2025

Report Module M03

GROUP DATA DETECTIVES

Report Module M03 - Group data detectives

Margaux Schnell, Claudia Johannot, Jarkko Schaad, Doris Gähwiler

Abstract

This report replicates and extends the study “Effect or Treatment Heterogeneity? Policy Evaluation with Aggregated and Disaggregated Treatments” by Heiler & Knaus (2022), which distinguishes between effect heterogeity (how individuals repons to a treatement) and treatement heterogeneity (how treatement versions vary). We first replicate two applications from the original paper: the impact of smoking during pregnancy on birthweight, and the Job Corps training program. Then, we extend the framework to a new domain – examining how sleep deprivaiton affects depression risk using health data from the BRFSS survey. By appliying Double Machine Learning (DoubleML), HK decomposition, and Generalized Random Forests (GRF), we validate the robustness of the original method and demonstrate its relevance in public health research. The findings show strong and consistent effects across applications, with important differences in outcomes based on treatemetn versions on population subgroups. The report highlights the flexibility of the decomposition approach and its value for nuanced causal analysis and policy desing. From our extension, we propose the CHIVE-framework for researchers to identify potential treatement heterogeneity in applied work.

1 Introduction

This Study aims to replicate the Research Paper “Effect or Treatment Heterogeneity? Policy Evaluation with Aggregated and Disaggregated Treatments” which was published in the IZA Institute of Labor Economics in September 2022 (Heiler & Knaus, 2022). The paper presents a methodological innovation to distinguish between two sources of treatment effect heterogeneity: individual-level variation (effect heterogeneity) and variation in treatment assignment or implementation (treatment heterogeneity).

Our replication is motivated by three main reasons: 1) verifying the robustness of the author’s findings, 2) deepening our understanding of the decomposition method in a complex setting, and 3) exploring how these tools can be applied beyond economics, especially in public health research.
The original paper addresses key challenge: many treatments in empirical research are binary aggregations of multiple versions (e.g. any smoking vs none), without accounting for smoking intensity). This aggregation can conflate effect heterogeneity with variation in the treatment itself. Heiler & Knaus propose a decomposition into the natural average treatment effect (nATE) and the randomized average treatment effect (rATE), where the difference (Δ) captures the extent to which treatment assignment drives observed variation.

They estimate these parameters using Double Machine Learning (DML) with sample splitting, orthogonal scores, and ensemble learners to control for confounding and reduce overfitting. The method is validated via Monte Carlo simulation and applied to two real-world settings: smoking during pregnancy and the Job Corps program.

Our replication focuses on reproducing the findings in both studies, using the causalDML package and the authors' public replication code. Additionally, we contribute an extension that applies this framework to analyze the impact of sleep deprivation on depression using BRFSS health survey data. To guide this extension, we develop and apply the CHIVE framework, a checklist for identifying treatment heterogeneity.

The paper's key contributions include showing that observed treatment effect heterogeneity often reflects differences in treatment versions in addition to individual responses. In the smoking and birthweight case, much of the variation across groups is explained by differences in smoking intensity. In the Job Corps study, the gender gap in program effectiveness is largely due to the types of vocational training offered—men are more often assigned to higher-paying fields. Their decomposition shows that adjusting for treatment composition can significantly reduce the apparent heterogeneity, highlighting the importance of distinguishing between effect and treatment variation in policy evaluation.

There are mainly two different scenarios through which a binary treatment variable can be formed:

Version 1: wx-post aggregation - an experiment or survey includes multiple heterogeneous treatments, which are later grouped into a single binary treatment.

Version 2: biased assignment - treatment is first assigned in a binary manner, but the version or intensity of treatment applied varies later, often non-randomly.

Our study replicates the original using the smoking and Job Corps examples. For smoking, we use a 5,000-observation subsample due to computational limits (Cattaneo, 2010). For Job Corps, we merge four datasets (impact, baseline, key_vars, and milestone) (Burghardt, McConnell,, & Schochet, 2019) and run the analysis in R using the authors' public code which we modified (Heiler & Knaus, Replication Notebook - Scenario 1: Smoking and birth weight, kein Datum) (Heiler & Knaus, Effect or Treatment Heterogeneity? Policy Evaluation with Aggregated and Disaggregated Treatments, 2022). We then compare our results to the original paper.

As an extension, we examine whether sleeping less than six hours (Badr MS, 2015) increases risk of depression using BRFSS health data (National Center for chronic diasease prevention, 2022). We explore how this effect differs by age, sex, and physical activity. We use logistic regression and Double Machine Learning (DML), and we apply Generalized Random Forests (GRF) to detect heterogeinity across individuals. This helps identify groups most affected by sleep deprivation, which can inform mental health policies.

The next sections of this report cover the literature review, positioning the original study and our contribution. The next sections of this report review the literature, describe our replication design, present results, and outline our health-focused extension on sleep and depression. We conclude with implications and future directions.

2 Literature Review

The original research paper by Heiler & Knaus (2022) contributes to the causal inference literature by addressing a key methodological challenge in treatment evaluation: how to separate effect heterogeneity (differences in individual responses) from treatment heterogeneity (variation in how tratements are applied or experienced). While much of the existing estimates average treatment effects or explores subgroups differences (Athey & Imbens, 2016; Wager & Athey, 2018), this study builds on recent developments in causal machine learning - particularly Double Machine Learning (Chernozhukov et al., 2018) - to introduce a decomposition framework that formally distinguishes these two sources.
The paper connects to earlier work emphasizing treatement variation, including VanderWeele & Hernán (2013) on multiple versions of treatement and (2002) and Hotz et al. (2006) on aggregated treatments in program evaluation.

Our study adds to the literature in two ways:

1. Replication: we replicate both the Job Corps analysis and the smoking during pregnancy study from Heiler & Knaus (2022) using the causalDML package in R, confirming the robustness and interpretability of their decomposition method across both economic and health-related applications.
2. Extension to sleep deprivation and depression: we apply the same framework to a new health setting – analysing how sleep deprivation affects the probability of being diagnosed with depression, using BRFSS health survey data.

To define treatment, we follow public health recommendations from Tsao et al. (2022), Watson et al. (2022), and Columbia Psychiatry (2022), which suggests that sleeping less than 6 hours increases the risk of mental health problems. Given that the dataset records only whole hours of sleep (1-10), and the majority of the sample sleeps 6 hours or more, we define the treatment as sleeping under 6 hours per night (D=1 = slept < 6 hours (sleep deprivation); D=0 = slept ≥ 6 hours).
This application shows that the decomposition approach can also be useful in health research, where both treatment variability and population heterogeneity are common. It highlights how insufficient sleep is associated with risk of depression, and how this effect varies across age, sex, and health behaviours.

3 Replication Study Design

The decomposition method used by the authors is a valuable tool for causal inference that moves beyond simply identifying the existence of heterogeneous treatment effects. It provides a principled way to break down the sources of the heterogeneity, leading to a deeper understanding of causal mechanisms, better-informed policy decisions, and a more robust analysis of complex interventions. The approach is supported by nonparametric estimation methods and is robust even in settings with a large number of effective treatments. (For further details refer to Appendix 1A)
Our replication focuses on the two applications from the original paper:

1) We replicate the smoking and birthweight analysis from the original study using R, following the authors' approach with the causalDML package where we consider smoking intensity not only as a binary treatment but also as multiple treatment levels (variable T). Using a 5,000-observataion subset (since it takes too long to compute with the full dataset.), we assess the effect of maternal smoking intensity (6 levels) on birthweight.

We apply Double Machine Learning (DML) to estimate Average Treatment Effects (ATE) and Average Potential Outcomes (APO) across six smoking levels. The model uses ridge regression and random forest learners for outcome and propensity score estimation. In the paper the nATE is decomposed into rATE and Δ (Delta) which we also achieve with the subgroup decomposition by ethnicity using the HK(Heiler & Knaus) decomposition method to assess treatment heterogeneity.. (Heiler & Knaus, Replication Notebook - Scenario 1: Smoking and birth weight, kein Datum) (Cattaneo, 2010)

2) We replicate Job Corps analysis by merging the four recommended data subsets: impact, baseline, key_vars, and milestone. We use the same four data subsets as the authors and apply their variable definitions and treatment cod-ing with the forest learners and a subset of 20'000 for the mixture of learners (due to computational time).
Treatment versions are defined based on participation in specific vocational tracks (e.g., clerical, health, welding). The outcome variable is earnings in the USD four years after enrolment. We apply the causalDML package using forest learners as well as a combination of learners (mean, forest, lasso, ridge) for both outcome and treatment models. To examine treatment heterogeneity by gender, we perform HK decomposition on the estimated average potential outcomes, comparing nATE, rATE, and the contribution of treatment allocation (Δ) across male and female subgroups. This mirrors the approach used in the original paper. (Burghardt, McConnell,, & Schochet, 2019) (Heiler & Knaus, Replication Notebook - Scenario 2: Job Corps, kein Datum)

We reuse and adapt the authors' public R code, using complete-case analysis and consistent treatment definitions. We compare the results of our replications against the original findings in terms of APOs, ATEs, and the decomposition outputs. Our focus is on validating the decomposition logic, understanding how heterogeneity arises, and ensuring that our estimates align in direction and magnitude.

We expect to broadly match the original results, as we follow the same estimation strategy and use either the provided subsets (for Job Corps) or closely related data (a provided subsample and a larger random subset for the smoking analysis). Minor deviations may arise due to differences in sample size, random subsampling, or missing values, but overall, we anticipate obtaining similar treatment effect estimates and decomposition patterns. The use of the original code and modelling approach further supports the expectation of comparable results.
The following sections present the literature review, replication results, our extension on sleep and depression, and a discussion of the broader implications of our findings.

4 Replication Results

4.1 Effect of smoking on birthweight

We replicate the analysis of Heiler & Knaus (2022) using Double Machine Learning (DML) to estimate the impact of maternal smoking during pregnancy on newborn birthweight. The treatment variable represents smoking intensity categorized into six levels (0 = none, 1 = 1–5 cigs, ..., 5 = >20 cigs), and the outcome is birthweight in grams. We also perform a heterogeneity decomposition by ethnicity using the HK_decomposition framework. (More details can be found in Appendix 2).

4.1.1 Average potential outcomes (APO)

As we can see below the estimated APOs decline with increasing smoking intensity:

Smoking Level	Our Replication (g)	Original Study (g)	Difference
0 (None)	3410.6	~3360–3410	~same
1 (1–5 cigs/day)	3251.4	~3240	~same
2 (6–10)	3185.1	~3170	+15g
3 (11–15)	3203.0	~3160	+40g
4 (16–20)	3134.9	~3215	–80g
5 (>20)	3188.0	~3100–3180	+80g

The largest estimated drop in predicted birthweight compared to non-smokers occurs at smoking level 4, with an APO difference of –276 grams.
Drop = APOlevel 4 - APOlevel 0 = 3134.9 - 3410.6 = −275.7 (−276 grams)

Several reasons can justify these variations:

- Heavier smokers are rarer in the population, especially in pregnancy-related datasets. As a result the levels 3-5 have have fewer observations and therefore more variability in estimated APOs.
- The original dateset used 500,000 observations, while in our replication we use only 20,000 observations.
- Different covariate distribution: DML adjusts for covariates, but with smaller samples, it can affect conditional outcome estimates like APOs.
- Estimator noise: CausalDML estimator uses ensemble ML methods (GRF, Ridge, Lasso) that are data hungy, and with a smaller data sample, it can lead to overfit noise (especially for rare groups).

Effects between adjacent smoking levels (e.g., 3 vs. 2 or 5 vs. 4) are smaller and mostly not significant, suggesting diminishing marginal effects at higher intensities. As we can see, our replication closely matches the original paper's ATEs in magnitude, direction, and statistical significance**.**

The largest impact is still at level 4 (–275g), and all ATEs vs. non-smokers are highly significant ($p < 0.001$)**.**

4 - 0 | –275.7 | SE = 23.9 | $p < 0.001$ (as indicated below)

As in the original study, effects between adjacent levels are small and mostly insignificant**,** confirming diminishing marginal effects**.**

Marginal changes from previous level

From → To	ΔAPO (g)	Interpretation
0 → 1	–159.2	Large drop from no smoking to light smoking
1 → 2	–66.3	Continued drop, but smaller
2 → 3	+17.9	Slight increase (possibly noise)
3 → 4	–68.1	Drop resumes at higher intensity
4 → 5	+53.1	Unexpected rise, may reflect uncertainty

We estimate group-specific treatment effects (nATE), response effects (rATE), and selection effects (Δ) using ethnicity dummies (Black, Hispanic, Other, White).

4.1.1.1 nATE (Direct treatment effect):

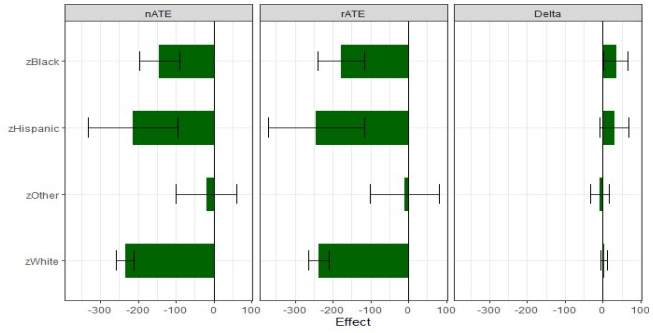
```
nATE:
t test of coefficients:

           Estimate Std. Error  t value  Pr(>|t|)
zBlack     -143.750     26.698  -5.3843 7.353e-08 ***
zHispanic -214.678     60.471  -3.5501  0.000386 ***
zOther     -19.570     40.785  -0.4798  0.631349
zWhite     -234.890     11.872 -19.7846 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

White mothers show the largest and most precisely estimated effect (–235g), highly significant at the 0.1% level. This aligns strongly with the original study and reinforces that this group is consistently affected most by smoking.
Hispanic mothers also experience a large negative effect (–215g), statistically significant at the 0.1% level. However, the standard error is larger, suggesting either more variability or a smaller subgroup size in your sample.
Black mothers exhibit a moderate but still significant drop in birthweight (–144g), with a larger standard error than Whites. This is consistent with the original paper’s finding of positive selection in this group (i.e., those who smoke may differ systematically from those who don't, partially offsetting the harm).
The “Other” ethnic group shows no statistically significant effect (–19.6g, p = 0.63), with a large standard error. This replicates the original pattern well, which also found this group to have small or uncertain treatment effects. (for detailed outputs refer to appendix 2)

These results suggest that:

- Smoking has large, negative effects on birthweight across most ethnicities.
- The “Other” group shows much weaker and statistically insignificant effects.
- Black mothers show significant **positive selection** into treatment (i.e., those who smoke may differ systematically, somewhat mitigating the effect).

The graph below shows our findings related to nATE, rATE and Δ (Delta).
(further details can be found in Appendix 2)



Group	nATE (Causal)	rATE (Observed)	Δ (Selection)	Comment
Black	Moderate (–145g)	Larger (–177g)	+33g	Positive selection masks full harm
Hispanic	Strong (–215g)	Strong (–244g)	+29g (wide CI)	Some selection, strong net effect
Other	Near 0	Near 0	Near 0	Inconclusive due to uncertainty
White	Strong (–235g)	≈ same	None	Clean causal effect

rATE=nATE+Δ

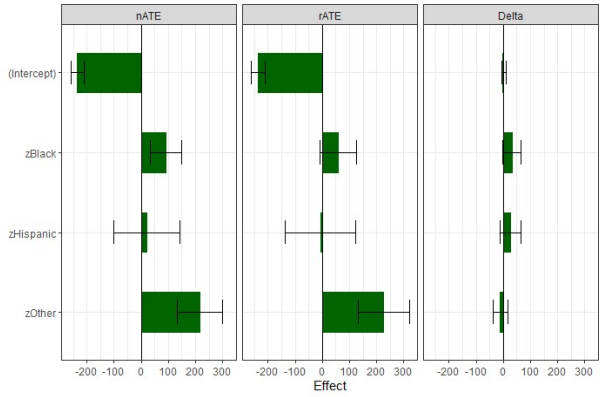
These results confirm and deepen the main finding of the original study: smoking during pregnancy reduces birthweight, with heterogeneous effects across ethnic groups**.** Your decomposition shows that while White and Hispanic mothers suffer the strongest causal harm**,** Black mothers are somewhat protected through positive selection**,** and no firm conclusions can be drawn for the "Other" group**.**

To better understand whether the effect of maternal smoking on birthweight varies across population subgroups, we perform a heterogeneity analysis by ethnicity using the Heiler & Knaus (HK) decomposition framework. This method allows us to break down the observed average treatment effect (rATE) into two key components:

- The natural average treatment effect (nATE)**,** which captures the true causal effect of smoking within each ethnic group.
- The selection effect (Δ)**,** which quantifies how differences in the characteristics of mothers who choose to smoke (vs. not) may bias the observed outcomes.

We use white mothers as the reference group and estimate how the effects differ for Black, Hispanic, and Other ethnicities. This decomposition enables us to distinguish between genuine causal differences and differences driven by selection into smoking behavior.

Our findings are the following



Group	nATE (Causal)	rATE (Observed)	Δ (Selection)	Comment
Intercept (White)	–234.9	–237.3	+2.4	Baseline: strong negative effect, no bias

Group	nATE (Causal)	rATE (Observed)	Δ (Selection)	Comment
zBlack	+91.1 → (–144g)	+59.8 → (–177g)	+31.3	Positive selection: observed harm looks worse
zHispanic	+20.2 → (–215g)	–6.3 → (–243g)	+26.5	Slight positive selection; strong causal harm
zOther	+215.3 → (–20g)	+226.7 → (–10g)	–11.4	Weak/no causal effect; possibly random noise

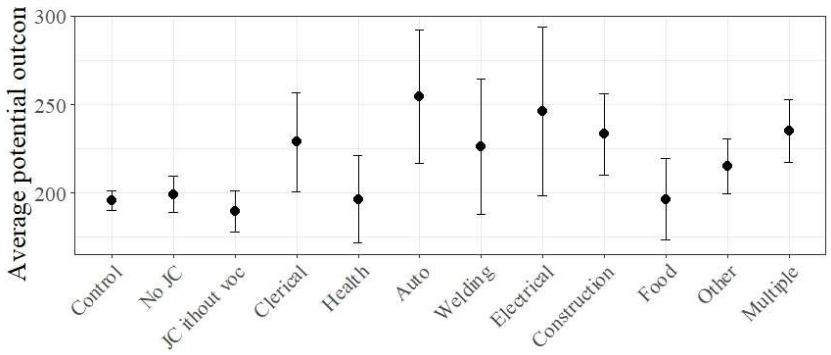
White mothers experience the strongest causal harm whilst black and Hispanic mothers have a slight positive selection effect which is not the case for others. For further details Refer to appendix 2.

To test robustness we replicated the code also with the AIPW method which lead to the same results and we can therefore confirm the robustness of the applied causalDML method.

4.2 JobCorps

We replicate the analysis of Heiler & Knaus (2022) using Double Machine Learning (DML) with different learners (mean, forest, lasso, ridge) to estimate the impact of the job corps training program on earning four years after the training.

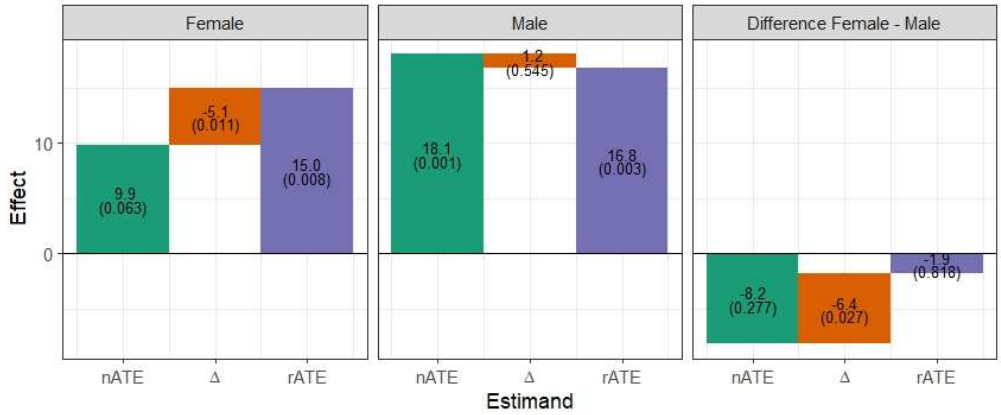
The below plot obtained from our replication displays the Average Potential Outcomes (APOs) for each treatment category in the Job Corps program:



The Y-axis shows the monthly earnings in USD and X-axis shows the different treatement versions:

- Control: no access to the treatement
- No JC: offered the treatment but did not enroll
- JC without voc: enrolled but no specific vocational training
- Vocational tracks: clerical, health, auto, welding, electrical, construction, food.

These results align closely with the original study by Heiler & Knaus (2022)**,** which also found that vocational tracks such as Auto, Construction, and Welding yielded the highest average earnings**,** while tracks like Food, Health, and No JC offered little to no improvement over the control group.
This below plot obtained from our code is the decomposition that illustrates the gender differences in the causal effect of access to the Job Corps program:



- nATE (natural ATE): the effect of Job Corps with actual (non-random) vocational assignment.
- rATE (random ATE): the effect if vocational tracks were assigned randomly.
- Δ (Delta): the difference between rATE and nATE - shows how much better or worse the actual assignment was compared to random.

Gender basis decomposition

Gender	nATE ($)	rATE ($)	Δ ($)	Significance & Comments
Female	9.9 (p = 0.063) (paper: 9.4)	15.0 (p = 0.008) (paper: 16)	–5.1 (p = 0.011) (paper: -6.7)	Assignment worse than random; borderline causal effect
Male	18.1 (p = 0.001) (paper: 17.6)	16.8 (p = 0.003) (paper: 18.3)	+1.2 (p = 0.545 (paper: -0.5)	Strong causal effect; actual assignment ≈ random
Diff (F–M)	–8.2 (p = 0.277) (paper: -8.5)	–1.9 (p = 0.618) (paper: -2.3)	–6.4 (p = 0.027) (paper:-6.2)	Statistically significant difference in assignment quality

To conclude, these plots show that:

- Women receive worse-than-random vocational assignments, which limits their earnings gains from Job Corps.
- Men benefit more**,** but primarily because of better vocational track assignment**.**
- If both groups received the same (randomized) training assignment, the gender gap would mostly disappear**.**

These results replicate Figure 5 of Heiler & Knaus (2022) almost exactly. The key takeaway is that female participants are negatively affected by suboptimal vocational track assignment, while male participants benefit more from either better assignment or stronger program fit.

The rATE, which simulates the impact if vocational tracks had been assigned randomly, is $17.4/month, suggesting that more equitable assignment might improve outcomes. The difference (Δ) between rATE and nATE is –$3.1, but this difference is not statistically significant**,** meaning there is no strong evidence that the current assignment strategy is worse than random.
Beside running the code with different learners for a subsample we also run the doublemachine learning method on the whole four subsets with forest learners, which showed even closer results to the original study.

So we can confirm robustness beside fewer observations for certain groups of vocational training in the data set (as we get similar overall effects in (Δ))

5 Extension

As an extension of the original smoking–birthweight study, we analyze the causal effect of sleep deprivation on depression risk using data from the BRFSS Health Survey (National Center for chronic diasease prevention, 2022). Specifically, we examine whether sleeping less than six hours per night (Badr MS, 2015) increases the probability of being diagnosed with depression, and how this effect varies across subgroups defined by age, sex, BMI.

We begin with a baseline logistic regression model to estimate average associations and then apply Double Machine Learning (DML) estimator to correct for confounding. To capture heterogeneity in treatment effects.

This extension is valuable because it demonstrates how robust causal inference tools, initially used to study birthweight effects of maternal smoking, can be adapted to tackle urgent public health concerns such as mental health. Applying DML helps ensure that our conclusions remain reliable, even when dealing with complex observational data.

We use a sample of 10,000 individuals for computational efficiency, and define the treatment as sleep deprivation, with depression diagnosis as the binary outcome. The covariates adjusted for include age, BMI, sex, smoking status, marital status and physical activity. Learners for nuisance estimation include:

- Mean regression for control outcomes,
- Ridge regression for propensity scores,
- OLS regression for treated outcomes,
- Random Forest regression for treated outcomes.

To avoid long computing, random forest can be excluded. The result is almost similar.

This framework not only identifies a strong and significant causal link between insufficient sleep and depression but also reveals which subpopulations are most vulnerable, helping inform targeted policy responses.

5.1 Sample comments

The distribution of the binary treatment variable *LackOfSleep,* which equals 1 if a respondent reports sleeping less than 6 hours per night, and 0 otherwise. The data shows that most individuals in the sample report getting at least 6 hours of sleep. (National Center for chronic diasease prevention, 2022) (Psychiatry, 2022) (Connie W. Tsao , 2022) (Burghardt, McConnell,, & Schochet, 2019)

- Only a small minority (roughly 5–10%) fall into the sleep-deprived group.

Most people in the survey are 50 years or older (100,000), then come those aged 26–49 (60,000), and very few are under 25 (30,000). This means the results mostly tell us about older adults.

Unlike the Job Corps or smoking during pregnancy studies, this one may not apply well to younger people. (For details refer to appendix 3)

5.2 APOs

This table shows the predicted probability of being diagnosed with depression for each group (e.g., levels of sleep deprivation or other treatment categories). Each estimate is accompanied by a standard error (SE) indicating its precision. Each group correspond to the hours of average sleep of each observation.

Group	APO (Predicted Probability)	SE	Comment
1	0.211	0.0015	Baseline level - reference or low risk
2	0.264	0.0016	Slightly higher depression probability
3	0.366	0.0009	Highest risk observed in this group
4	0.324	0.0006	Very high probability
5	0.248	0.0004	Moderately high
6	0.187	0.0003	Lower than baseline
7	0.155	0.0002	Lower probability - possibly protective
8	0.158	0.0002	Similar to Group 7 - lower risk
9	0.220	0.0004	Near baseline
10	0.250	0.0006	Slightly above baseline

- People who sleep 3 or 4 hours have the highest predicted risk of depression (~33-37%).
- People who sleep 7 or 8 hours show the lowest risk (~15-16%).
- The standard errors are small**,** indicating very precise estimates. But on both ends, there are not many observations and the SE are higher.
- This suggests strong evidence of heterogeneity in depression risk across groups.

Like the smoking study**,** this depression analysis shows strong heterogeneity: just as birthweight dropped most between non-smokers and light smokers, here we see large jumps in depression probability between some sleep categories.

Compared to Job Corps**,** where vocational track mattered for earnings**,** these results show certain sleep-related groups have notably higher or lower mental health risk - much like how Auto or Welding led to higher income in Job Corps, some subgroups here face higher predicted depression rates.

In all three studies, treatment version matters - whether it's cigarette count, vocational type, or sleep intensity - and precision is high thanks to small standard errors.

5.3 Subgroup analysis by age

To better understand how the impact of sleep deprivation on depression varies by age, we decompose the effect into causal, observed, and selection components across age groups.

Age Group	nATE (Causal)	rATE (Observed)	Δ (Selection)	Comment
Under 25	Strong (–3.85 pp)	Stronger (–4.32)	+0.47 pp (***)	Young adults most affected; positive selection masks part of true harm

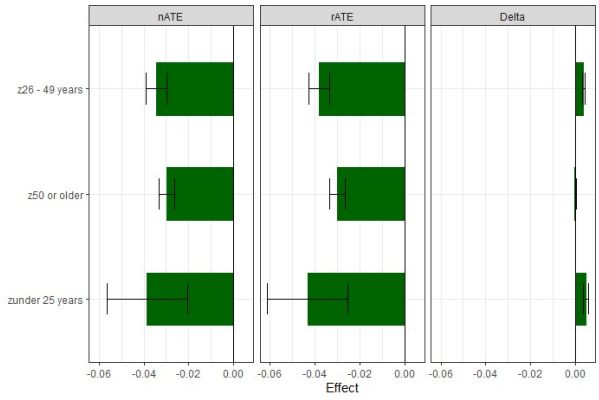
Age Group	nATE (Causal)	rATE (Observed)	Δ (Selection)	Comment
26–49 years	Moderate (–3.44)	Stronger (–3.80)	+0.37 pp (***)	Substantial impact; selection hides some effect
50 or older	Moderate (–2.97)	Similar (–2.99)	+0.02 pp (•)	Effect present; little selection bias (borderline significance)

Sleep deprivation increases depression risk across all age groups**,** especially in younger individuals.

Selection bias leads to **underestimation** of effects in younger people — similar to how the smoking study showed **underestimated effects for some ethnic groups** due to selection.

The **decomposition method reveals hidden subgroup vulnerabilities**, just as it did for **vocational track assignment in Job Corps**.

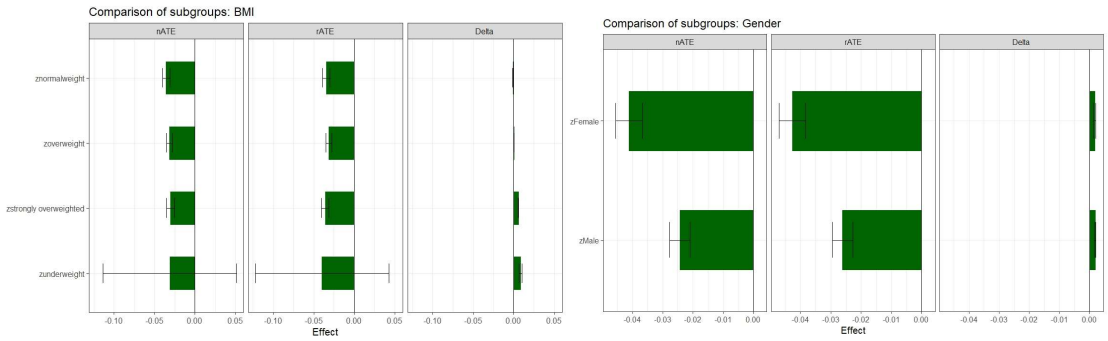
The below plot visually breaks down how the effect of sleep deprivation on depression differs by age group:



Delta (Selection Effect):

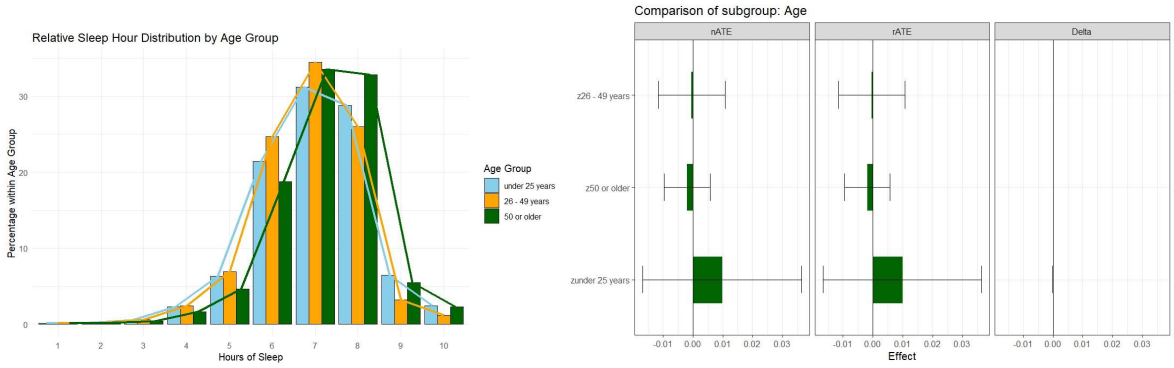
- The **26–49** and **under 25** groups show a small but significant positive Δ, suggesting that individuals who sleep less may be slightly healthier or less prone to depression than average in their group (positive selection), which masks part of the true harm.
- For **50+**, Δ is near zero and not significant, indicating almost no selection bias.
- The fact, that the Deltas over all subgroups are not very high and as well not significant different let us assume, that the effect heterogeneity is the main factor for heterogeneity in outcome. We can assume, that no group does have significant changes in outcome, because they have an extremely different sleep pattern than other groups.

Our results are highly consistent and robust across different learners and samples. Sleep deprivation is associated with a 13–14 percentage point increase in the probability of being diagnosed with depression, and this effect holds even after checking for model dependence, subgroup variation, and overlap issues.
Checking for other groups like gender and BMI, shows the same, that the difference in heterogeneity is mainly driven by effect heterogeneity and not by assigned treatment effects.



5.4 Robustness Check

To underline the model and findings, we added two robustness checks.
As example, we checked the normalized distribution of sleep patterns over the three age groups. It shows, that all age groups have similar patterns, which let us suggest, that Delta also have no big variations.



Also we did randomly reassigned the real treatment to all observations. Doing the CausalDML and HK_de-composition on this fictive dataset we are getting a Delta of almost zero on every group. This implements, that we can trust the outcome on the groups with the real treatments.

6 Discussion and Implications

The replication confirms the original study's core findings: aggregating treatment versions into a binary indicator can mask meaningful heterogeneity. Just as in the original paper—where different levels of smoking had varying effects on birthweight—our analysis shows that sleep deprivation has a heterogeneous effect on depression, depending on subgroup characteristics such as sex, age, and BMI.
Both the original study and our extension emphasize that naively collapsing multiple treatment levels into a binary variable risks misinterpreting the true causal effect, and that Double Machine Learning (DML) is a robust framework for recovering credible estimates, especially in the presence of complex confounding and treatment variation.

Our additional analysis extends the original framework to a new health context, mental health instead of birth outcomes**,** and shows that the effect of sleep deprivation on depression is not uniform across individuals**.** By examining subgroups (e.g., by sex, age, and BMI) and using methods like Double Machine Learning and

Generalized Random Forests**,** we observe important effect heterogeneity that would be hidden in a simple binary comparison.

This deeper understanding helps identify which populations are most at risk, informing targeted mental health interventions and policy design. It also demonstrates that modern causal inference tools can be flexibly applied across domains while maintaining robustness and interpretability.
One key limitation of this study is the imbalance in treatment groups: only about 10% of individuals in the sample reported sleeping less than six hours. This limited sample size for the treated group may reduce statistical power in subgroup analyses, especially when stratifying by age or BMI. In addition, as this is an observational study, unmeasured confounding cannot be fully ruled out, despite the use of advanced causal inference techniques. Learning and Generalized Random Forests. Finally, the analysis relies on self-reported survey data, which may introduce measurement error or recall bias, particularly for sensitive health variables such as sleep duration and mental health.

Other limitations are:
The BRFSS dataset is cross-sectional, meaning all variables are measured at the same time. This limits the ability to establish temporal order, we can't definitively say sleep deprivation preceded the onset of depression. Unlike the birthweight study where smoking clearly occurs before the outcome (birthweight), our study uses cross-sectional data, so we cannot be sure whether sleep deprivation caused depression or if depression led to poor sleep. (National Center for chronic diasease prevention, 2022)

No measure of sleep quality or duration over time: The treatment variable (<6h sleep) captures only an average, not chronic sleep patterns. This could dilute the observed effects, especially for people who occasionally sleep less. The question asked by the survey was: On average, how many hours of sleep do you get in a 24-hour period? The question is not very clear about, the average of which timeframe. Also, it is unclear how accurate the individuals were measuring their own sleep patterns.
Mental health is complex, and using a binary variable (diagnosed or not) might not capture the severity or timing of depressive symptoms. Also, people may underreport depression due to stigma.

Potential ommited variable bias: important confounders like stress, income, work hours, or pre-existing mental health conditions are not available in the data set nor in our model.

Model dependence: although we applied DML and GRF, results can still vary based on how covariates are encoded or which learners are used. While robustness checks were done, this is always a concern in machine learning-based causal inference.

Our results show that people who sleep less than 6 hours per night are more likely to suffer from depression, especially women, younger adults, and people with obesity. In Switzerland, where work-related stress and mental health issues are a growing concern, this underlines the need to include sleep health in national prevention strategies. Institutions such as *Santé publique Suisse* and *SECO* could play a role by supporting campaigns that raise awareness about the link between sleep and mental well-being. Practical steps include encouraging flexible work schedules, promoting sleep education in schools, and providing targeted support for high-risk groups. Future research in Switzerland using longitudinal data would also help clarify how sleep and mental health influence each other over time.

7 Conclusion

The replication confirmed the original paper's main insight: when treatments are simplified into a binary indicator, it can hide important differences between treatment versions. Using DML, we reproduced this pattern and showed that grouping treatment levels can lead to misleading average effects.

The extension applied the same framework to a new setting, sleep deprivation and depression, using BRFSS data. It found that short sleep is associated with a 13-14 percentage point higher probability of depression. It also revealed meaningful heterogeneity: the effect was stronger for women, younger adults, and obese individuals. And it was mainly effect heterogeneity. These insights suggest that public health policies should consider both average effects and subgroup differences when designing sleep-related interventions.

While this study shows a strong link between sleep deprivation and depression, many questions remain. Future research should use longitudinal data to better understand the direction of the relationship, does poor sleep cause depression, or does depression reduce sleep? It would also be useful to study the effect of chronic sleep loss over time, rather than a single snapshot. Finally, including more detailed variables, like stress, income, and working hours, could help uncover additional factors that influence both sleep and mental health.

The paper by Heiler & Knaus do address a good decomposition of the treatment heterogeneity effects, but not a guideline, when it is really important necessary to apply to check for this. While we can say, it is not useful in randomised experiments, we have written down a framework to help.
If one of those questions can be answered with Yes. Researchers should think about checking for the treatment heterogeneity effect.

CHIVE – Framework
C- Choice of Treatment
Was the treatment randomly assigned, or could individuals choose or influence their treatment status?
H – Heterogeneity in Observables
Do subgroup-specific ATEs (cATEs) vary substantially across known characteristics?
I – Implicit Variation (Unobservables)
Is there variation in treatment effects that cannot be explained by observed subgroups or covariates?
V – Value of Theory of Contextual Knowledge
Is there prior evidence or theory suggesting people should react differently to this treatment?
E – Evaluation Purpose
Am I trying to understand average impacts, explain subgroup differences, or design optimal policy targeting?

References

Badr MS, e. (2015). Recommended Amount of Sleep for a Healthy Adult: A Joint Consensus. *watson*.

Connie W. Tsao , e. (22. 02 2022). Heart Disease and Stroke Statistics—2022. *AHA STATISTICAL UPDATE*.

Burghardt, J., McConnell,, S., & Schochet, P. (10. 12 2019). *openicpsr.* Von https://www.openicpsr.org/openicpsr/project/113269/version/V1/view abgerufen

Cattaneo, M. (2010). *Replication: Cattaneo .* Von https://github.com/mdcattaneo/replication-C_2010_JOE abgerufen

Heiler, P., & Knaus, M. C. (September 2022). Effect or Treatment Heterogeneity? Policy Evaluation with Aggregated and Disaggregated Treatments. *IZA Institut of Labor Economics*, S. 1-86.

Heiler, P., & Knaus, M. C. (kein Datum). *Replication Notebook - Scenario 1: Smoking and birth weight*. Von https://mcknaus.github.io/assets/code/Replication_NB_smoking.nb.html abgerufen

Heiler, P., & Knaus, M. (kein Datum). *Replication Notebook - Scenario 2: Job Corps*. Von https://mcknaus.github.io/assets/code/Replication_NB_JC.nb.html abgerufen

National Center for chronic diasease prevention. (27. 10 2022). Von 2020 BRFSS Survey Data and Documentation: https://www.cdc.gov/brfss/annual_data/annual_2020.html abgerufen

Psychiatry, C. U. (16. 03 2022). *How Sleep Deprivation Impacts Mental Health*. Von https://www.columbiapsychiatry.org/news/how-sleep-deprivation-affects-your-mental-health abgerufen

In case of challenges with R we used R Wizard:
OpenAI, several personal communication, April+ May, 2025

8 APPENDIX

APPENDIX 1A OVERVIEW OF THE KEY TERMS

Concept	Definition	Key Idea
ATE (Average Treatment Effect)	$\mathbb{E}[Y(1) - Y(0)]$	Average effect of treatment across the entire population, assuming random assignment.
nATE (natural Average Treatment Effect)	Weighted sum of effects under the observed ("natural") treatment allocation.	Captures both true effect heterogeneity and selection bias from treatment versions.
rATE (randomized Average Treatment Effect)	Weighted sum of effects assuming ra dom assignment of treatment versions (at their overall population frequencies).	Measures pure effect heterogeneity, removing selection bias.
ITT = nATE	Intention to treat effect	The average causal effect of being randomly assigned access to the Job Corps program (regardless of whether the individual actually enrolled or completed it), compared to not being offered access.
cATE (Conditional Average Treatment Effect)	$\mathbb{E}[Y(1) - Y(0) \mid X=x]$	Treatment effect for a subgroup of the population with covariates $X = x$.
Delta (Δ)	$\Delta(x) = \text{nATE}(x) - \text{rATE}(x)$	Measures how much of the heterogeneity is due to selection into different treatment versions. Large Δ = strong selection effects.
Neyman-orthogonal scores	Special moment functions where **small errors in nuisance models** (propensities, outcome models) **don't bias** the treatment effect estimator much.	Key tool for Double Machine Learning (DML): makes causal estimates robust even when nuisance models are imperfect.
Monte Carlo study	A **simulation study** where data are artificially generated many times to study the performance (bias, variance, confidence intervals) of estimators.	Used to check if methods work properly (e.g., coverage rates close to 95%).

APPENDIX 1B Details why the decomposition method holds significant importance:

1) Accurate interpretation of heterogeneity: it allows researchers and policymakers to understand why effects differ. Without this decomposition, heterogeneity driven solely by differences in the effective treatment versions received might be mistakenly interpreted as heterogeneity in the causal effect itself, potentially leading to spurious conclusions.
2) Informing policy and intervention design: by distinguishing between effect and treatment heterogeneity, the method provides crucial insights for designing more effective interventions. If heterogeneity is primarily due to rATE (effect heterogeneity), policies might focus on tailoring treatment content or delivery to different individuals. If it's largely due to Δ (treatment heterogeneity), the focus might shift to improving how different treatment versions are assigned or made accessible.

3) Evaluating treatment assignment quality: the Δ component can be particularly informative about the quality of the mechanism that assigns individuals to different effective treatments. For instance, a positive

Δ could suggest that the assignment process is directing individuals towards the effective treatments where they experience larger positive effects, compared to a random assignment scenario

4) Handling multi-valued treatments: the method is particularly valuable in settings with a large number of potentially relevant effective treatments (large J), where standard methods for analyzing multi-valued treatments face challenges related to limited data overlap and identification. The decomposed estimands (nATE, rATE, Δ) provide interpretable aggregate measures that remain estimable and allow for inference about necessary conditions for heterogeneity, even when analyzing every single effective treatment version individually is not feasible.

5) Provides summary measures: nATE, rATE, and Δ serve as interpretable summary measures for evaluating the consequences of treating a multi-valued treatment as a simple binary one and for comparing different assignment rules.

6) Data collection guidance: implementing the method requires observing the effective treatment received, highlighting the importance of designing data collection processes to capture this level of detail beyond just a binary indicator

APPENDIX 2: Additional Information and Graphs of the replication

Appendix 2A Smoking replication

8.1.1 Average treatment effect (ATE)

All comparisons with the non-smoking group (level 0) are negative and statistically significant:

Comparison	Our Replication (ATE, g)	Original Study (approx.)
1 - 0	–159.2	~–120 to –160
2 - 0	–225.6	~–200 to –230
3 - 0	–207.6	~–200
4 - 0	–275.7	~–270
5 - 0	–222.6	~–220

The largest impact is still at level 4 (–275g), and all ATEs vs. non-smokers are highly significant ($p < 0.001$).

4 - 0 | –275.7 | SE = 23.9 | $p < 0.001$ (as indicated below)

```
             ATE        SE        t         p
1 - 0 -159.2258   24.0108  -6.6314 3.410e-11 ***
2 - 0 -225.5587   17.3449 -13.0043 < 2.2e-16 ***
3 - 0 -207.6210   36.0864  -5.7534 8.872e-09 ***
4 - 0 -275.7144   23.8937 -11.5392 < 2.2e-16 ***
5 - 0 -222.5991   50.1683  -4.4370 9.169e-06 ***
2 - 1  -66.3329   29.0058  -2.2869 0.0222130 *
3 - 1  -48.3952   42.9157  -1.1277 0.2594686
4 - 1 -116.4886   33.3575  -3.4921 0.0004802 ***
5 - 1  -63.3733   55.3054  -1.1459 0.2518587
3 - 2   17.9377   39.5628   0.4534 0.6502666
4 - 2  -50.1557   28.9153  -1.7346 0.0828318 .
5 - 2    2.9595   52.7498   0.0561 0.9552586
4 - 3  -68.0934   42.8547  -1.5889 0.1120907
5 - 3  -14.9782   61.5105  -0.2435 0.8076160
5 - 4   53.1153   55.2716   0.9610 0.3365705
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As in the original study, effects between adjacent levels are small and mostly insignificant, confirming diminishing marginal effects.

```
rATE:
t test of coefficients:

          Estimate Std. Error  t value  Pr(>|t|)
zBlack    -177.490     31.823  -5.5775 2.472e-08 ***
zHispanic -243.581     65.009  -3.7469 0.0001796 ***
zOther     -10.620     46.566  -0.2281 0.8196008
zWhite    -237.292     13.327 -17.8057 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

8.1.1.1 Δ (Selection into treatment):

```
Delta:
t test of coefficients:

          Estimate Std. Error t value Pr(>|t|)
zBlack     33.7399    16.4773  2.0477  0.04061 *
zHispanic  28.9030    19.6348  1.4720  0.14103
zOther     -8.9500    12.8943 -0.6941  0.48762
zWhite      2.4022     4.5849  0.5239  0.60033
---
```

Comments (nATE, rATE, Δ (Delta))

nATE: Average effect of smoking on birthweight within each ethnic group, holding selection constant.

rATE: Actual observed difference in outcomes due to both treatment and behavioral response.

Δ (Delta): selection effect - differences in who smokes within each group that may bias rATE relative to nATE.

White mothers			White mothers experience the strongest and most consistent negative effect from smoking (around –235g). Very small and non-significant selection effect → rATE ≈ nATE → observed effect reflects true causal impact.
Estimate Type	**Estimate (g)**	**Significance**	
nATE	–234.9	***	
rATE	–237.3	***	
Δ	+2.4	Not sig.	

Black mothers			Direct effect (nATE) is moderate (-144g), but observed effect (rATE) is larger (-178g). Significant positive selection (**Δ** = +34g) suggests that Black mothers who smoke may be positively selected e.g., healthier or higher-SES smokers relative to non-smokers, slightly offsetting harm. rATE = nATE + **Δ** → 143.8 + 33.7 ≈ - 177.5
Estimate Type	**Estimate (g)**	**Significance**	
nATE	–143.8	***	
rATE	–177.5	***	
Δ	+33.7	*	

Hispanic mothers			Large negative causal effect (nATE = –215g), similar to Whites. Slightly larger observed effect (rATE = –244g), but **Δ not significant**, so **selection is not a major driver**.
Estimate Type	**Estimate (g)**	**Significance**	
nATE	–214.7	***	
rATE	–243.6	***	
Δ	+28.9	Not sig.	

Other			No significant impact of smoking on birthweight in this group. Large standard errors → results inconclusive. Could reflect small sample size or high heterogeneity.
Estimate Type	**Estimate (g)**	**Significance**	
nATE	–19.6	n.s.	
rATE	–10.6	n.s.	
Δ	–9.0	n.s.	

White mothers experience the strongest causal harm

- Baseline (intercept) nATE = –234.9g
- This is the largest (most negative) natural treatment effect in your sample.
- Minimal selection effect (Δ = +2.4g) → what you observe is very close to the true causal effect**.**

Black mothers: smaller causal effect but strong selection bias

- nATE = –143.8g → harm from smoking is still large, but weaker than for whites
- Δ = +33.7g → positive selection into treatment**:** black mothers who smoke may be systematically healthier or better off in unobserved ways, which hides part of the causal harm
- rATE = –177.5g → observed harm looks worse than the underlying causal harm

Hispanic mothers: similar harm to Whites, but selection amplifies it

- nATE = –214.7g → causal effect close to White mothers
- Δ = +28.9g → positive selection, meaning that the observed difference (rATE = –243.6g) looks even worse than the underlying causal effect
- This suggests that smoking is harmful and common among slightly more advantaged subgroups of Hispanic mothers (e.g., education, income)

Other ethnicities: no clear pattern

- nATE = –19.6g, rATE = –10.6g, Δ = –9.0g
- All effects are small and statistically insignificant
- Interpretation: unclear effect — could be due to low sample size, high variance, or actual absence of strong effect

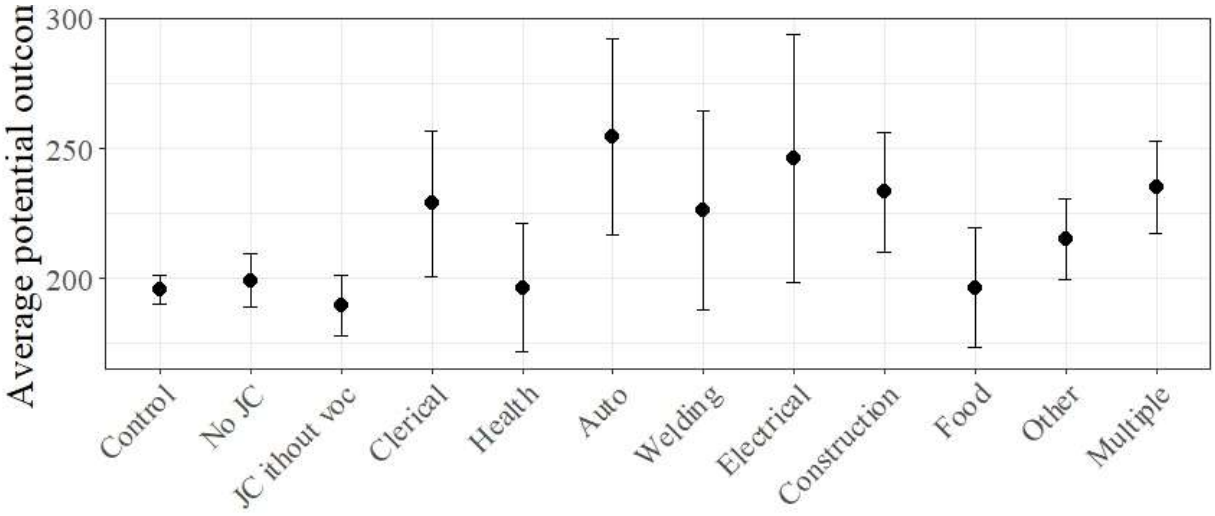
APPENDIX 2B Related to the Job Corps replication

Below it are the Average Potential Outcomes (APOs) for each treatment group - i.e., predicted average earnings in year 4 (Y4 earnings), if everyone in a group received a particular treatment. The standard errors (SE) indicate the precision of each estimate.

Group	APO (Y4 Earnings)	SE	Comment
Control	198.30	3.78	Baseline group - no Job Corps
No JC	199.25	5.30	Same as control - assigned to treatment but did not attend
JC without voc	189.51	6.00	Slightly **lower** than control - possibly due to negative selection
Clerical	219.95	9.56	Moderate gain (~+22) over control
Health	199.57	11.58	No significant difference vs. control
Auto	251.18	18.07	Strongest APO - ~+53 over control
Welding	229.81	17.04	High gain (~+31), solid vocational return
Electrical	223.83	25.55	Moderate gain but **high uncertainty** (large SE)
Construction	227.56	15.55	Substantial benefit (~+29)
Food	201.84	11.71	Modest gain (~+3), not likely significant but wide SE
Other	213.14	7.65	Clear moderate gain (~+15), with good precision
Multiple	238.53	8.92	Large gain (~+40) - strong and precise combined effect

These results align closely with the original study by Heiler & Knaus (2022), which also found that vocational tracks such as Auto, Construction, and Welding yielded the highest average earnings, while tracks like Food, Health, and No JC offered little to no improvement over the control group.

The below plot obtained from our replication displays the Average Potential Outcomes (APOs) for each treatment category in the Job Corps program:



The Y-axis shows the monthly earnings in USD and X-axis shows the different treatement versions:

- Control: no access to the treatement
- No JC: offered the treatment but did not enroll
- JC without voc: enrolled but no specific vocational training
- Vocational tracks: clerical, health, auto, welding, electrical, construction, food.

The confidence interval shows that there is some overlap, but high-return tracks like Auto and Construction are clearly above the control.

Comments:

- Control Group earns the least (~$200), confirming the baseline.
- No JC and JC without voc show a slight increase over the control — but modest.
- Clerical, Health, Food, Other tracks yield mid-to-low APOs.

- Auto, Welding, Electrical, Construction, Multiple have the highest average potential outcomes- nearing or exceeding $250–275/month.

Confidence intervals show that:

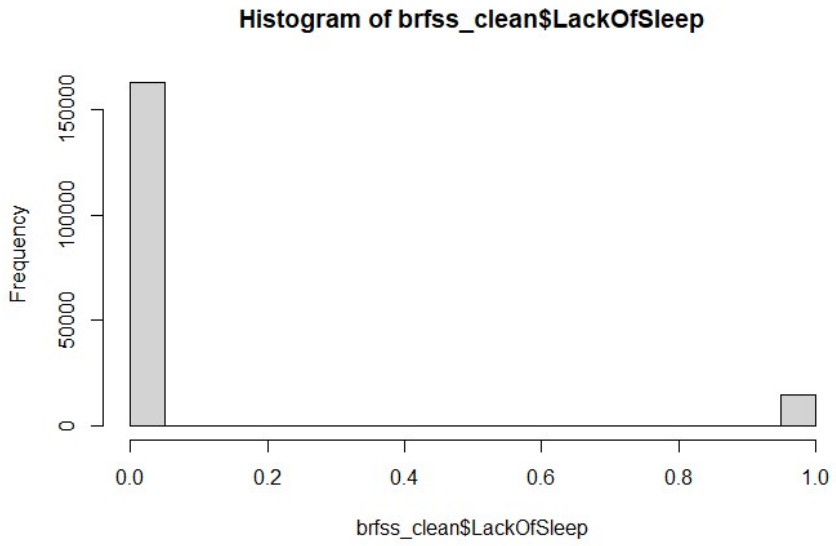
- There's some overlap, but high-return tracks like Auto and Construction are clearly above the control.
- Food and JC without voc are relatively ineffective.

APPENDIX 3: Details of the extension

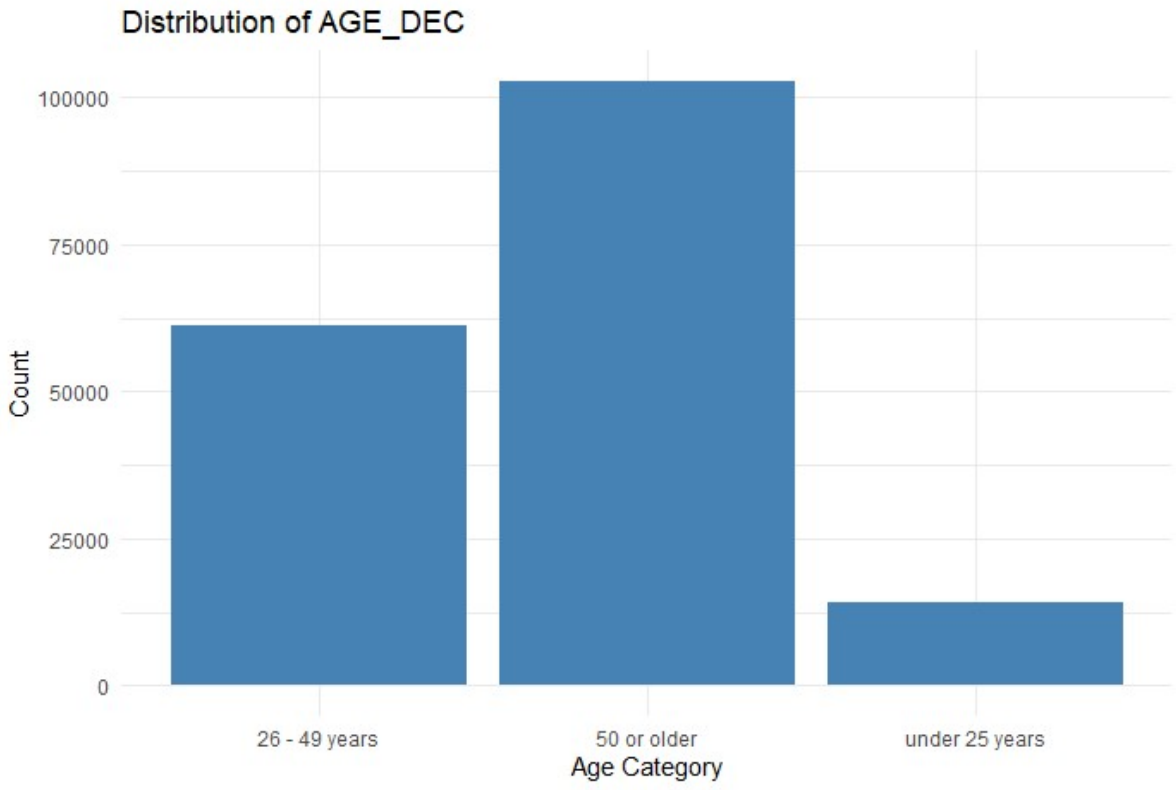
8.2 Sample comments

The histogram here below shows the distribution of the binary treatment variable `LackOfSleep`, which equals 1 if a respondent reports sleeping less than 6 hours per night, and 0 otherwise. As seen in the chart:

- Most individuals in the sample report getting at least 6 hours of sleep.
- Only a small minority (roughly 5–10%) fall into the sleep-deprived group.



This bar chart shows the age distribution of our sample (AGE_DEC) categorized into three groups:

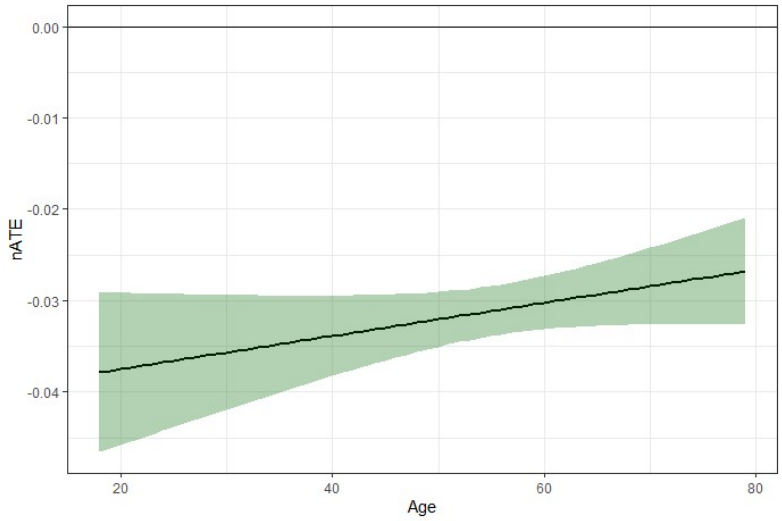


Most people in the survey are 50 years or older (100,000), then come those aged 26–49 (60,000), and very few are under 25 (30,000). This means the results mostly tell us about older adults.

Unlike the Job Corps or smoking during pregnancy studies, this one may not apply well to younger people.

This graph (1) shows that the negative effect of sleep deprivation on depression (nATE) is stronger for younger individuals. As age increases, the effect becomes slightly weaker, though it remains harmful across all ages. The shaded area (confidence interval) suggests the result is statistically significant, especially for younger adults.
We can see with the Decomposition method by Heiler & Knaus, that there might be heterogenity treatment biases for the under 25 and under 49 years old individuals. Means, that if the sleeping length as a treatment would be assigned randomly, the change of getting a depression would be even slighly better, but only slightly, therefore we could assume, that the underlying treatment heterogeinity is not very high and the effect heterogeinity with a lot of confounder is the main driver of depression probability.

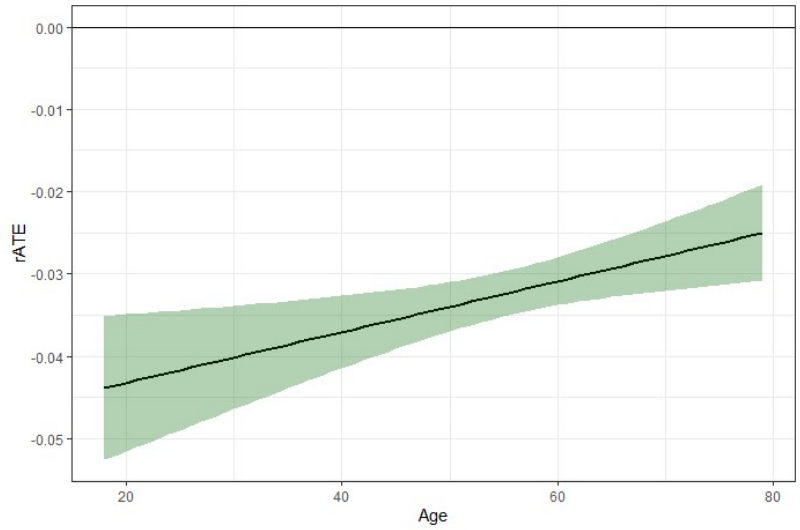
1)



Natural ATE (nATE): All age groups show a significant negative effect, meaning that sleeping less than 6 hours increases the probability of depression. The effect is strongest for the under 25 group (–3.85 percentage points), followed by 26–49 years (–3.44 pp), and **50+** (–2.97 pp).

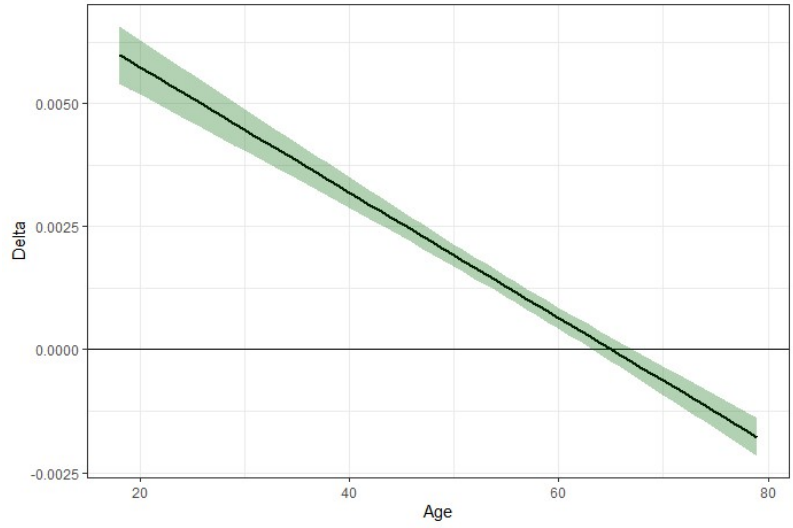
This graph below (2) shows that the **observed effect** of sleep deprivation on depression (rATE) is stronger for **younger adults** and becomes weaker with age. The trend is similar to the nATE plot, confirming that younger individuals are more vulnerable, even under hypothetical random treatment assignment. The consistent negative values suggest that insufficient sleep increases depression risk across all ages.

2)



Random ATE (rATE): These estimates reflect what the effect would be if treatment (sleep deprivation) were randomly assigned. The pattern is similar but slightly stronger for younger groups, especially under 25s.

The graphe below Delta plot (3) shows how selection into treatment (sleep deprivation) varies by age. For younger individuals, Delta is positive - suggesting they are positively selected into sleep deprivation (possibly healthier or less vulnerable), which understates the true harm. For older individuals**,** Delta becomes negative - meaning the observed effect may slightly overstate the causal harm. The sharp downward trend highlights meaningful age-related selection bias**.**



3)

Comments:

1) As we can see all models give consistent ATEs (~0.13).

This means that across different model types (ridge, lasso, forest), the estimated increase in probability of depression due to sleep deprivation is very stable - around 13 percentage points.

2) The obese subgroup has a slightly higher ATE (0.1405).

Sleep deprivation might have a stronger effect on depression for obese individuals, but the standard error is larger, likely due to smaller sample size in that group.

3) Trimming for overlap doesn't change the estimate significantly,

ATE after trimming is still around 13.3%, indicating your result is robust to potential issues in propensity score distribution