The Effect of Design Choices in Data Preprocessing on Estimated Treatment Effect: A Case Study on Academic Success

Liam Brinker 213215205 Yarden Adi 212585848

March 5, 2025

Abstract

Causal inference studies often rely on assumptions embedded in data preprocessing, yet the impact of these design choices on Estimated Treatment Effects (ETE) remains underexplored. Decisions such as defining treatment thresholds, categorizing outcomes, and selecting estimation methods can significantly influence causal estimates, thereby affecting the validity of conclusions. In our project, we assess how the ETE varies as a function of different preprocessing design choices. We explore this through the causal question: *What is the effect of enrolling as an adult student (age x or older) on academic success* Using the Predict Dropout or Academic Success dataset from Portugal, we analyze how variations in the definition of adulthood, alternative classifications of graduation status, and different causal inference models impact the Average Treatment Effect (ATE). This project serves as a proof of concept for the broader issue of how preprocessing choices influence causal effect estimation. By quantifying the sensitivity of ATE estimates to different design choices, we emphasize the importance of methodological transparency and rigor in causal inference. Our findings provide insights into best practices for preprocessing in causal studies and highlight the need to assess the robustness of conclusions across multiple design choices. Project resources are available in this GitHub repository.

1 Data Review and Preprocessing

To investigate the relationship between age at enrollment and academic success, we needed data that captures students' academic progress and personal characteristics at the time of enrollment. This section details the dataset we used and the preprocessing steps taken to prepare it for causal analysis.

1.1 Data Review

The data used in our project originated from research done in Portugal by the Polytechnic Institute of Portalegre (higher education institute that focuses on technology), as an attempt to provide information to the tutoring team about the risk of students' dropout and failure (Realinho et al., 2022). Commonly, it is used to build machine learning models for predicting academic performance and dropout (see relevant Kaggle competitions in this subject).

The data contains information about students' pre-academic background, age, academic performance, social and economic status, and other relevant variables. It consists of 4424 records and 37 variables, including the enrollment age. The dataset includes a trinary outcome variable indicating whether a student dropped out, graduated, or is still enrolled in the academic program after its predetermined period.

The data was created by joining three primary data sources:

1. CNAES (National Competition for Access to Higher Education) - contains information about students' academic backgrounds, demographics, and course applications at the time of their enrollment in Portuguese higher education institutions.
2. AMS (Academic Management System) - provides student records data, including demographic information, course enrollments, and academic performance throughout their studies.
3. PORDATA (Contemporary Portugal Database) - provides macroeconomic data, including unemployment, inflation, and GDP figures.

Traditionally, each attribute used in the dataset is associated with one of the following classes: demographic, socioeconomic, macroeconomic, academic data at enrollment, and academic data at the end of the first and second semesters.

1.2 Data Preprocessing

1.2.1 Design Choices for Treatment and Outcome Variables

Treatment Variable: Defining Adulthood Thresholds In line with our causal question, we introduce a binary treatment variable: "enrolling as an adult student." Recognizing that the definition of adulthood can vary, we consider multiple age thresholds to assess their impact on graduation outcomes:

- **Age 20:** This threshold captures students who may have taken a brief period after secondary education before enrolling in higher education.
- **Age 21:** Aligning with common definitions of non-traditional students, this age reflects individuals who delayed enrollment to pursue other opportunities or due to personal circumstances.
- **Age 23:** This threshold includes students who may have engaged in extended activities such as work, travel, or family commitments before commencing their academic programs.

For each threshold, we preprocess the data by creating a binary indicator: students aged greater than or equal to the specified threshold at enrollment are labeled as adult students (treatment = 1), while those younger are considered traditional students (treatment = 0). The original numerical age variable is excluded from subsequent analyses to focus on the binary treatment effect. Figure 1 in Appendix A presents the distribution of the treatment variable in the raw data and across the different adulthood thresholds considered.

Outcome Variable: Defining Academic Success Our analysis defines success as the completion of an academic program and focuses on estimating its probability. To explore how different definitions of success influence our findings, we preprocess the outcome variable in two distinct ways:

1. Strict Success Definition:
   - Successful: Students who have completed their academic program within the expected duration.
   - Unsuccessful: Students who are still enrolled to their academic program beyond the standard timeframe or have dropped out.

   In this approach, the outcome variable is binary, with 1 indicating successful completion and 0 representing non-completion.
2. Inclusive Success Definition:
   - Successful: Students who have either graduated from their academic program within the expected timeframe or are still enrolled.
   - Unsuccessful: Students who have discontinued their studies in the academic program.

   Here, the binary outcome assigns 1 to successful students and 0 to those who have dropped out.

Figure 2 in Appendix A presents the distribution of the outcome variable in the raw data and under both strict and inclusive definitions.

Combining Treatment and Outcome Definitions To comprehensively assess how preprocessing choices influence the ATE, we analyze all possible combinations of treatment and outcome definitions. This involves pairing each adulthood threshold (ages 20, 21, and 23) with both the strict and inclusive success definitions, resulting in six distinct analytical scenarios.

1.2.2 General Design Choices Applied to All Configurations

Apart from the design choices that vary across configurations, there are several fixed design choices applied to all configurations. These choices are intended to reduce the number of features and simplify the dataset, making it more suitable for applying causal inference methods. We created a processed dataset for each treatment and outcome configuration.

Removal of post-treatment variables The dataset contains multiple variables gathered throughout the student's academic journey, e.g., academic accomplishments, debts, and payment tracking. Despite being informative for machine learning models trying to predict academic dropout, these accomplishments were recorded after the treatment was determined and, therefore, cannot be safely used in the causal inference procedure. We eliminated those variables to maintain the integrity of our results and prevent any post-treatment interference.

Removal of treatment-correlated features During the one-hot encoding of categorical variables, we deliberately excluded the dummy variable representing "Application mode 39 - Over 23 years old" since this feature is strongly correlated with our treatment variable (being a certain age or older at enrollment: 20, 21, or 23). Including this dummy variable could have introduced redundancy and potentially biased our analysis.

Clustering categorical values The dataset is of impressive complexity and detail, as evident by the categorical variables with over 30 unique values. To perform meaningful analysis, we manually cluster similar categories into one broader category to represent them, instead of simply removing the less frequent value. For example, we merged the values "Armed Forces Professions", "Armed Forces Officers", and "Armed Forces Sergeants" into a single "Armed Forces" category, which functions as a single value in the analysis. We performed this procedure on five variables within the dataset - the qualifications and occupations of the parents (mother and father) and the student's previous qualifications.

Pruning categorical outliers Even after performing the previous step, some variables still contained rare values. After careful consideration and visual inspections, we decided to prune some of the rare values to ease the analysis. The pruning resulted in a sample exclusively consisting of individuals with Portuguese nationality, which may limit the generalizability of this study to a broader population.

After the preprocessing stage, each configuration's dataset comprises 4249 records, each containing 21 variables, including the treatment and outcome variables. Five of these variables are numerical, and the rest are categorical.

2 Assumptions for Causal Inference

In this section, we formally present and discuss four assumptions regarding the nature of our data. Combined, these assumptions guarantee the trustworthiness of an observational causal experiment's results.

2.1 Stable Unit Treatment Value Assumption (SUTVA)

The SUTVA assumption consists of two parts. The first one is *no interference*, which requires that the potential outcomes of each unit are not affected by the treatment assignment of any other unit; the second is *no hidden variations of treatment*, which forbids the existence of different forms or versions of each treatment level (within the same configuration), which lead to different potential outcomes.

In our experiment, the treatment variable, "enrolling as an adult student," is well-defined for each configuration (age thresholds: 20, 21, or 23) and has only two versions. Hence, the second part of the assumption safely holds. The assumption's first part holds if we presume one student's adulthood does not affect the probability of fellow students graduating from the program, which is controversial. There are several aspects in which the varied spectrum of ages on campus might affect students' success. Such elements include but are not limited to 1) peer effects - encountering different perspectives and learning methods (that can be related to one's age) might affect one's abilities; and 2) In competitive programs, students' success may be directly affected by their peers' accomplishments. Therefore, if age is indeed related to academic achievements, one student's age can affect other students' success.

2.2 Consistency

The Consistency assumption states that an individual's potential outcome under their observed exposure history is the outcome that would actually be observed for that person. Formally, for a unit that receives treatment T, we observe the corresponding potential outcome $Y = TY_1 + (1 - T)Y_0$. We believe this assumption holds for all treatment and outcome configurations in our study for several reasons:

1. **Data Source Reliability:** Our data comes from three well-established institutional databases (CNAES, AMS, and PORDATA), each with standardized data collection procedures and quality control measures.
2. **Clear Treatment Definition:** The treatment (being above a certain age at enrollment: 20, 21, or 23) is precisely defined and measured without ambiguity. Age at enrollment is an objective measure that cannot be misinterpreted.
3. **Outcome Measurement:** The graduation outcomes (strict or inclusive definitions) are documented in the academic records system (AMS) and follow standardized institutional definitions.

4. **Data Processing Transparency:** All our data preprocessing steps are well-documented and reproducible, ensuring the transformation from raw data to analysis variables maintains the integrity of the treatment and outcome measurements.
5. **Stable Treatment:** Age at enrollment is a stable characteristic that cannot change retrospectively, ensuring that the treatment status remains consistent throughout the study period.

2.3 Ignorability - No Unmeasured Confounders

The Ignorability assumption states that the treatment assignment T is independent of the potential outcomes Y_0, Y_1 given the observed covariates X, i.e., $Y_0, Y_1 \perp T|X$. This assumption is essential for ensuring that the estimated treatment effect is unbiased and not confounded by factors affecting both the treatment and the outcome.

However, this assumption is inherently unverifiable in practice. As noted by Hernán and Robins (2006), we cannot account for unmeasured confounders, as we do not observe them. Despite this limitation, correlations between observed covariates and unmeasured confounders can reduce bias associated with missing information (Schulz et al., 2023). If our dataset contains variables from diverse domains possibly affecting the treatment or potential outcomes, we may mitigate the effect of unmeasured confounders.

Based on Alyahyan and Düştegör (2020), we identify several potential confounder groups that may affect both age of enrollment (20, 21, or 23) and the probability of academic success: pre-academic performance, student demographics, student environment, psychological factors, and macroeconomic indicators.

Our dataset contains most relevant confounders, including pre-academic performance, student demographics, student environment, and macroeconomic indicators. However, we lack psychological factors, which could be a significant limitation in our analysis, as they may influence both age at enrollment and academic success. Academic progression is also an essential confounder, but being a post-treatment variable, it cannot be used without risking post-treatment bias.

2.4 Common Support (Overlap)

The Common Support assumption states that each unit has a non-zero probability of receiving each treatment level, i.e., $\forall x \in X, P(T = 1|X = x) > 0$ and $P(T = 0|X = x) > 0$.

We empirically validate this assumption for each treatment configuration (adulthood thresholds: 20, 21, 23) using propensity scores. For each configuration, we trained logistic regression models to predict the treatment assignment based on covariates, using predicted probabilities as propensity scores. We plotted the propensity scores of treated and control groups, ensuring significant overlap in each case. Results are summarized in Table 1 in Appendix B.

Based on these empirical validations summarized in Table 1 and visualized in Figure 3 (both are in Appendix B), we conclude that the common support assumption holds across all treatment configurations.

3 Causal Analysis Methodology

After validating the assumptions for causal inference, we describe the methodology used to perform the causal analysis. We present the relevant measures of causal effects and the methods we use to estimate them.

3.1 Measures

Average Treatment Effect (ATE) To quantify the causal effect of the treatment on the outcome, we estimate the ATE, which measures the overall impact of the treatment across the entire population. The ATE is defined as the difference between the expected outcome under treatment and the expected outcome under control: $ATE = E[Y_1 - Y_0]$. This measure provides a comprehensive understanding of the causal effect, capturing its impact on both treated and untreated individuals.

Bootstrap Confidence Intervals To assess the uncertainty in our estimated treatment effect, we employed bootstrap resampling to calculate confidence intervals for the ATE. We perform 1000 bootstrap iterations by default, where in each iteration, we resample the entire dataset with replacement, maintaining the original sample size. We then apply our causal inference methods to this resampled dataset, computing the ATE. After collecting these bootstrap estimates, we calculate the 95% confidence intervals using the percentile method, taking the 2.5th and 97.5th percentiles of the bootstrap distribution.

3.2 Methods

The fundamental problem of causal inference is that directly observing causal effects is impossible. For any given individual, we can only observe the outcome under one treatment condition—either treated or untreated. We never observe both potential outcomes for the same unit simultaneously, making it impossible to directly calculate individual treatment effects. In the remainder of this section, we present the estimation methods we used in our analysis, explain how they address this problem, present the assumptions required for estimating causal effects, and detail the implementation of each method.

3.2.1 Covariate Adjustment

Covariate adjustment methods use statistical models to regress the missing potential outcomes based on the observed covariates and the treatment. The regressed values are then used to estimate the ATE. We applied these methods across all treatment and outcome configurations.

S-Learner The S-Learner approach treats the treatment variable T as one of the covariates. First, a statistical model predicts the conditional expected outcome as a function of both the covariates X and the treatment T, i.e., $\hat{\mu}(X,T) = \hat{\mathbb{E}}[Y^{obs}|X,T]$. Any machine learning method can be used for this purpose, but importantly, the entire population is used to train the model. The ATE is estimated as follows:

$$\widehat{ATE}^{SL} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}(x_i,1) - \hat{\mu}(x_i,0)\right).$$

Hahn (1998) demonstrated that the S-Learner is consistent under the assumptions made in Section 2, provided that the model $\mu(X,T)$ is correctly specified. However, it may be prone to model misspecification, as noted by Rubin (1979). If the relationship between the covariates and the outcome is incorrectly modeled, it can lead to biased estimates of the ATE.

T-Learner To mitigate potential model misspecification issues in the S-Learner, we also employ the T-Learner. The T-Learner trains separate models for the treatment and control groups, using observations in the treatment group to estimate the response under treatment, $\hat{\mu}_1(x) = \hat{\mathbb{E}}[Y^{obs}|X,1]$, and observations in the control group to estimate the response under control, $\hat{\mu}_0(x) = \hat{\mathbb{E}}[Y^{obs}|X,0]$. The ATE is then estimated as:

$$\widehat{ATE}^{TL} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_1(x_i) - \hat{\mu}_0(x_i)\right).$$

Greenland and Robins (1986) showed that the T-Learner is consistent under the assumptions made in Section 2 and is more robust against misspecification of the outcome model compared to the S-Learner.

Implementation Details Before applying our models, we standardized all numerical features using StandardScaler and one-hot encoded categorical variables. We utilized Gradient Boosting classifiers for both the S-Learner and T-Learner approaches. This choice was made because Gradient Boosting can capture complex patterns in the data and handle interactions between variables automatically. For the S-Learner, we trained a single Gradient Boosting model on the entire dataset, including the covariates and the treatment variable. For the T-Learner, we trained separate Gradient Boosting models for the treated and control groups.

3.2.2 Propensity-Based Methods

To address potential model specification issues in covariate adjustment methods, we also employed propensity score methods, which adjust for confounding using estimated propensity scores rather than direct covariate adjustments.

Inverse Probability Weighting (IPW) This method estimates the probability of receiving treatment based on covariates and uses these probabilities as weights to balance the treated and untreated groups. Our analysis employs the Horvitz–Thompson estimator (Horvitz and Thompson, 1952). Intuitively, it estimates the outcome variable's mean in the treatment and control groups by weighting the observations based on their likelihood of treatment

assignment. First, a classification model is trained to predict the propensity score $\hat{e}(x) = \hat{P}(T = 1|X = x)$. The estimated propensity scores are then used to weigh each sample, allowing for the estimation of the ATE:

$$\widehat{ATE}^{IPW} = \frac{1}{n}\sum_{i=1}^{n}\frac{t_i y_i}{\hat{e}(x_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-t_i)y_i}{1-\hat{e}(x_i)}.$$

Propensity Score Matching (PSM) This method pairs treated units with control units that have similar propensity scores, thereby creating a matched sample where the distribution of observed baseline covariates is more balanced between treated and untreated subjects (Austin, 2011). Our implementation uses a nearest-neighbor matching approach with replacement, allowing multiple matches per unit. For each unit i, we identify the n closest matches based on propensity scores. Let $\mathcal{M}_i$ denote the set of indices of the n nearest neighbors for unit i in the opposite treatment group. We calculate weights w_{ij} for each match $j \in \mathcal{M}_i$ inversely proportional to the distance in propensity scores:

$$w_{ij} = \frac{1/d_{ij}}{\sum_{k\in\mathcal{M}_i} 1/d_{ik}},$$

where d_{ij} is the absolute difference in propensity scores between units i and j. The individual treatment effect for unit i is then computed as:

$$\tau_i = \begin{cases} y_i - \sum_{j\in\mathcal{M}_i} w_{ij} y_j & \text{if } t_i = 1 \\ \sum_{j\in\mathcal{M}_i} w_{ij} y_j - y_i & \text{if } t_i = 0 \end{cases}.$$

Following Basu et al. (2023), the ATE is estimated as:

$$\widehat{ATE}^{PSM} = \frac{1}{N}\sum_{i=1}^{N}\tau_i,$$

where N is the total number of units and τ_i represents the estimated individual treatment effect obtained from matched pairs.

Implementation Details The implementation utilizes the same scaled and encoded features as in the covariate adjustment methods, with a Random Forest classifier estimating the propensity scores. This choice was made because Random Forest captures complex relationships between covariates and treatment assignment and provides stable probability estimates, making it well-suited for propensity score estimation. The IPW method applies these propensity scores to weigh the observations. To prevent issues with very small denominators, the implementation includes safeguards that clip propensity scores to a range of $[10^{-5}, 1 - 10^{-5}]$. The PSM implementation adopts a nearest-neighbor approach with replacement, allowing multiple matches per unit. Using 11 nearest neighbors per unit balances bias reduction and variance. The matching process incorporates weights inversely proportional to the propensity score distances, assigning greater importance to closer matches. This strategy leverages more information from the data than one-to-one matching, potentially yielding more precise estimates while maintaining effective bias reduction from confounding variables.

3.2.3 Doubly-Robust Method

The doubly robust method combines both outcome regression and propensity score weighting, providing a consistent estimator for the ATE even if either the propensity score model or the outcome regression model is misspecified, as long as one of the two is correctly specified (Bang and Robins, 2005). The estimator first requires the estimation of the propensity score, $\hat{e}(x_i)$, and the outcome models, $\hat{\mu}_1(X_i)$ and $\hat{\mu}_0(X_i)$. These estimates are then combined in the following way to estimate the ATE:

$$\widehat{ATE}^{DR} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{t_i(y_i - \hat{\mu}_1(x_i))}{\hat{e}(x_i)} + \hat{\mu}_1(x_i)\right) - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(1-t_i)(y_i - \hat{\mu}_0(x_i))}{1-\hat{e}(x_i)} + \hat{\mu}_0(x_i)\right).$$

The doubly robust estimator incorporates "augmentation" terms that correct errors in the estimation process. In the ATE estimator, the terms $(y_i - \hat{\mu}_1(x_i))$ and $(y_i - \hat{\mu}_0(x_i))$ represent the difference between observed outcomes

and predicted outcomes. These augmentation terms serve as "error corrections": if the outcome regression model is misspecified, the propensity score weighting helps correct the bias, and if the propensity score model is misspecified, the outcome regression model helps correct the bias. This dual correction mechanism provides the estimator's "double robustness" property.

Implementation Details We leveraged the strengths of both the outcome regression and propensity score models identified in our previous analyses. We used Gradient Boosting to model the potential outcomes and Random Forest for the propensity score model. This decision aligns with the implementation of the previous methods.

4 Results

In this section, we examine the estimated Average Treatment Effect (ATE) of adult enrollment on academic success under different preprocessing design choices. We first analyze the causal effect of adult enrollment using different estimation methods, and then explore the sensitivity of ATE to variations in the preprocessing configurations.

4.1 Causal Analysis of the Effect of Adult Enrollment on Academic Success

To estimate the ATE, we employed multiple causal inference methodologies, including S-Learner, T-Learner, IPW, PSM and Doubly Robust estimation.

The results in Table 2, Appendix C, consistently indicate a negative effect of adult enrollment on academic success. The choice of estimation method affects the magnitude of the effect, with PSM exhibiting greater variability and sometimes yielding confidence intervals that include zero, indicating weaker significance. In contrast, methods such as the S-Learner, T-Learner, and Doubly Robust estimators provide relatively stable negative estimates. This suggests that, across methodologies, adult enrollment is associated with a decreased probability of completing academic programs.

4.2 Sensitivity of ATE to Design Choices in Preprocessing

While the causal effect remains consistently negative, its magnitude is sensitive to preprocessing design choices, particularly the definition of adulthood and academic success.

Figure 4 in Appendix C visualizes the distribution of ATE estimates across different configurations, which derive from specific design choices made in the preprocessing stage. These design choices include the selection of the age threshold defining adulthood and the definition of academic success. The histogram shows a range of estimated effects, reinforcing how methodological choices impact the measured causal effect. By rounding ATE values to two decimal places, we simplify the visualization and avoid an excessive number of unique values that would make interpretation more challenging.

A key observation is that increasing the adulthood threshold intensifies the negative ATE. Similarly, the definition of academic success has a profound impact on the estimated causal effect, but this effect varies depending on the chosen adulthood threshold. At an adulthood threshold of 20, a strict definition (graduation only) yields more negative ATE values compared to an inclusive definition (graduation or continued enrollment). However, for higher adulthood thresholds, this pattern does not hold consistently. These differences underscore that different design choices in the data preprocessing stage can lead to different estimations of the causal effect.

Overall, our findings highlight that while the negative impact of adult enrollment is consistent across methods, its precise magnitude and significance depend on design choices in the preprocessing stage. Researchers must be mindful of these choices when interpreting ATE estimates in causal studies.

5 Conclusions and Discussion

This study investigates how different design choices in data preprocessing influence the estimation of the ATE. Our approach serves as a proof of concept, using a specific dataset from Portuguese higher education institutions and focusing on the causal question of how enrolling as an adult student impacts academic success. The two primary design choices we explored were the selection of an age threshold for adulthood and the definition of academic success.

Our results demonstrate that seemingly minor preprocessing decisions significantly impact the estimated causal effect. Across different configurations, the ATE varied considerably, indicating that the estimated effect of adult enrollment

is not invariant but depends on how key variables are defined and processed. The methodological choices involved in causal inference—such as defining treatment groups and outcomes—can meaningfully alter the conclusions drawn from an analysis, emphasizing the importance of transparency and justification in preprocessing decisions.

5.1 Implications of Design Choices on ATE Estimation

The selection of an age threshold affects both the magnitude and statistical significance of ATE estimates. As the age threshold increases, the estimated negative impact of adult enrollment becomes more pronounced. This suggests that older students face greater challenges in completing their academic programs, potentially due to increased external responsibilities such as work and family obligations. A lower threshold, by contrast, produces smaller ATE estimates, reflecting a less distinct separation between traditionally younger students and those just above the defined adulthood cutoff.

Similarly, the definition of academic success has a profound impact on the estimated causal effect, but this effect varies depending on the chosen age threshold. At an age threshold of 20, a strict definition (graduation only) yields more negative ATE values compared to an inclusive definition (graduation or continued enrollment), as the latter accounts for students who are still making progress toward completion. However, for higher age thresholds, this pattern does not hold consistently, suggesting that as students enroll at later ages, the impact of the outcome definition becomes less predictable.

Our findings highlight that different preprocessing choices may lead to different interpretations of the causal effect. This variability underscores the necessity of conducting robustness checks across multiple configurations and explicitly stating the rationale for preprocessing decisions. Without such considerations, causal conclusions may be contingent on arbitrary or context-specific definitions rather than reflecting an inherent treatment effect.

5.2 Limitations and Future Research Directions

While our study provides valuable insights into the sensitivity of ATE to preprocessing choices, several limitations must be acknowledged:

1. **Scope of Application**: This work serves as a proof of concept for a broader question—how the ATE is affected by different design choices in data preprocessing. Our findings are based on a specific dataset from Portuguese higher education institutions, and results may not generalize to different datasets and causal inference contexts.
2. **Preprocessing Bias**: The predefined categories of adulthood and success influence the estimated effect, but alternative ways of defining these variables might yield different conclusions. Future research should explore a wider range of possible definitions.
3. **Unobserved Confounders**: Despite our inclusion of various demographic and academic variables, important unmeasured factors—such as motivation, financial stability, and study habits—could influence both treatment assignment and outcomes, potentially biasing the ATE estimates.

Future research should aim to generalize these findings by examining how the ATE is influenced by preprocessing design choices across different datasets with various causal contexts, helping to determine whether the effects observed in this study hold more broadly. Additionally, further exploration of alternative definitions for treatment and outcome variables is necessary to assess the robustness of causal claims under varying preprocessing choices.

Our findings reinforce the necessity of careful and transparent preprocessing decisions in causal research. Without explicit acknowledgment and justification of preprocessing choices, different studies may arrive at divergent conclusions even when analyzing the same underlying phenomenon. The results of this study emphasize that preprocessing is not a neutral step in causal analysis—it is a crucial determinant of the estimated treatment effect and must be treated as such in research design and interpretation.

References

Eyman Alyahyan and Dilek Düştegör. Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education*, 17(1):3, 2020.

Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424, 2011.

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Anirban Basu, Aig Unuigbe, and Cristina Masseria. Understanding differences between what alternate propensity score methods estimate. *Journal of Managed Care & Specialty Pharmacy*, 29(4):391–399, 2023.

Sander Greenland and James M Robins. Identifiability, exchangeability, and epidemiological confounding. *International journal of epidemiology*, 15(3):413–419, 1986.

Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.

Miguel A Hernán and James M Robins. Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*, 60(7):578–586, 2006.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

Valentim Realinho, Jorge Machado, Luís Baptista, and Mónica V. Martins. Predicting student dropout and academic success. *Data*, 7(11), 2022. ISSN 2306-5729. doi: 10.3390/data7110146. URL `https://www.mdpi.com/2306-5729/7/11/146`.

Donald B Rubin. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74(366a):318–328, 1979.

Juliana Schulz, Erica EM Moodie, and Susan M Shortreed. No unmeasured confounding: Known unknowns or... not? *American Journal of Epidemiology*, 192(9):1604–1605, 2023.

A Treatment and Outcome Distributions Under Different Preprocessing Design Choices

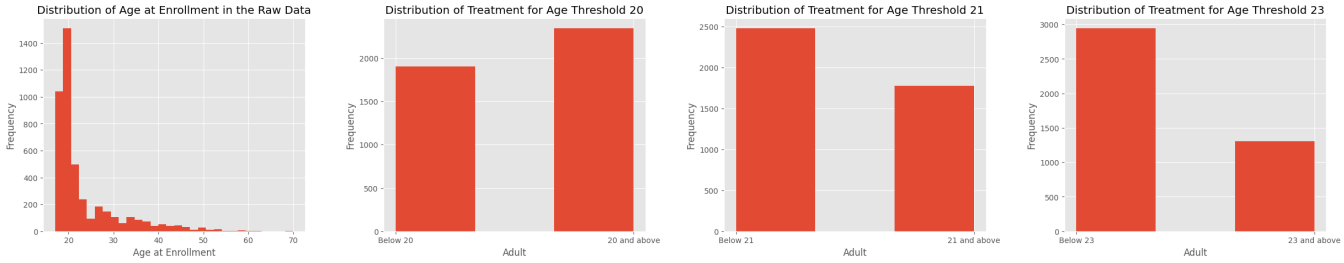

Figure 1: Distribution of the treatment variable in the raw data (left) and for different adulthood thresholds (right)

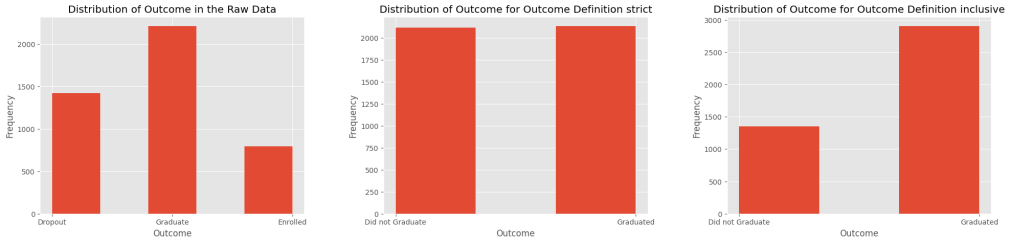

Figure 2: Distribution of the outcome variable in the raw data (left) and under strict and inclusive definitions (right)

B Common Support Assumption

Table 1: Propensity score ranges for treated and untreated groups across adulthood thresholds

Adulthood Threshold	Treated Range	Untreated Range	Common Support
Age $\geq$ 20	[0.0479, 1.0000]	[0.0178, 0.9874]	Holds
Age $\geq$ 21	[0.0129, 1.0000]	[0.0066, 0.9856]	Holds
Age $\geq$ 23	[0.0035, 1.0000]	[0.0005, 0.9859]	Holds

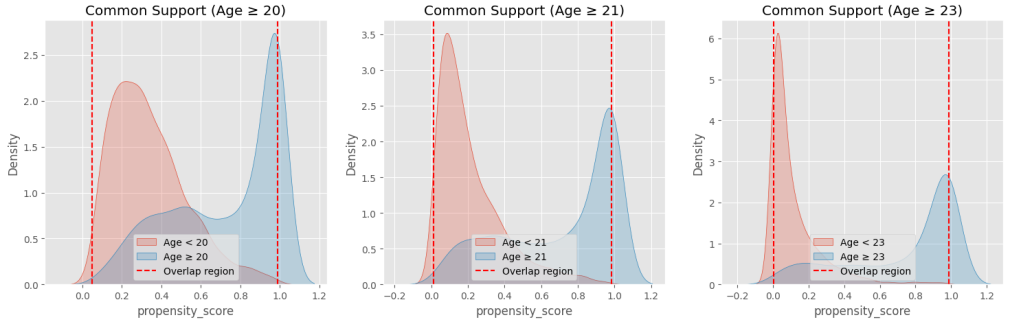

Figure 3: Common support of the propensity scores for each adulthood threshold

C ATE Results and Visualization

Table 2: ATE Estimates under Different Configurations

Config. No.	Adulthood Threshold	Outcome Definition	Method	ATE	95% Bootstrap CI	Statistically Significant?
1	20	Strict	S-Learner	-0.0807	[-0.1121, -0.0487]	Yes
2	20	Strict	T-Learner	-0.0962	[-0.1417, -0.0656]	Yes
3	20	Strict	IPW	-0.0586	[-0.0784, -0.0364]	Yes
4	20	Strict	PSM	-0.1425	[-0.2059, 0.1729]	No
5	20	Strict	Doubly Robust	-0.0995	[-0.1829, -0.0568]	Yes
6	20	Inclusive	S-Learner	-0.0720	[-0.0963, -0.0449]	Yes
7	20	Inclusive	T-Learner	-0.0765	[-0.1270, -0.0501]	Yes
8	20	Inclusive	IPW	-0.0278	[-0.0533, -0.0070]	Yes
9	20	Inclusive	PSM	-0.1362	[-0.1958, 0.1532]	No
10	20	Inclusive	Doubly Robust	-0.0632	[-0.1345, -0.0291]	Yes
11	21	Strict	S-Learner	-0.1134	[-0.1477, -0.0819]	Yes
12	21	Strict	T-Learner	-0.1288	[-0.1764, -0.0999]	Yes
13	21	Strict	IPW	-0.2056	[-0.2248, -0.1829]	Yes
14	21	Strict	PSM	-0.1971	[-0.3021, 0.0902]	No
15	21	Strict	Doubly Robust	-0.1472	[-0.2226, -0.1064]	Yes
16	21	Inclusive	S-Learner	-0.1378	[-0.1635, -0.1009]	Yes
17	21	Inclusive	T-Learner	-0.1462	[-0.1827, -0.1091]	Yes
18	21	Inclusive	IPW	-0.2404	[-0.2624, -0.2156]	Yes
19	21	Inclusive	PSM	-0.0980	[-0.2970, 0.0567]	No
20	21	Inclusive	Doubly Robust	-0.1238	[-0.1753, -0.0744]	Yes
21	23	Strict	S-Learner	-0.1598	[-0.1946, -0.1203]	Yes
22	23	Strict	T-Learner	-0.2020	[-0.2473, -0.1602]	Yes
23	23	Strict	IPW	-0.3127	[-0.3290, -0.2903]	Yes
24	23	Strict	PSM	-0.2538	[-0.4418, -0.0502]	Yes
25	23	Strict	Doubly Robust	-0.2602	[-0.3346, -0.1983]	Yes
26	23	Inclusive	S-Learner	-0.1825	[-0.2166, -0.1434]	Yes
27	23	Inclusive	T-Learner	-0.2110	[-0.2563, -0.1723]	Yes
28	23	Inclusive	IPW	-0.3961	[-0.4141, -0.3707]	Yes
29	23	Inclusive	PSM	-0.2633	[-0.4181, -0.0130]	Yes
30	23	Inclusive	Doubly Robust	-0.2356	[-0.3188, -0.1666]	Yes

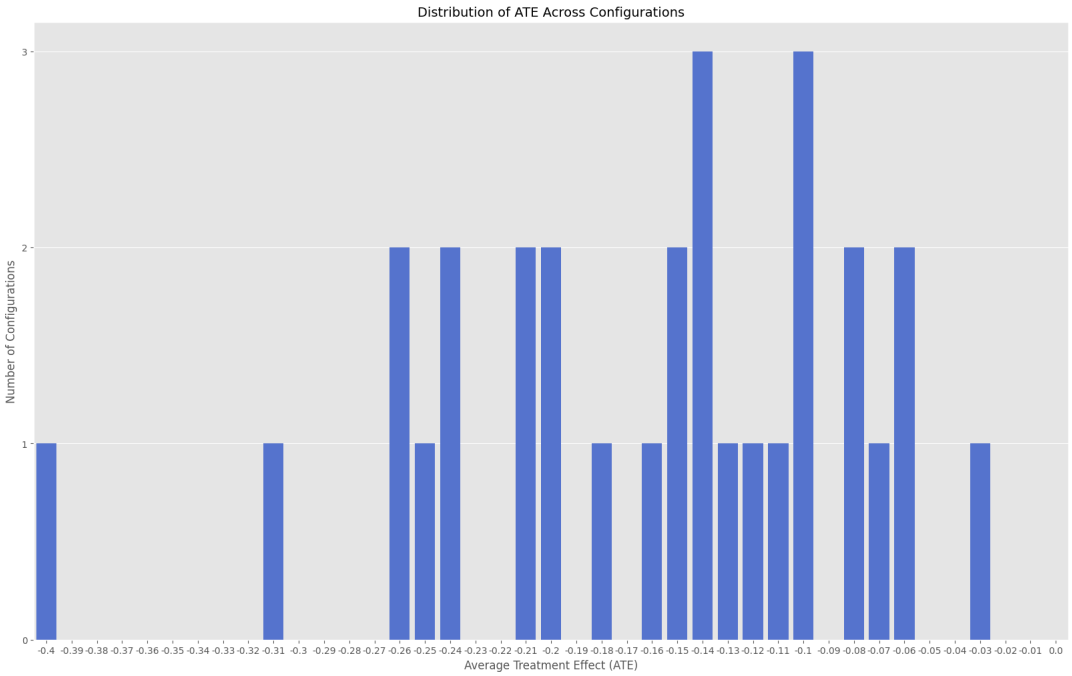

Figure 4: Distribution of ATE across configurations.