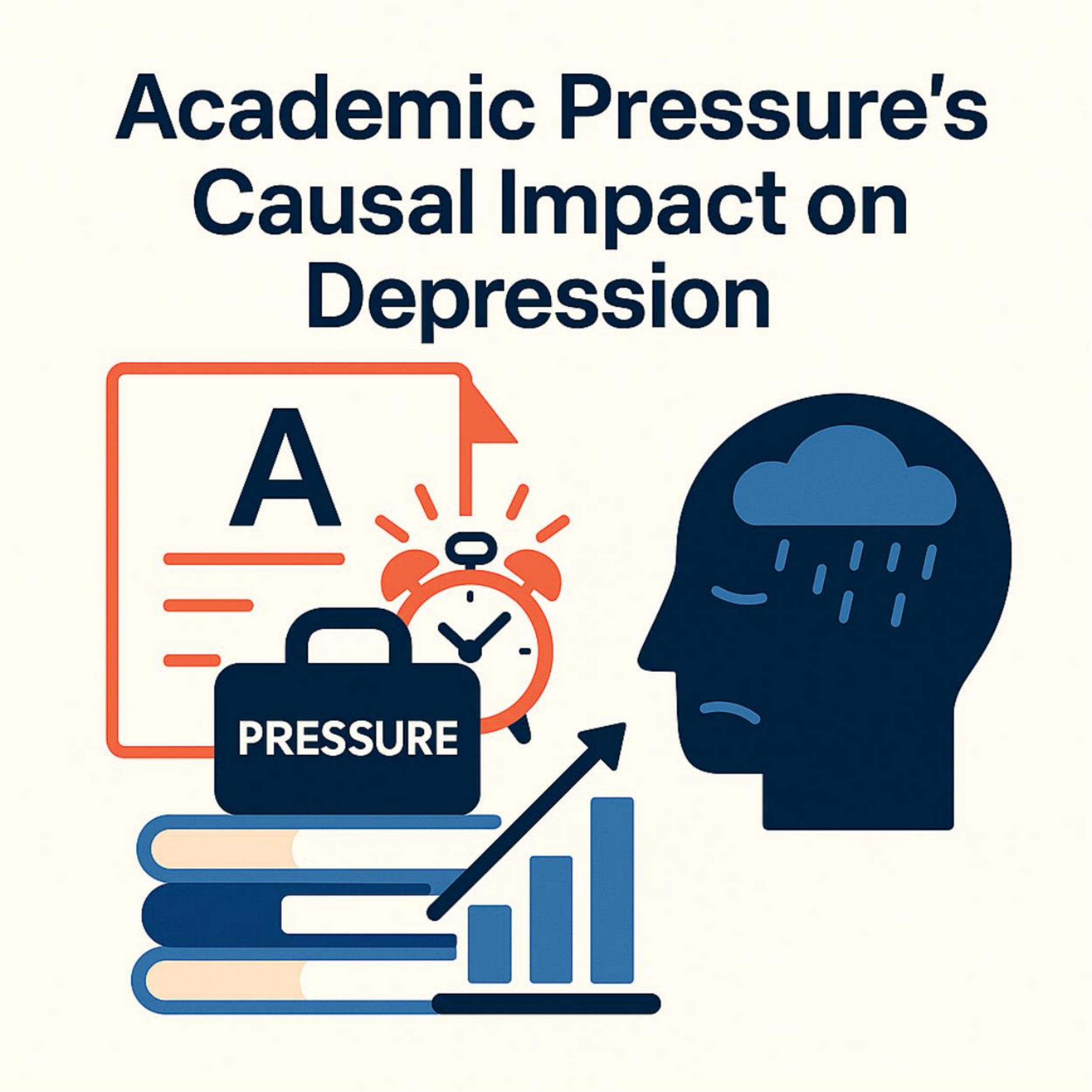
Applied Introduction to Causal Inference – Final Assignment Report – Student Depression Dataset

By: Tomer Rudnitzky, Yonathan Ghefter, and Idan Kanat

April 2025 – September 2025



Overview

Depression stands as one of the most significant and complex mental health challenges faced by individuals today. It is often difficult to identify clear causes, as it stems from a mix of psychological, biological, and social factors. The topic remains elusive, with no single explanation fully capturing its depth. This phenomenon is particularly relevant for students - the pressures of balancing academic demand, managing workloads, and preparing for future successful careers can often lead to heightened stress levels, which may contribute to mental health challenges.

Our study examined the causal relationship between academic pressure among students and the probability of experiencing depression. Formally, we addressed the following **causal question**:

"Among undergraduate students, does a "large enough" **Academic Pressure (Treatment)** have a causal impact over the probability of experiencing **Depression (Outcome)**?"

The Dataset

The dataset is called Student Depression Dataset and was obtained from Kaggle. It contains 27,901 non-null records, i.e. students, based in India. Notably, the dataset is synthetic, without an original real-world source. This suggests that the results of our analysis, including causal conclusions, may not fully generalize to real-world situations. The dataset contains a wide variety of features, which include (academic related factors in each category are highlighted in green):

- Demographics: Gender, Age, City, Degree.
- Mental Health History: Family History of Mental Illness, Suicidal Thoughts, **Depression Status**.
- Lifestyle & Well-being: Sleep Duration, Dietary Habits, Study Satisfaction.
- Academic Performance & Stress Factors: CGPA, **Academic Pressure**, Financial Stress.
- Behavioral Patterns: Study Hours.

The wide variety of features in the dataset offered substantial potential for causal analysis. Its unique focus on undergraduate students, just like us, made it especially relevant and relatable. This alignment allowed us to approach the causal analysis with a personal and informed perspective.

Causal Framework Specification

To ensure consistency throughout our work, we defined the following components:

- Population: Undergraduate university students, specifically from India.
- Units: Individual Students, enrolled in various academic programs.
- Time Zero: For simplicity, we assumed it to be the beginning of the academic semester.
- Treatment: Academic Pressure (will be further explained).
- Outcome: Depression Status - A binary variable indicating whether the student is experiencing depression or not.

Relevant Assumptions

To ensure clarity and coherence throughout our entire study, we explicitly state the key assumptions underlying our interpretation of the dataset. The following assumptions clarify how key variables are interpreted in the context of our causal analysis:

• Academic pressure is encoded in discrete 1-5 levels. Assumed to be self-reported during the semester.

• Self-reported values are assumed to be sufficiently reliable despite being subjective.

• Depression status is assumed to represent clinical or meaningful depressive symptoms, even though it's likely based on a simplified binary yes/no survey response. It is assumed to be measured by the end of the academic semester.

• CGPA reflects students' cumulative performance throughout the degree. Given by lecturers. Ranged at 1-10. Constant throughout the semester.

• Study Satisfaction is a general measure about the entire degree; not per semester, course / week.

• Timing - All remaining variables are measured in the beginning of the academic semester.

Exploratory Data Analysis (EDA)

To better understand the structure and characteristics of the dataset, we conducted a series of exploratory analyses focusing on key features such as academic pressure (treatment), depression status (outcome), CGPA, study satisfaction and more. We visualized the distributions of these variables and examined group-wise comparisons (e.g., by treatment status). Below lies a summary of the main visualizations and the insights they revealed.

Note: the dataset contained almost no missing values. This allowed us to proceed without serious imputation or exclusion, simplifying both exploratory analysis and causal estimation.

Treatment Distribution – Academic Pressure

Figures 1 and 2 present the distribution of academic pressure and its binary treatment version respectively. Figure 1a reveals non-uniform distribution over the discrete 1–5 scale, with peaks at both extremes (1 and 5) and at the center (3). This suggests that while many students report moderate levels of academic pressure, a substantial number also report experiencing either very low or very high pressure.

Based on the distribution in Figure 1, we observed that the closest representable value to the empirical median of academic pressure lay at 3. Given the discrete 1-5 scale, we selected 3 to guide the construction of a binary treatment indicator: such that values >= 4 indicated high academic pressure (treatment = 1), and values <= 3 indicated low or moderate pressure (treatment = 0). The threshold was chosen to reflect the upper range of the pressure scale while achieving as balanced a treatment allocation as possible, as seen in Figure 2, although moderate imbalance remained in favor of the untreated (lower academic pressure group).

This binarization simplified the causal analysis and aligned with our interest in studying the causal effects of elevated pressure. However, it also introduced potential limitations. Most notably, it combined multiple adjacent pressure levels into broad categories (e.g., low = 1–3, high = 4–5), potentially masking variation within each group. Beyond obscuring important distinctions, it could introduce bias if treatment effects vary within these aggregated levels. We proceeded under the assumption that any such intra-group variation was limited or ignorable for the purposes of our analysis. Additionally, we relied on covariate adjustment and stratification techniques in subsequent analyses to mitigate potential confounding and imbalance between treatment groups.

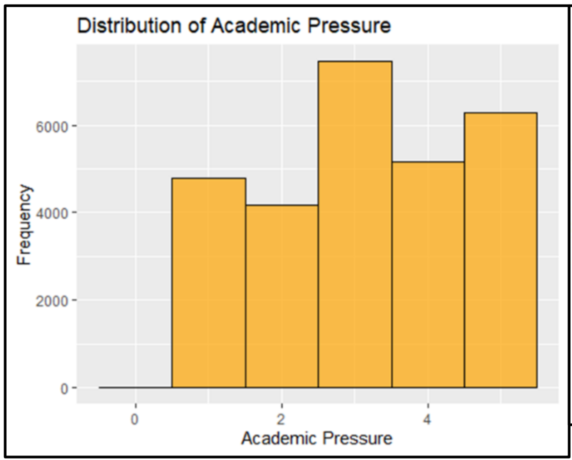


Figure 1 – Academic Pressure distribution

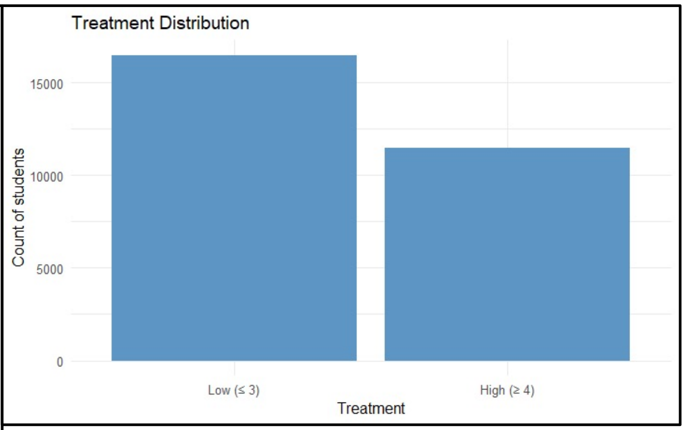


Figure 2 – Treatment distribution

Outcome Distribution – Depression Status

Figure 3 plots the marginal outcome distribution, depression status. 58.55% of the dataset's students are experiencing depression – a clear majority. However, Figure 4 provides further insight by stratifying depression status according to treatment groups. Among students with high academic pressure, most report experiencing depression. In contrast, among those with low or moderate pressure, the majority is not depressed.

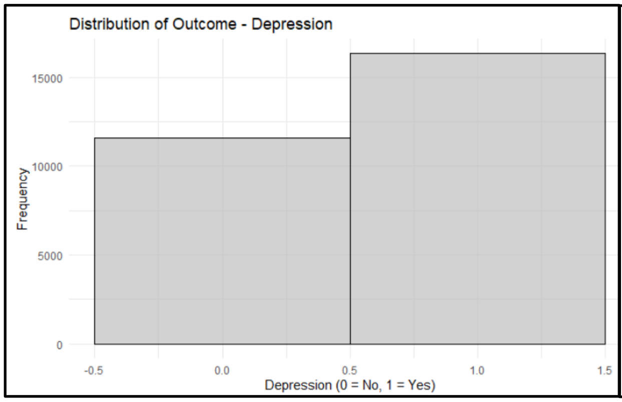


Figure 3 – Depression status distribution

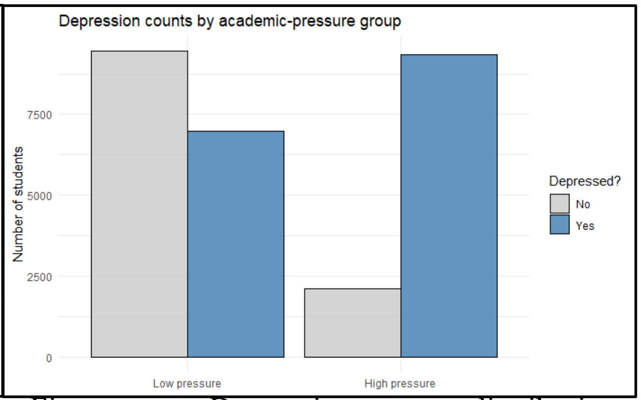


Figure 4 – Depression status distribution, stratified by treatment

To intuitively quantify the strength and direction of this association, we computed the naïve odds ratio (OR) using the observed frequencies in Table 1 below. Let $n_{i,j} = \sum_{k=1}^{N=27,901} 1_{\{A_k=i, Y_k=j\}}$ denote the number of students with depression status $Y = j$ and treatment $A = i$, where $i, j \in \{0,1\}$.

Table 1 – Frequency Table – Outcome (Depression Status) stratified by treatment

Academic Pressure (A) / Depression (Y)	Depressed (Y=1)	Not Depressed (Y=0)	
High Academic Pressure (A=1)	$n_{1,1} = 9{,}345$	$n_{1,0} = 2{,}106$	$n_{1,.} = 11{,}451$
Low Academic Pressure (A=0)	$n_{0,1} = 6{,}991$	$n_{0,0} = 9{,}459$	$n_{0,.} = 16{,}450$
	$n_{.,1} = 16{,}336$	$n_{.,0} = 11{,}565$	$N = 27{,}901$

The resulting naïve Odds Ratio turned out to be especially alarming:

$$OR = \frac{\dfrac{P(Y=1|A=1)}{P(Y=0|A=1)}}{\dfrac{P(Y=1|A=0)}{P(Y=0|A=0)}} = \frac{\dfrac{P(Depressed|"High\ Pressure")}{P(Not\ Depressed|"High\ Pressure")}}{\dfrac{P(Depressed|"Low\ Pressure")}{P(Not\ Depressed|"Low\ Pressure")}} = \frac{\dfrac{n_{1,1}}{n_{1,0}}}{\dfrac{n_{0,1}}{n_{0,0}}} = \frac{\dfrac{9{,}345}{2{,}106}}{\dfrac{6{,}991}{9{,}459}} \approx 6$$

This reveals that the odds of experiencing depression are 6 times higher among the treated (high pressure group), than among the untreated (low pressure group). Although this estimate does not account for potential confounding, it highlights a strong unadjusted positive association between academic pressure and depression status and motivates further causal analysis to determine whether this observed disparity is robust to confounder adjustment under the causal inference framework.

Distribution of Other Key Covariates

Figure 5 summarizes the distributions of four key covariates relevant to our analysis:

- Age is right-skewed with several anomalies, though most students fall within the 20-35 range, suggesting a young population - potentially including recent high school graduates.
- CGPA spans the full 5-10 range with moderate spread. Extremely low values (e.g., 5) are less common, while scores from 7 and above are relatively frequent.
- Degree is highly imbalanced, with a dominant majority reporting completion of Class 12, likely reflecting a pre-college or high-school level.
- Study Satisfaction is fairly balanced across the 1-5 scale, though level 4 is somewhat more frequent, and level 5 slightly less common.

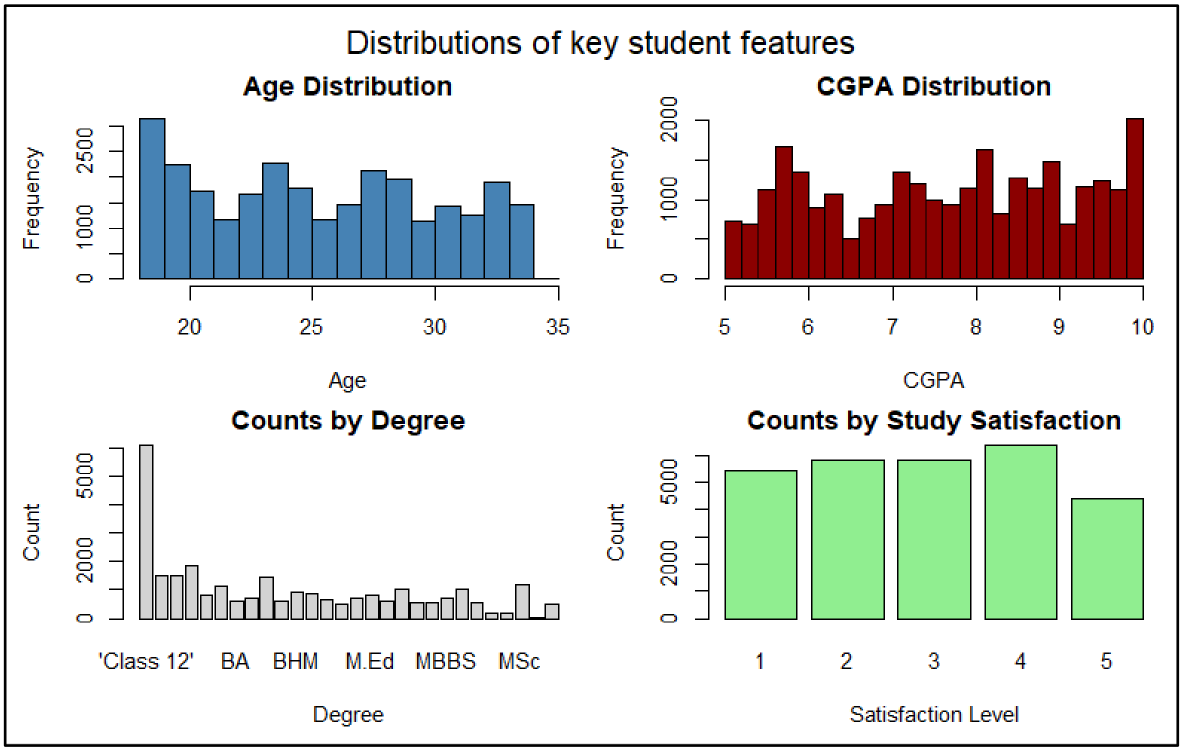


Figure 5 – Distribution of four key covariates: Age (top left), CGPA (top right), degree (bottom left), study satisfaction (bottom right).

These patterns informed our preprocessing decisions, further detailed in the corresponding section.

Preprocessing

To reduce sparsity and facilitate the forthcoming causal analysis - particularly with regard to the positivity assumption, we applied the following transformations on our data:

- Age and CGPA, originally continuous, were transformed into 4 quartile-based bins each to ensure approximately balanced groups.
- Degree categories were collapsed into 6 broader fields of study, addressing extreme imbalance and ensuring adequate representation across groups.
- Study satisfaction was retained as a 5-level ordinal variable.

These transformations simplified modeling and increased the likelihood of observing treated and untreated students across all covariate strata. In total, we defined 4 (Age bins) × 4 (CGPA bins) × 6 (Fields of study) × 5 (Satisfaction levels) = 480 covariate strata.

While perfect positivity is rarely achievable in high-dimensional settings, the above structure yielded only minor violations, which we discuss later. This preprocessing step was thus instrumental in supporting valid causal identification in the presence of treatment heterogeneity. Thus, we continued exploring the data using the preprocessed dataset, inspecting more fine-grained patterns, detailed below.

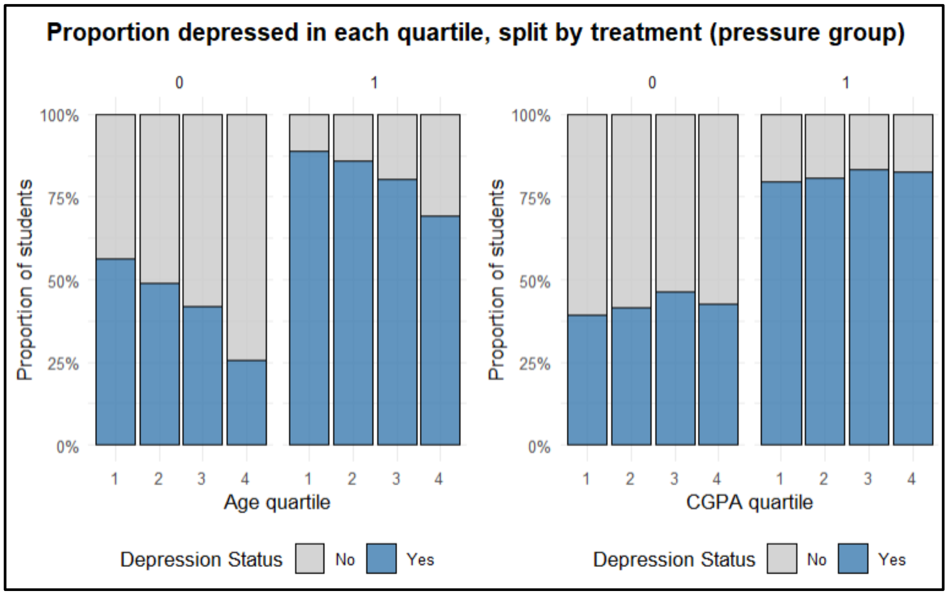
Outcome Distribution across Preprocessed Key Covariates

To better explore outcome patterns across subpopulations, we used the transformed dataset described in the preprocessing section. Figures 6a and 6b show the proportion of students experiencing depression within each age (left) and CGPA (right) quartile bucket, stratified by academic pressure (treatment group).

Several patterns emerge:

- Age: Within both treatment groups, depression prevalence decreases with age, with younger students report depression more frequently. Moreover, within every age quartile, the proportion of depressed students is consistently higher among students under high academic pressure than in the corresponding low-pressure group.
- CGPA: The proportion of students experiencing depression remains relatively stable across CGPA quartiles within each pressure group. However, within each CGPA group, those under high academic pressure consistently show a higher proportion of depression than students in the same CGPA group under low pressure.

These findings underscore both the overall burden of depression under pressure, and the presence of treatment effect heterogeneity (w.r.t to key features such as age and CGPA), further motivating the use of subgroup-specific causal methods.



Figures 6a & 6b: Depression status distribution by Age (left) & CGPA (right) quartiles, stratified by treatment.

Roles of Covariates

Directed Acyclic Graph (DAG) – Our Proposal

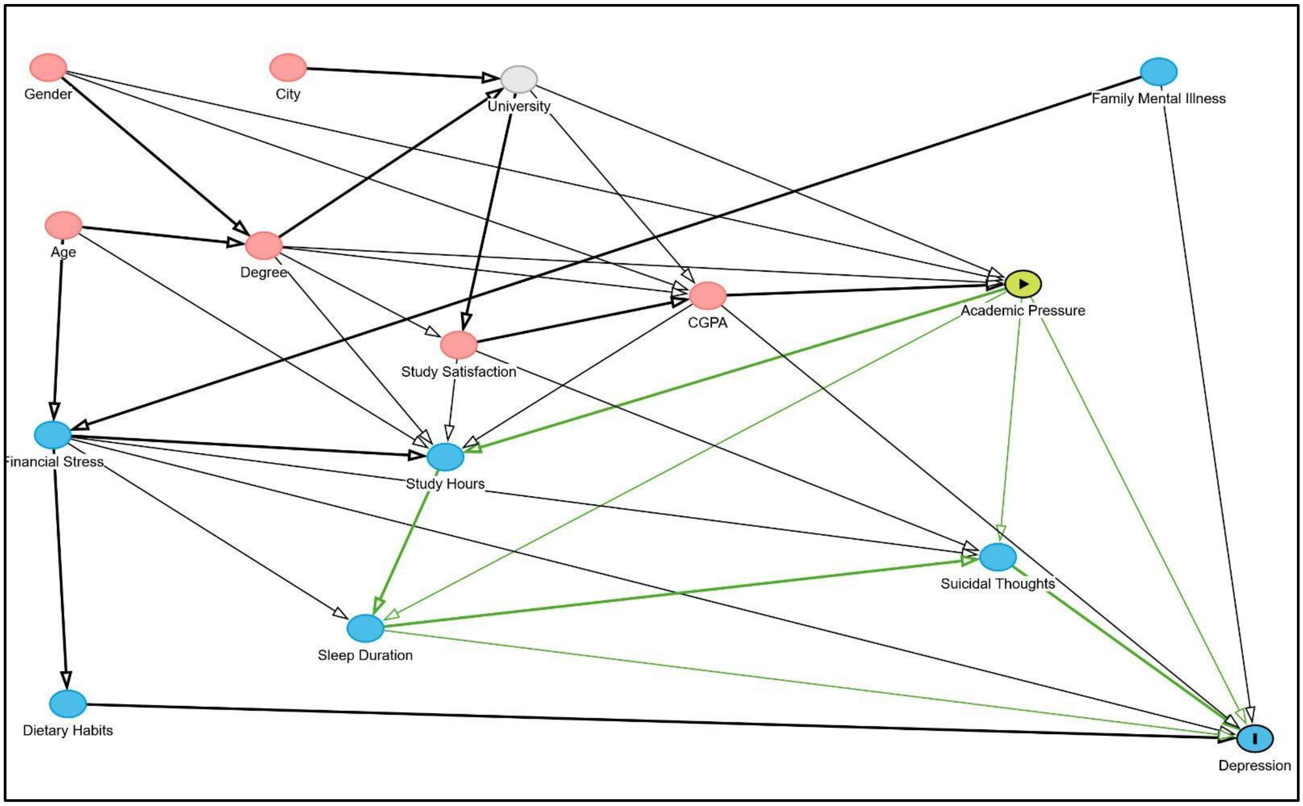


Figure 7: Our proposed DAG.

Main Assumptions from the DAG

Given the numerous paths in our DAG, we focused on some of the main encoded assumptions:

- Conditioning on CGPA, study satisfaction, and degree program (the parents of study hours), gender is independent of study hours. Gender influences those parents, but not study hours directly.
- Degree → University (and not vice-versa): Students are assumed to decide on their degree before choosing a university, which affects their choice of institution.
- Financial Stress → Dietary Habits → Depression: Financial stress influences diet (due to impact on food budget), and diet affects mood and risk of depression.
- Study Hours and Academic Pressure → Sleep Duration: More studying and higher pressure reduce sleep quantity and quality.
- CGPA and Family History of Mental Illness: are assumed independent.
- Family History of Mental Illness → Depression: Genetic vulnerability and shared household stresses increase depression risk.

- Degree → CGPA: Different majors have distinct grading norms and workloads, which shift GPA distributions.
- Suicidal Thoughts and CGPA: are assumed independent after conditioning on academic pressure, study satisfaction, financial stress, and sleep duration.
- Sleep Duration and Study Satisfaction → Suicidal Thoughts: Less of both may increase risk of suicidal thoughts.
- **Contentious discussion point**: The role of gender on CGPA was debated. Specifically, whether gender is independent of CGPA given degree, study satisfaction, and university. Including this edge or not does not affect the adjustment set, so it remained in the DAG .

Confounders

As the name suggests, confounders are defined as covariates which can distort, bias, or indeed, confound the relationship between the treatment A (high academic pressure), and outcome Y (depression status). Alternatively and more formally, confounders are any set of covariates X which satisfies the back door criterion, w.r.t to (A, Y), discussed in class. This ensures that when conditioning upon them, any non causal path from A to Y is blocked, allowing for identification of the Average Treatment Effect (ATE) and rendering its estimation a well defined task. Examples from the DAG include:

- **Age**: Older students may face additional responsibilities (e.g., jobs, family), raising both academic pressure and depression risk.
- **Gender**: Gender differences can shape exposure to academic pressure and depression risk.
- **CGPA**: Past academic performance affects academic pressure (students with lower CGPA may feel more pressure) and may strongly be associated with depression risk.
- **Degree**: Degree type (i.e. STEM/not) influences both workload and depression prevalence.
- **Study Satisfaction**: Satisfaction with studies may reduce perceived pressure and independently affect mental health outcomes, making it a confounder.
- **University (Unmeasured)**: University is not listed in our dataset, yet prestigious universities may increase academic workload and depression risk through competitiveness.

Effect Modifiers

Effect Modifiers are defined as variables which affect the strength and direction of the treatment effect. In other words, the Average Treatment Effect (ATE) may differ across levels of these variables. In our data, we have several potential effect modifiers:

- **Degree**: The effect of larger academic pressure on depression may be stronger in demanding STEM programs compared to project-based arts programs.

- **CGPA**: Additional academic workload might reduce stress for high-CGPA students (who feel rewarded by performance) but increase stress for low-CGPA students (who may not see improvements despite more effort).
- **Age**: Younger students may be more resilient to academic pressure, while older students (who may balance work or family responsibilities for example) might experience stronger negative effects.
- **Gender**: The impact of academic stress on depression could differ by gender, due to social expectations or support systems.
- **Study Satisfaction**: For satisfied students, academic pressure might be motivating, whereas dissatisfied students may experience greater risk of depression.
- **Family History of Mental Illness**: Students with such a history may be more vulnerable, amplifying the negative effect of stressors.
- **Study Hours**: Longer study hours may amplify the impact of high academic stress on depression risk, as sustained workloads can intensify mental health issues among students.
- **Sleep Duration**: Sleep loss may intensify the harmful effect of sustained academic pressure on depression, while sufficient sleep could buffer it.
- **Financial Stress**: Students under financial strain may be more affected by additional academic workload compared to their more financially secure peers.

Mediators

Mediators are defined as variables which mediate the treatment effect of A on the outcome Y. These variables lie on the causal path between A to Ysuch that part (or all) of the treatment's effect on the outcome is transmitted through these variables. The mediators in our DAG are:

- **Suicidal Thoughts**: Academic pressure can increase these negative thoughts, which might directly increase the likelihood of depression.
- **Sleep Duration**: Academic pressure might cause students to cut sleep time, negatively affecting their sleep quality, which in turn may turn students more vulnerable to depression.
- **Study Hours**: Together with Sleep Duration, it forms part of the pathway linking treatment to outcome. We assumed that higher academic pressure leads to more intensive study hours, which in turn reduces sleep duration. This increases the risk of depression, as discussed previously.

Colliders:

Colliders are variables that receive 2+ incoming arrows in a DAG, i.e., they sit at the intersection of multiple causal paths. Below we mention some examples for colliders that arise from the DAG:

- **Study Hours**: For instance, students who feel high academic pressure may increase their study time, while financial stress can reduce it if they need to work.

- **Sleep Duration**: Long study hours can cut into sleep. Stress can make good rest harder too.
- **Suicidal Thoughts**: For example, academic pressure can contribute to negative thoughts, and poor sleep can worsen mental state, potentially increasing the risk of suicidal thoughts.
- **CGPA**: For instance, certain degrees have stricter grading norms, and students more satisfied with their studies may often achieve higher scores than those who are not.
- **University (Unmeasured)**: For instance, students often choose their university based on the degree they wish to pursue, while the city they live in can also play an important role.

Instrumental Variables (IVs)

An instrumental variable (IV) is a variable that influences treatment assignment but does not directly affect the outcome or share unmeasured causes with it. A ***causal IV*** must affect treatment (relevance - i.e. change the probability of receiving it), affect the outcome only through that treatment (exclusion restriction), and remain independent of unmeasured confounders that jointly affect treatment and outcome (independence). A ***non-causal IV*** is a weaker version: it only needs to be associated with treatment, and not associated with the outcome except through that link. In both cases, an IV generates exogenous variation in treatment assignment - that is, variation not driven by confounders, which can be used to identify causal effects even when confounding prevents valid estimation through standard adjustment methods. From the DAG, there exist no such variables, since no variable satisfies the exclusion restriction condition, i.e. none of the variables influencing depression do so exclusively through academic pressure.

Identification Assumptions

To ensure the validity of causal effect estimation from our observational data, several key assumptions must hold:

SUTVA (Stable Unit Treatment Value Assumption)

SUTVA builds on two components. By the data's design, we believed it's reasonable to assume:

- No Interference: One student's depression status is independent of others' academic pressure. Although students can be influenced by shared academic environments or peer effects, we considered them marginal and ignorable for our analysis.

  We assumed so for several reasons: (1) the data has no group or network structure - students are treated as independent units; (2) academic pressure is mostly driven by course load and not social dynamics; (3) the university setting is diverse and loosely coupled, which limits peer influence; and (4) shared courses might raise pressure for multiple students, but that's a common cause - not interference. One student's pressure doesn't influence the joint course, nor another student's depression status. So it's not a violation of SUTVA - but rather just shared exposure.

- No Multiple Versions of Treatment: Academic pressure was measured on a 1–5 scale and encoded as a binary variable: high (4–5) vs low (1–3). While this may group students with different pressure levels (e.g., 1 vs. 3), we assumed the causal effect within each group is similar enough to justify a single treatment version.

Conditional Exchangeability

Formally, this assumption states that for all $a \in \{0,1\}. Y^a \perp A|X$. Essentially, conditional on a set of observed confounders X we adjust on, the counterfactual depression outcomes Y^a are independent of the actual treatment allocation (i.e. high / low academic pressure) A. This implies that given X, treated and untreated students are exchangeable w.r.t to their outcomes!

Based on our DAG, we identified two minimal covariate sets that satisfy the backdoor criterion:

1. {**CGPA, Degree, Age, Study Satisfaction**}.

2. {**CGPA, Degree, Gender, Study Satisfaction**}.

To reduce random positivity violations in our analysis, we chose to condition on the first set of confounders. Thus, the final adjustment set X (i.e. confounders we adjusted for) for which we have conditional exchangeability is: **Study Field**, **Age / Quartile**, **CGPA / Quartile**, and **Study Satisfaction Level**.

Positivity

For our binary treatment case, this assumption formally states that there are treated and untreated students (i.e. the propensity score is never 0 or 1) given any value of X, where X is the set of confounders we adjusted for. Mathematically, for all $x \in S_x, a \in \{0,1\}. P(A = a|X = x) \in (0,1)$.

A sufficient condition which would satisfy positivity would be: for every covariate combination $X = x$, there exist both treated and untreated students - i.e. students under both high and non-high academic pressure. However, lack of matching observations does not necessarily mean structural violations.

To empirically assess whether the positivity assumption holds in our dataset, we initially divided the data into all 480 possible strata of covariates X (after discretizing continuous variables such as age and CGPA to quartiles to reduce sparsity). We then checked whether each stratum (i) was present in the data, and (ii) contained both treated and control units. Table 2 summarizes the number of strata with violations.

Table 2 – Summary of *X* strata with positivity violations

Violation Type	Without Treated	Without Control	Nonexistent in the Data	Total
Count	4	9	3	16

Out of 480 possible strata, 477 existed in the data, leaving 3 nonexistent strata in our data. Among the 477 observed strata, an additional 13 lacked either treated or control units. In total, this initial check revealed **16 violations out of 480 (≈3.3%)**, a negligible share. Nevertheless, since this was only an initial assessment, we proceeded with deeper inspections to ensure robustness.

Beyond looking at the raw violation counts, Figure 8a illustrates a deeper inspection and a core challenge: students younger than 21 are overwhelmingly concentrated in the ″Class 12″ (high school) group, suggesting a source of potential positivity violations. These younger individuals likely represent a distinct population segment not comparable to older university-level students. This implies that certain covariate combinations; such as young students in STEM or Health are hard to find from the dataset. These gaps are possibly due to sampling noise: students under 21 typically come from high school. As a result, we cannot empirically observe both treated and untreated units for all strata of X for students under 21, imposing a random violation of the positivity assumption.

To address this, we considered restricting the analysis to students aged 21 and above, where treatment assignment is less tightly bound by structural age–field of study entanglement. This did not mean excluding all "Class 12" students, since many aged 21+ remained in that group and were still relevant to our target population. Also, since we had so few observations with age>35 we decided to remove these from our analyses, as they would've negatively impacted the variance of our estimators while not providing meaningful results in the >35 age range.

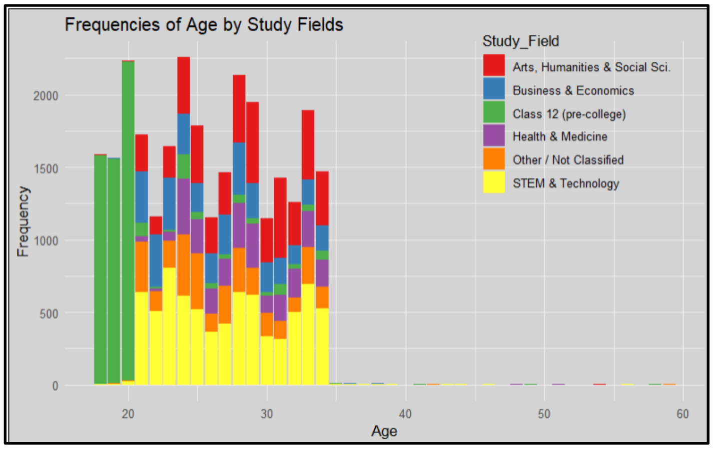


Figure 8a: Age distribution, stratified by field of study.

To further probe positivity, we also examined the distribution of CGPA, shown in Figure 8b. To reduce sparsity and focus on the main student population, we trimmed at CGPA ≥ 5, which removed a miniscule amount of observations with missing CGPA data. After trimming, most fields exhibit reasonable overlap, representing various study-fields across the 5–10 CGPA range, although certain study-field/CGPA combinations still have relatively few observations.These gaps are less structural than the Age/Class 12 case, but they underscore that positivity should be assessed not only theoretically but also empirically. Note: We trimmed observations with CGPA < 5 and Age > 35, treating them as anomalies or data mismatches, since these values fell outside of typical student ranges / grades and did not affect quartile definitions, due to their extremity.

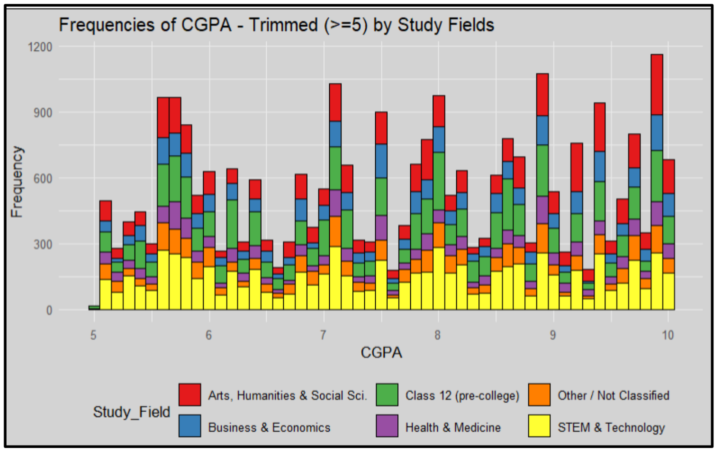


Figure 8b: Trimmed CGPA (>= 5) distribution, stratified by field of study

Propensity Score Models

Formally, the Propensity Score: $\pi(x) = P(A = 1|X = x)$ = the probability of receiving treatment given the covariates X. If positivity holds, it implies: $\pi(x) = P(A = 1|X = x) \in (0,1)$. This quantity is unknown and must be estimated from the data. We estimated it via a logistic regression model; appropriate for binary outcomes like treatment A, using X - the set of confounders we adjusted for, as predictors. This facilitated positivity checks in high dimension like ours and downstream causal effects estimation (i.e. ATE estimation using IPTW).

Specifically, we model: $\pi(x) = P(A = 1|X = x) = expit(\beta^T x) = \frac{e^{\underline{\beta}^T x}}{1+e^{\underline{\beta}^T x}}$, where:

$$\underline{\beta} = \left(\beta_0, \beta_{Study-Field}, \beta_{Age}, \beta_{CGPA}, \beta_{Satisf.}\right)$$, defined loosely for modeling flexibility purposes.

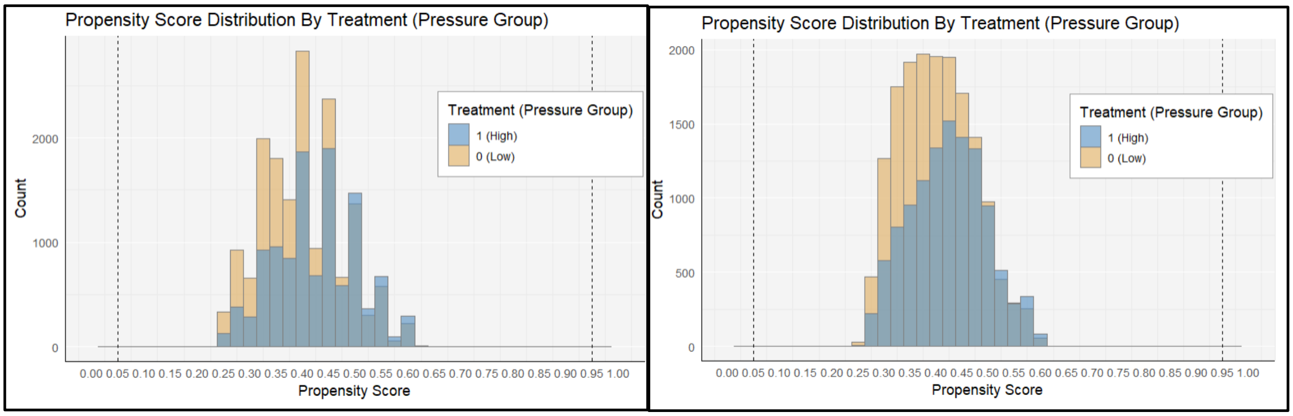
$$x = (1, Field\_of\_Study, Age\_Quartile, CGPA\_Quratile, Study\_Satisfaction\_Level) \Longrightarrow$$

$$\underline{\beta}^T x = \beta_0 + \beta_{Study-Field} \times Field\_of\_Study + \beta_{Age} \times Age\_Quartile + \beta_{CGPA} \times CGPA\_Quartile + \beta_{Satisf.} \times Study\_Satisfaction\_Level$$

We estimated the propensity score under **three different confounder specifications**:

1. **Quartile dummy variables for Age & CGPA**: (flexible, does not assume linearity, more parameters).
2. **Continuous Age & CGPA**: (fewer parameters, assumes linear effect on log-odds). Essentially using: $\underline{\beta}^T x^*, x^* = (1, Field\_of\_Study, Age, CGPA, Study\_Satisfaction\_Level)$.
3. **Numeric quartiles (1-4) for Age & CGPA**: (assumes linear effects across quartiles).

Across all specifications, the predicted propensity scores, shown in the histograms in Figure 9a & 9b below & appendix A1 for model (3), exhibit desirable properties, expanded below.



Figures 9a & 9b: Predicted Propensity Score distribution, stratified by treatment. Model (1) (left), Model (2) (right).

Empirical Overlap (Positivity): Across all model specifications, the two treatment groups exhibit substantial overlap in their PS distributions. For most values of $\hat{\pi}(x)$, we observe both treated and untreated students, supporting the empirical validity of the positivity assumption.

Absence of Extreme Values (Weight Stability): The predicted scores are contained within the [0.25, 0.6] range, with few observations at the tails. This mitigates the risk of large inverse-probability weights in methods like IPTW, which as noted in class, could inflate estimator variance.

Taken together, the agreement across all PS model specifications highlights the robustness of our conclusions, regardless of how age and CGPA are assumed to be modeled (processed dummy quartiles, raw - continuous without processing, or numeric quartiles). Given redundancy across specifications, we excluded Model (3) from further analysis.

Causal Effects Estimation (ATE = Average Treatment Effect)

We estimated the ATE, as well as a 95% confidence interval, using 3 methods.

1. Naïve ATE Estimation

The naïve ATE Estimation is defined as:

$$\widehat{ATE}_{Naive} = \frac{\sum_{i=1}^{N=27,901} Y_i \times A_i}{\sum_{i=1}^{N=27,901} A_i} - \frac{\sum_{i=1}^{N=27,901} Y_i \times (1 - A_i)}{\sum_{i=1}^{N=27,901} (1 - A_i)}$$

This can be directly estimated from the data with no need of additional modeling using high-dimensional X confounders. However, since it does not adjust for confounding, it may be biased in the presence of systematic differences between treatment groups - hence the need for adjusted estimators such as Standardization and IPTW, detailed below.

2. ATE Estimation using Standardization

The ATE estimation formula using Standardization is defined as:

$$\widehat{ATE}_{Standardization} = \frac{1}{N}\sum_{i=1}^{N=27,901} \hat{E}[Y_i|A = 1, X_i = x_i] - \frac{1}{N}\sum_{i=1}^{N=27,901} \hat{E}[Y_i|A = 0, X_i = x_i] =$$

$$\frac{1}{N}\sum_{i=1}^{N=27,901} \hat{P}(Y_i = 1|A = 1, X_i = x_i) - \frac{1}{N}\sum_{i=1}^{N=27,901} \hat{P}(Y_i = 1|A = 0, X_i = x_i)$$

Since our outcome (depression status) is binary, we fitted logistic regression models of the form $\boldsymbol{Depression \sim Treatment + Confounders}$, using variations of our minimal adjustment set. For each model, we predicted the potential outcomes, artificially setting $A = 1, A = 0$ for every student in the dataset, then averaged these predictions across the whole sample. The ATE Standardization estimate was defined as the difference between these two averages.
This approach adjusts for confounding on the chosen minimal set of covariates, because treatment and control predictions are made conditional on X. We used **three different covariate specifications** for the outcome model, all treating Study Satisfaction as continuous (1–5):

1. **Quartile dummy variables for Age & CGPA**: (flexible but potentially biased, does not assume linearity, more parameters).
2. **Continuous Age & CGPA** (fewer parameters, assumes linear effect on log-odds).
3. **Extension of Model (2) + Interactions between Treatment & Age / CGPA**: (more flexible, allows treatment effect heterogeneity by age & CGPA).

For each specification, we reported the ATE, variance, and 95% CI using nonparametric bootstrap ($B = 300$ for robust estimation).

3. Estimation using Inverse Probability of Treatment Weighting - IPTW (Stabilized Weights)

We also used IPTW (specifically the stabilized weights version to moderate our estimator's variance) to estimate the ATE. This estimator reweights each observation by a variant of the inverse of its probability of getting treatment given its covariates. The ATE estimator is defined as:

$$\widehat{ATE}_{IPTW}^{SW} = \frac{1}{N}\sum_{i=1}^{N=27,901}\frac{W_i^{SW} \times Y_i \times A_i}{W_i^{SW} \times A_i} - \frac{1}{N}\sum_{i=1}^{N=27,901}\frac{W_i^{SW} \times Y_i \times (1-A_i)}{W_i^{SW} \times (1-A_i)}$$

Where: $\forall 1 \leq i \leq N = 27,901.\, W_i^{SW} = \begin{cases} \frac{\hat{P}(A=1)}{\hat{\pi}(x_i)}, A_i = 1 \\ \frac{\hat{P}(A=0)}{1-\hat{\pi}(x_i)}, A_i = 0 \end{cases}$

Note: This formula requires estimating $\hat{P}(A=1), \hat{\pi}(x_i)$ for all $1 \leq i \leq N = 27,901$ beforehand! The latter was obtained from the (2) propensity score models described above.
For each specification, we reported the ATE, variance, and 95% CI using nonparametric bootstrap ($B = 300$ for robust estimation).

Results: Below are the results of the estimates (+ 95% CIs for each method):

Table 3 – Summary of ATE Estimates (Mean, SE & 95% CI) – using different methods

ATE		Mean ATE	Standard Error	95% CI
Naive		0.3911	0.0052	[0.3806, 0.4014]
Standardization	Model (1) - Quartile Variables	0.3635	0.0052	[0.3523, 0.3734]
	Model (2) - Original Variables	0.3652	0.0052	[0.3549, 0.3744]
	Model (3) - Original Variables + Interactions	0.3652	0.00555	[0.3543, 0.3755]
IPTW	PS Model (1)	0.3627	0.0055	[0.3531, 0.3748]
	PS Model (2)	0.3638	0.0051	[0.3529, 0.3719]

All methods produced similar ATE estimates, consistently positive around [0.35, 0.38], except the naive estimator with a CI between [0.38, 0.4] and is clearly biased upwards compared to other models. This agreement across the various estimation methods indicates robustness and increases confidence that the estimated effect is reliable and not driven by model-specific assumptions. More interestingly, none of the 95% confidence covered 0, providing statistical evidence that the causal effect of high academic pressure on depression is statistically significant at the 5% level. The magnitude of the estimated ATE - close to 0.36, implies that on average, exposure to high academic pressure increases the probability of experiencing depression by approximately **0.36**, a substantial effect in practical terms Moreover, the relatively narrow confidence intervals across methods & covariate specifications reflect high precision and further underscore estimated ATE's stability.

CATE Estimation

We set out to estimate the **Conditional Average Treatment Effect (CATE)** function of high academic pressure on depression, while allowing the effect to vary according to a set of effect modifiers upon which we condition. We used two common meta-learner strategies (S-Learner and T-Learner). Each meta-learner was paired with two different base learners:

1. **Logistic Regression** - parametric, easy to fit & interpret.
2. **Random Forest** - flexible, nonparametric, (implemented w. 500 trees via Ranger library).

For each meta-learner and base learner, we used 2 effect modifier specifications:

1. **Raw, Continuous Age and CGPA.**
2. **Processed, Quartile-based Age and CGPA.**

Other covariates suspected as effect modifiers are specified in the 'Effect Modifiers' section.

We wanted to focus on the results for continuous age, especially since quartiles, although can aid interpretability (by allowing results to be expressed in simple group comparisons), impose cutoffs & assume that treatment effects change in stepwise jumps between groups. This risks oversimplifying effects and masking heterogeneity in reality.

S-Learner Approach:

- Fitted a single outcome model of the form: $\boldsymbol{Depression \sim Treatment * Effect\ Modifiers,}$ essentially using treatment, modifiers, and their interactions as covariates.
- For each student, predicted potential outcomes under $\boldsymbol{A = 1}$ and $\boldsymbol{A = 0}$, "fixing" modifiers.
- Estimated **individual CATEs:** $\boldsymbol{\widehat{\tau(Z_\iota)} = E[Y|A \widehat{= 1,} Z = Z_\iota] - E[Y|A \widehat{= 0,} Z = Z_\iota],}$ reported key statistics - mean CATEs, standard errors, 95% CIs of the mean CATEs (we also stratified by covariate strata and computed those on the stratified data, with boxplots too).

T-Learner Approach:

- We fitted two models of the form: $\boldsymbol{Depression \sim Treatment + Effect\ Modifiers}$, each - based on the treated ($A = 1$) and untreated ($A = 0$)groups (i.e. $E[Y|A \widehat{= 1, Z} = Z_I]$,
- $E[Y|A \widehat{= 0, Z} = Z_1]$, for each base learner.
- Estimated **individual CATEs** $\widehat{\tau(Z_1)} = E[Y|A \widehat{= 1, Z} = Z_1] - E[Y|A \widehat{= 0, Z} = Z_1]$, reported key statistics - mean CATEs, standard errors and 95% CIs of the mean CATEs (we also stratified by covariate strata and computed those on the stratified data).

<u>CATE Estimation Results</u>:

For brevity, we did not report all stratified mean CATEs. Instead, to illustrate the methodology and enable comparison with the earlier estimation methods, we focused on mean CATE statistics across meta-learners and base-learner variants. We excluded the quartile-based effect modifier specification (except for Figure 12a), given its redundancy and the close similarity of results to those obtained using continuous age and CGPA.

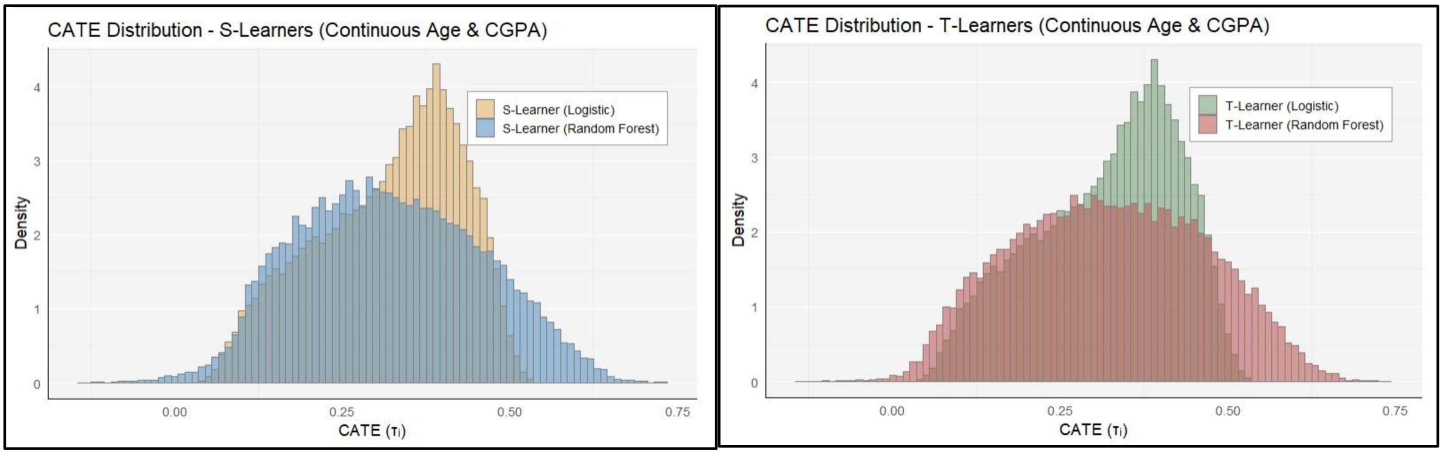
Table 4 – Summary of Mean CATE Estimates (+ SE & 95% CI) – using different methods

Meta-Learner	Base-Learner	Mean CATE	Standard Error	95% CI (of Mean CATE)
S-Learner	Logistic Regression	0.3192	0.0006	[0.318, 0.3205]
	Random Forest	0.3201	0.0008	[0.3185, 0.3217]
T-Learner	Logistic Regression	0.3192	0.0006	[0.318, 0.3205]
	Random Forest	0.3221	0.0008	[0.3204, 0.3237]

Table 4 shows that mean CATEs (practically equivalent to ATEs) were consistent across meta-learners and model variants. With the logistic regression base learner, results were identical across both meta-learners (~0.32), and similarly for Random Forest. This is slightly lower than the 0.35-0.40 range obtained via Standardization and IPTW. This difference likely reflects variations in covariate specification, inclusion of effect modifiers like sleep duration and financial stress, and, in the S-Learner, interaction terms. Importantly, as with the other estimation approaches, all four confidence intervals excluded zero, providing clear statistical evidence that the causal effect of high academic pressure on depression is significant at the 5% level. The stability of these estimates across modeling approaches reinforces the robustness of our findings and strengthens confidence that the observed effect is not an artifact of any single meta-learner specification.

To further investigate observed trends in Table 4, we visualized the CATE distributions across meta-learners and base learners (Figures 10a & 10b). The results show that Logistic Regression base learners yielded left skewed but more concentrated distributions around ~0.32, while Random

Forest–based learners produced wider, more dispersed distributions, extending further into both lower and higher values. It is worth noting the mean remained consistent across all 4 learners.



Figures 10a & 10b: CATE distribution by base learners. S-Learner (left), T-Learner (Right).

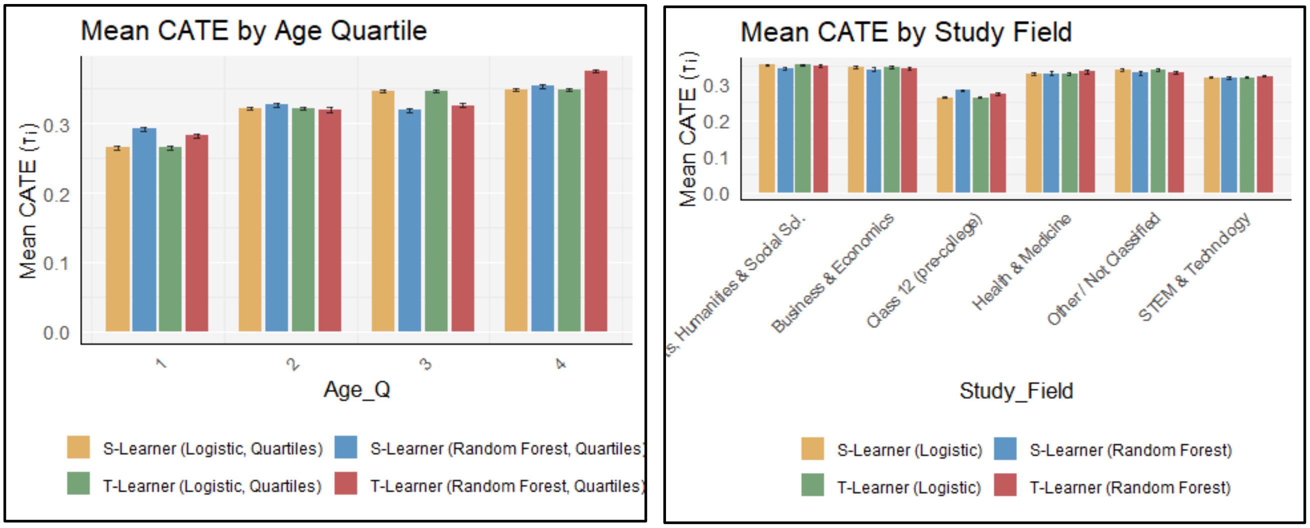
Given space constraints, we did not report all mean CATEs and associated statistics stratified by every effect modifier. Instead, we compared, focused and visualized three representative modifiers where heterogeneity was interpretable: **age quartile, study field, and financial stress**.

Age Quartiles (Figure 11a): Using quartile-based CATE models, we observed a gradual increase in mean CATE with age. For instance, mean effects rose from ~0.28 in Q1 (youngest) to ~0.38 in Q4 (oldest), suggesting that older students may be more vulnerable to academic pressure.

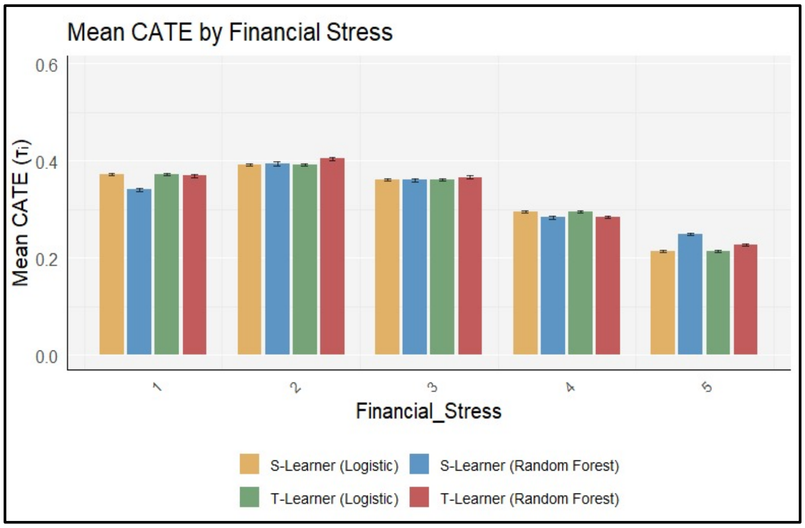
Study Field (Figure 11b): Some variation was visible across fields. Notably, Class 12 (pre-college) students had lower mean effects than other fields, across all learners. Other fields clustered more tightly around the overall mean (~0.32). Their results appeared very similar across all 4 models.

Financial Stress (Figure 11c): A decreasing gradient emerged: students with low-to-moderate financial stress displayed higher mean CATEs than other groups, whereas those with the highest and lowest stress reported substantially lower effects (~0.2-0.25).

Taken together, these results show that while the overall effect of academic pressure on depression is consistently positive and centered around **~0.32**, heterogeneity exists. Clear signals of variation arise stratifying by **age quartile, study field, and financial stress**, highlighting that certain subgroups may experience stronger or weaker treatment effects. Graphs visualizing the CATE by raw CGPA & Age can be found in the appendix A3 with similar means to the quartile-based models.



Figures 11a, 11b & 11c: Mean CATEs + corresponding 95% CIs across all 4 meta learners & base learners. By: Age quartile (left), Study Field (right), Financial Stress (below).



Additional Theoretical Questions

A. Potential Unobserved Confounder

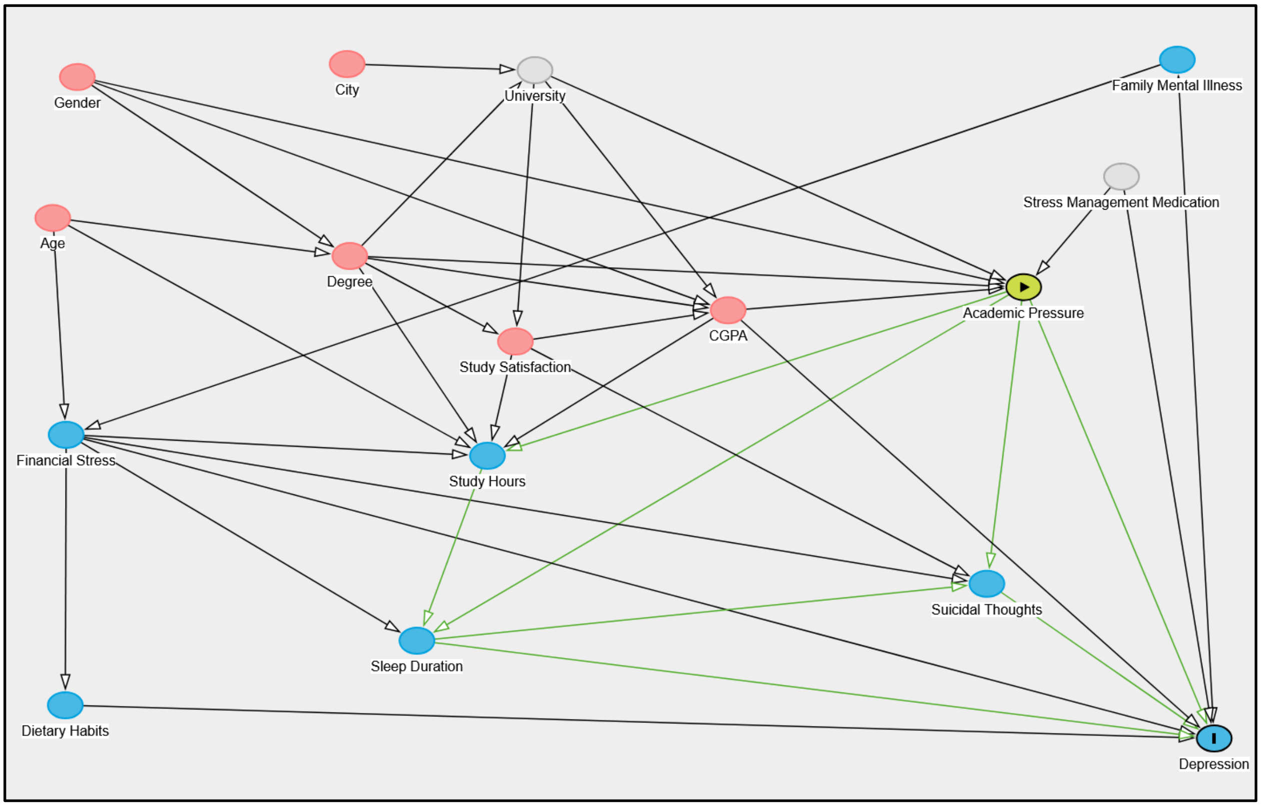


Figure 12: Our proposed DAG, with a hypothetical unobserved confounder - "Stress Management Medication".

In this DAG we have included **"Stress Management Medication"** as an **unobserved confounder** that is a parent of both the treatment "Academic Pressure" and the outcome "Depression". This confounder is categorical and indicates which type of stress management medication if any an individual takes certain types of stress medication can also be used to treat depression symptoms and thus usage of this medication may inhibit depression symptoms and change the result of the depression diagnosis. This unobserved confounder opens a backdoor path between the treatment and the outcome hence had it been measured we would have to include it as a confounding variable.

B. Potential Instrumental Variable (IV)

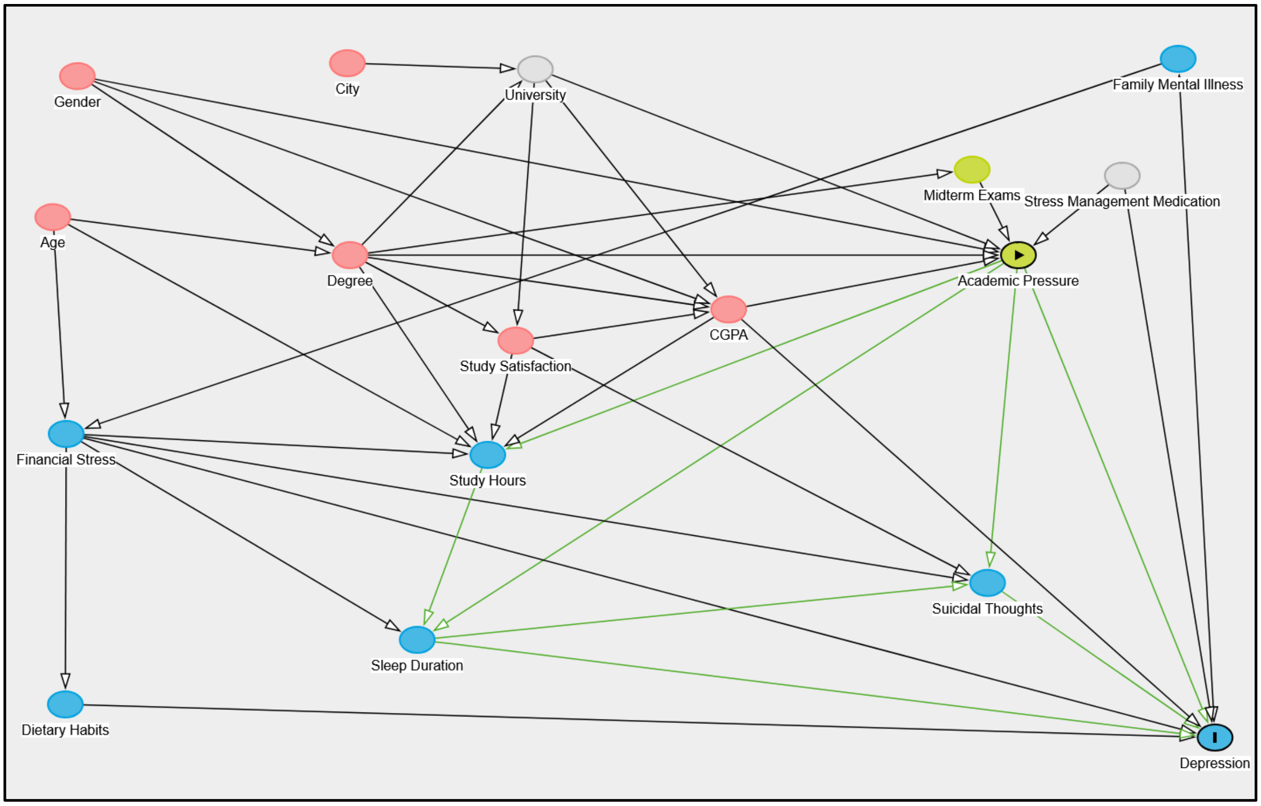


Figure 13: Our proposed DAG, with a hypothetical instrumental variable - "Midterm Exams".

In this DAG we added the **instrumental variable "Midterm Exams"** indicating whether or not individuals have at least 1 midterm exam this semester. This depends on the degree as certain degrees might have more courses with midterm exams than others. This variable satisfies the definition of causal IV: (i) Relevance: "Midterm Exams" affects "Academic Pressure" (treatment). (ii) Exclusion restriction: "Midterm Exams" affects "Depression" only through "Academic Pressure", there are no other directed paths between the two. (iii) Independence: all undirected paths from "Midterm Exams" to "Depression" are blocked, any undirected path involving the treatment will have an unadjusted collider and any backdoor path will be blocked by "Degree" as a confounder. This causal IV can be used to identify the causal effect if the unobserved confounder from the previous section had existed.

C. Potential Negative Control Outcome (NCO)

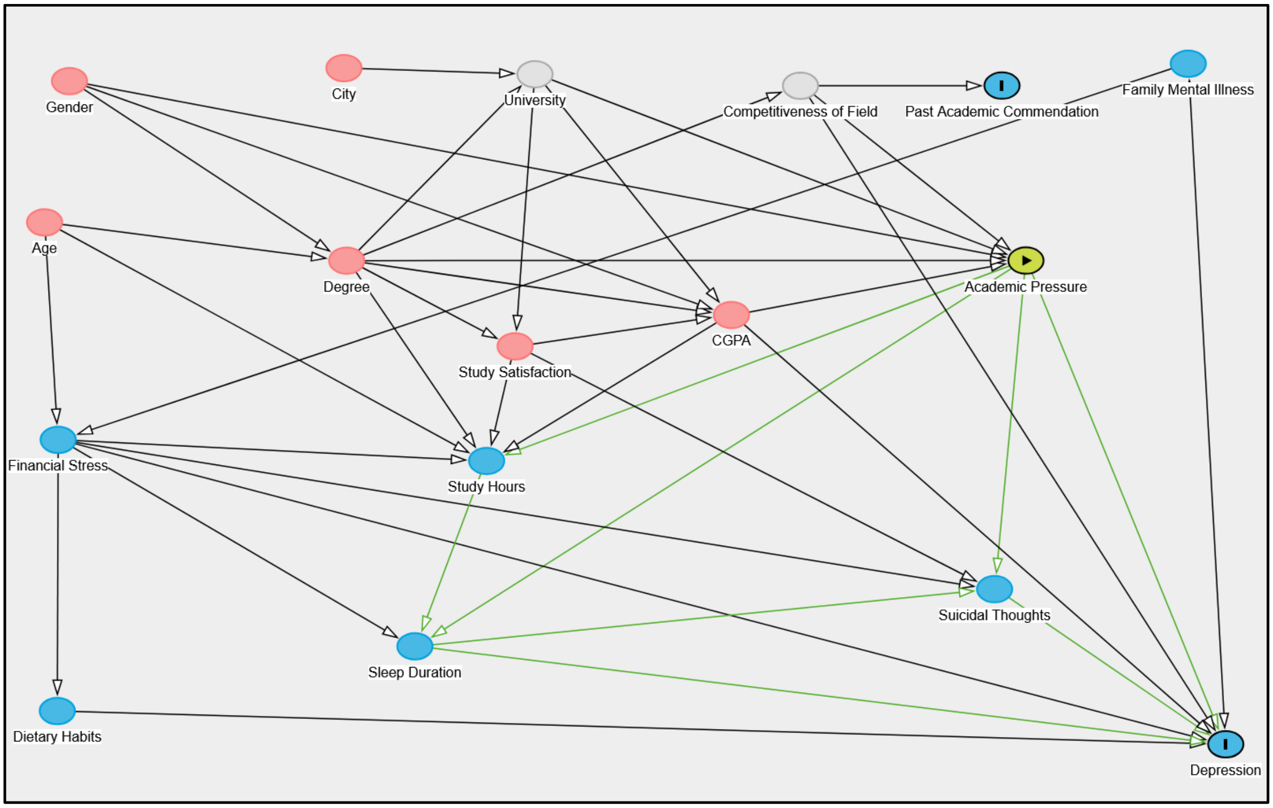


Figure 14: Our proposed DAG, with a hypothetical negative control outcome - "Past Academic Commendation".

We suggest **"Past Academic Commendation"** for a possible **NCO**, if it had been measured it could have been used to test unmeasured confoundness caused by "Competitiveness of Field". This unmeasured confounder indicates how much competition is experienced in one's field of study, hence we believe that it directly relates to "Degree", "Academic Pressure", "Past Academic Commendation" and to "Depression" because competition may create a hostile environment affecting one's mental health directly. "Past Academic Commendation" is an NCO according to the definition: (i) "Past Academic Commendation" is not a descendant of "Academic Pressure" (treatment). (ii) "Academic Pressure" is independent of "Past Academic Commendation" given the set of confounders we adjust on and "Competitiveness of Field". (iii) "Past Academic Commendation" is not independent of "Competitiveness of Field" given our chosen set of confounders because it is a child of the unmeasured confounder.

D. Potential Negative Control Exposure (NCE)

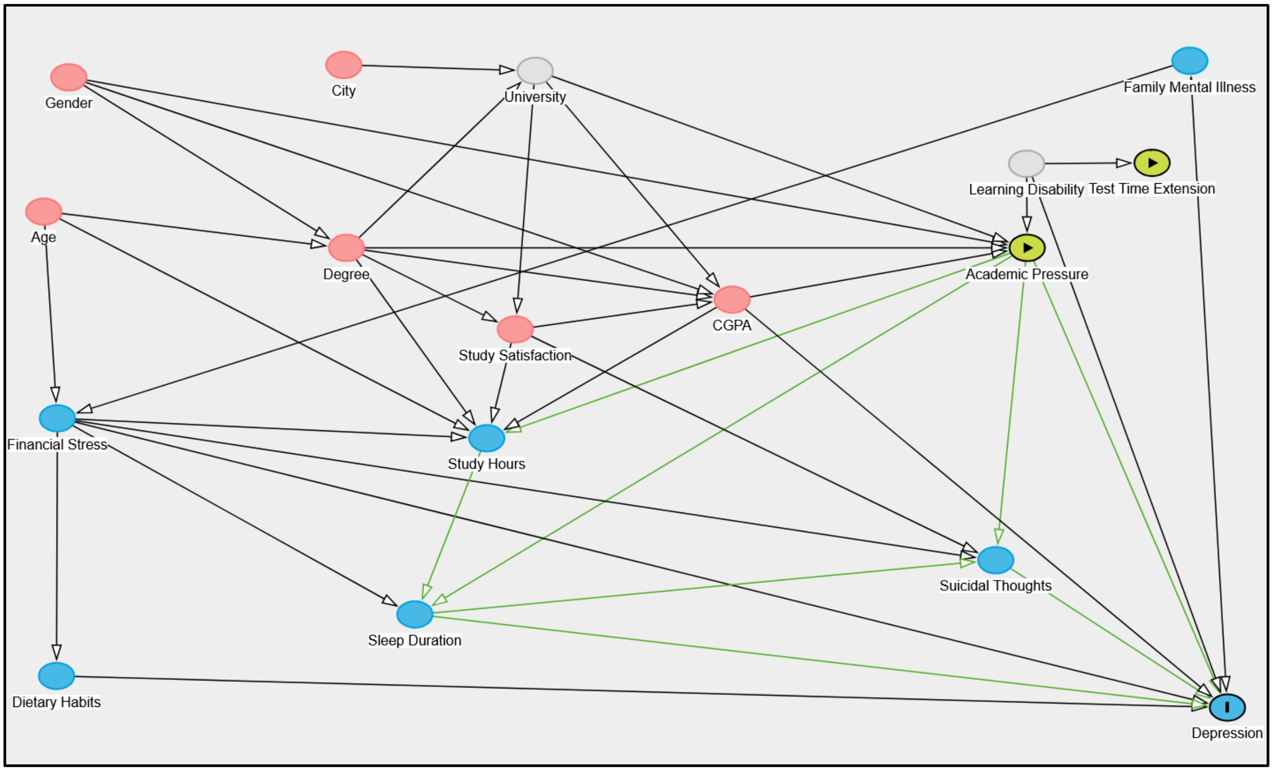


Figure 15: Our proposed DAG, with a hypothetical negative control exposure - “Test Time Extension”.

Our hypothetical **NCE** is **“Test Time Extension”** , a variable for whether or not one is given time extensions during exams in their institution. If it had been measured we could have tested for unmeasured confoundness caused by ״Learning Disability”, because it matches the definition for NCE: (i) “Depression” is not a descendant of “Test Time Extension”. (ii) “Academic Pressure” and “Depression” are d-separated by “Learning Disability” and the confounder set we chose. (iii) “Test Time Extension” is not independent of “Learning Disability” given the confounders and “Academic Pressure” (treatment) because “Learning Disability” is a parent of “Test Time Extension”. “Learning Disability” is a variable indicating if a student has a learning disability, this can directly affect both whether or not they have a time extension and how much academic pressure they experience, it can also directly affect depression status because of possible frustration in dealing with the disability in an academic setting and potential alienation from peers.

E. Mediators

Yes, as can be seen in our original DAG **several mediators exist** in our data; "Study Hours", "Sleep Duration" and "Suicidal Thoughts". Since we have multiple mediators in our DAG, satisfaction of the identification assumptions for the CDE, NDE and NIE is difficult to assess. For the CDE assuming positivity and assuming no interference and no multiple versions of levels for each mediator (same quality of sleep and an hour of study has the same effectiveness for every one for example), assuming SUTVA in other words, the other assumptions (conditional exchangeability of the outcome and mediators and outcome and treatment) relating to the DAG are satisfied by considering $X = \{Age, CGPA, Study\ Satisfaction, Degree, Financial\ Stress\}$, this set satisfies the back-door criterion for the treatment and when combined with the treatment, satisfies the back-door criterion for all of the mediators jointly. For the NDE and NIE, maintaining SUTVA and positivity assumptions as well, there is still cond. Exchangeability between the treatment and the outcome and mediators and outcome using the X set above, and the assumption of cond. exchangeability between the mediators and treatment given X is also satisfied. That leaves the last assumption $Y^{a,m} \perp M^{a'}|X$ which is a cross world assumption and cannot be verified unless under specific conditions under SPSEM-IE which is beyond the scope of this paper. We do not know whether or not these assumptions are likely to hold.

F. Selection bias and collider bias

We **do not have selection or collider bias** in our data, since our data is synthetic and simulates a questionnaire answered by randomly chosen students there is no selection bias, as for collider bias, under our current DAG we do not adjust for any post-treatment colliders so there is no collider bias either. For our data to include selection bias theoretically our data would represent questionnaire answers given after the semester has finished yet still asking questions about its beginning and middle, but the questionnaire would only be taken by students who suspect they have depression before taking the survey, this will induce post-outcome selection bias to the data.

Summary – Discussion and Conclusions

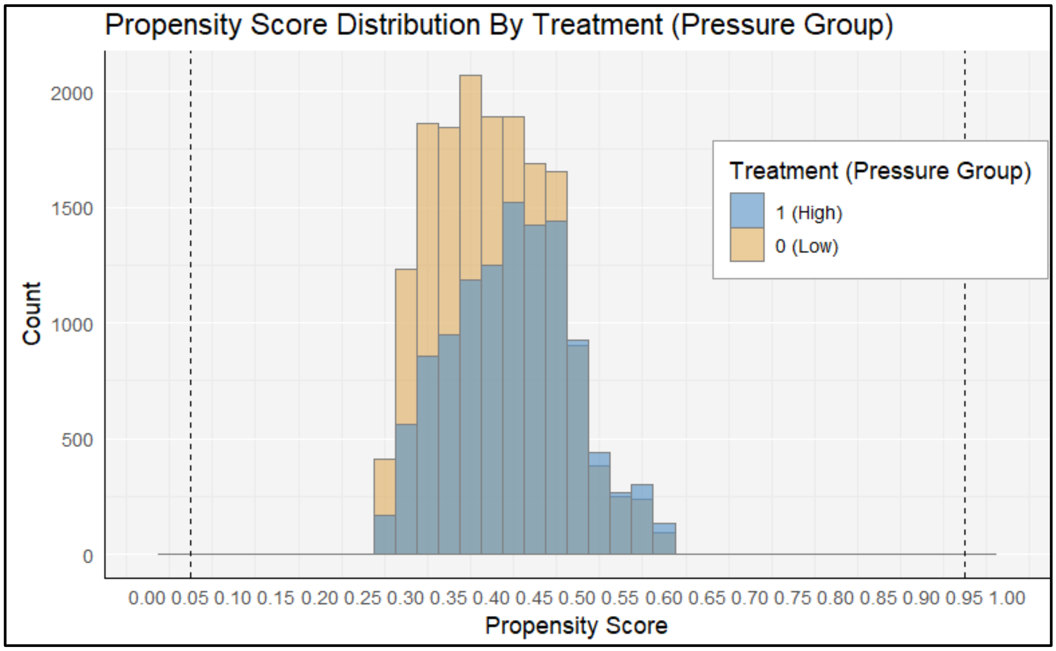
The results of the study strongly suggest that academic pressure exerts meaningful causal influence on students' mental health. Across all estimators: naïve comparison, standardization, and IPTW, we found consistent results. This fits with our hypothesis that psychological academic stress has a factor in causing depression. In addition, we observe that the effect may differ across groups, such as age categories and study fields.

Notwithstanding this, we would like to mention several important limitations that should be taken into consideration in order to treat our conclusions more carefully:

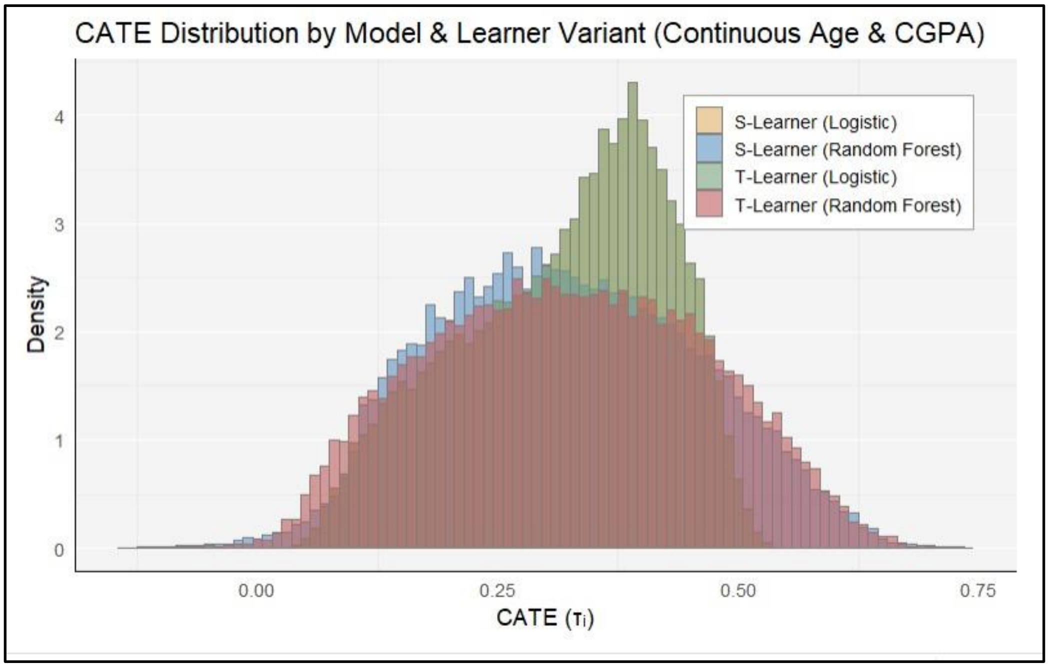
1. **Synthetic Dataset**: As mentioned earlier, the dataset is synthetic. Even though it appears realistic based on our observations, there is a possibility that it does not accurately reflect reality. More complex relationships between confounders may exist in real-world settings that the dataset does not capture. Therefore, future research should also focus on gathering real-world student data, using large student populations across diverse cultural contexts.
2. **Measurement Error**: Academic pressure and depression were self-reported. This type of data may carry the risk of additional biases, which must be treated carefully.
3. **Binary Treatment**: Reducing academic pressure to a binary indicator may have masked meaningful variations between the levels of the original covariate. For example, it is possible that only students under extremely high pressure are at elevated risk of depression, while moderate pressure could be neutral or even beneficial. Likewise, very low pressure might indicate disengagement or lack of motivation, which could also be harmful. Such potential non-linear relationships are lost when collapsing the 1–5 scale into two categories. In future work, it can be interesting to use a three-layer split (e.g., low, medium, and high), or even to separately compare extremely high versus the rest, or extremely low versus the rest, in order to better capture non-linear effects.
4. **Correctness of DAG**: All of our results are interpreted under the assumption that our DAG is correct, should it be false our results will include bias that may significantly affect the results and mask the true causal effect. We did not consult with expert psychologists when constructing our model of reality, hence our DAG should be questioned by further research done in this area to perhaps build a better model for reality.

Taken together, our study demonstrates a consistent and statistically significant causal effect of high academic pressure on depression, with estimated increases in the probability of experiencing depression of roughly 0.3-0.4. In practical terms, this means the observed effect is unlikely to be due to chance, and that high academic pressure substantially raises students' risk of depression. The effect is heterogeneous, varying across age, study field, financial stress, and more, highlighting that certain student groups are especially vulnerable to academic pressure. Methodologically, our work shows how DAG-guided confounder selection, combined with standardization and IPTW under multiple specifications, can yield robust causal evidence in complex educational data, under the standard causal identification assumptions of exchangeability, positivity, and SUTVA. At the same time, important limitations remain - including the use of a synthetic dataset, self-reported measures, the reduction of academic pressure to a binary treatment, and the fact that our 27,000+ students all come from India, which may limit cultural generalizability. These considerations call for cautious interpretation of our results.

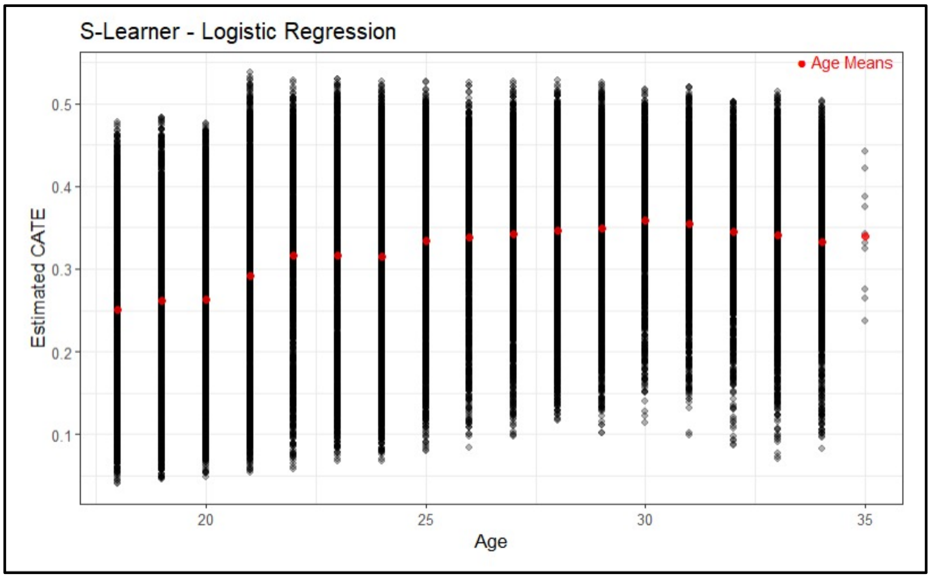
Appendices



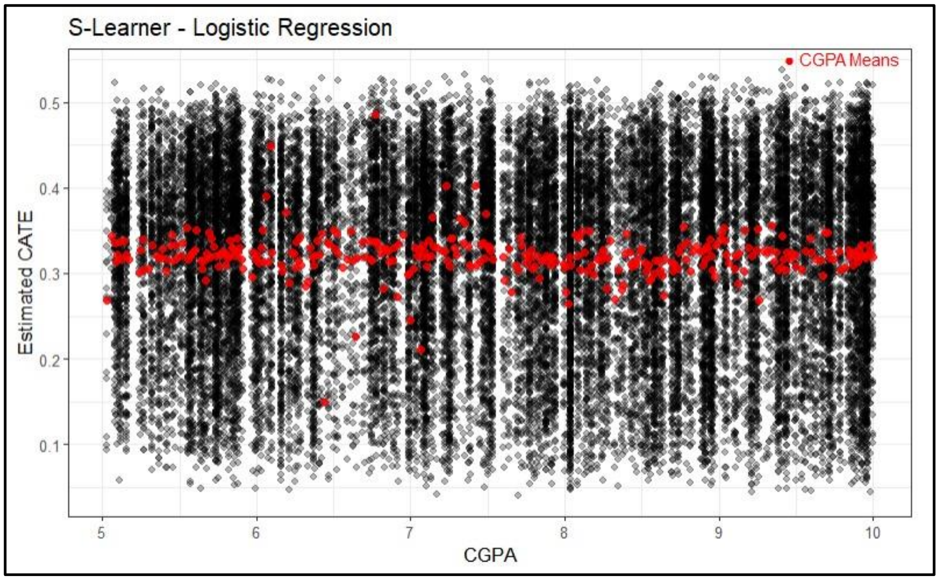
Appendix A1: Predicted Propensity Score distribution, stratified by treatment. Model (3) (as specified in the Propensity Score section).



Appendix A2: Predicted CATE distribution, stratified by meta-learner & base-learner.



Appendix A3a: Predicted CATE scatterplots, by Age (with red means). Logistic Regression - S-Learner.



Appendix A3b: Predicted CATE scatterplots, by CGPA (with red means). Logistic Regression - S-Learner.