

Although effect directions varied across exposures, the presence of systematic deviations highlights fairness considerations when applying risk models in public health contexts.

#### **4. Discussion**

This analysis integrates survey methodology with causal machine learning to examine the relationship between substance use and MDE among reproductive-age women. Several insights emerge.

First, while baseline GLM models suggested positive associations, DML-based ATEs were generally modest and frequently compatible with near-zero effects. This contrast illustrates the importance of flexible confounding control, particularly when exposures correlate with socioeconomic and behavioral covariates.

Second, subgroup heterogeneity was pronounced. Pregnancy status, income terciles, and race each displayed distinct patterns of vulnerability. These findings align with prior public health literature emphasizing structural determinants of mental health and suggest that population-averaged estimates may obscure meaningful subgroup differences.

Third, fairness assessments indicated disparities in predictive performance across demographic groups, underscoring the need for algorithmic vigilance. The differential behavior of risk metrics across subpopulations has implications for clinical screening guidelines, data-driven risk stratification, and the ethical application of predictive models in reproductive health settings.

Several limitations warrant acknowledgment. NSDUH data remain cross-sectional and self-reported, limiting causal interpretation despite adjustment. Sparse subgroups contribute to wide uncertainty in effect estimates, and measurement constraints prevent nuanced characterization

of substance use patterns. Future work may incorporate longitudinal designs, domain-specific priors, or alternative machine learning architectures to improve robustness.

Despite these limitations, this study demonstrates a reproducible and transparent pipeline for integrating causal machine learning with complex survey data. The approach facilitates nuanced understanding of subgroup vulnerabilities and provides a framework for evaluating algorithmic fairness in public health contexts.