Github Repository: https://github.com/data5student/ST313-Data-Ethics

Title: Evaluating Causal Assumptions and Ethical Transparency in Observational Health Research Papers

Topic: Discussing ethical challenges in using large national health datasets and creating models to determine causal relationships, specifically by evaluating claims from the paper:

> "Exploring the impact of coffee consumption and caffeine intake on cognitive performance in older adults: a comprehensive analysis using NHANES data and gene correlation analysis",

about the relationship between coffee consumption and cognitive performance using NHANES data in the US.

Candidate Numbers: 60276, 67742, 66120

Table of Contents:

Introduction

Motivation and Ethical Context

Contemporary health research is becoming increasingly reliant on observational studies, moving away from the 'golden standard' of randomised controlled trials (RCTs) (Bajwa et al. 2021). It is not always feasible to conduct an RCT, especially in studies involving long-term dietary and lifestyle behaviours, where researchers must resort to alternative experimental designs. It is important to examine these methodologies as these observational findings can carry substantial weight; influencing public understanding, clinical recommendations, and health policies, yet often lack rigorous validation of its statistical processes. A prominent case study of this is the 1998 Wakefield MMR-autism report, which "falsely claimed causative links" between autism and vaccination, sparking an international anti-vaccination movement and an increase in preventable disease outbreaks (Flaherty, 2011). Moreover, there is a systemic problem in science with growing incentive for questionable research practices (QRP) such as: HARKing, selective reporting, p-hacking (Andrade, 2021). This can lead to publication or citation bias and the emergence of what Stark and Saltelli call "cargo cult statistics", which is the ritualistic application of statistical methods without understanding assumptions, limitations, or proper interpretation, contributing to the replication crisis (Stark & Saltelli, 2018). Whilst some cases involve deliberate fraud, even well-intentioned observational research can be ethically ambiguous due to the challenges of methodology.

The National Health and Nutrition Examination Survey (NHANES) conducted by the CDC is a commonly used database in the United States to study the relationship between lifestyle factors and long-term health outcomes. Despite the influential results, the data contains a high-dimensional set of interacting factors, many of which are correlated, noisy, incomplete, and measured with error. As such, modelling assumptions can substantially affect the predictions and interpretations of important variables, which then leads to fluctuating accuracy and results across studies.

When investigating observational data, a statistician must utilise perceptive judgement for causal inference in the multidimensional NHANES datasets which can mean a variety of conflicting assumptions across multiple studies, even within the specific topic of caffeine consumption and cognitive function. One study adjusted for sex, age, and smoking, "the most important potential confounders for coffee consumption…" (Di Maso et al., 2021), while others adjust for upward of ten variables as confounders (Li et al., 2025). Thus, there is no consensus within publications with "results from these studies conflicting, and the study designs including many possible confounders" (Mendoza et al., 2023) with "the possibility of residual and unmeasured confounding… not completely ruled out." (Wang et al., 2025). However these choices have an impact. Most noticeably: "The initially observed significant associations between coffee consumption and performance… were no longer statistically significant after additional adjustments for potential confounders" (Li et al., 2025), which highlights just how the results and conclusions are influenced by the chosen model.

Research Objective

To demonstrate the impact of methodological decisions on results and discuss the ethical implications of these choices, we will use pre-existing data with real applicable complications.

Specifically in the context of a previously published paper, we examine "*Exploring the impact of coffee consumption and caffeine intake on cognitive performance in older adults: a comprehensive analysis using NHANES data and gene correlation analysis"* by Li et al (2025). This paper studies the causal relationship between coffee consumption and cognitive performance using NHANES data. We will attempt to directly reproduce parts of it while commenting on the challenges of observational research and propose an alternative model to contrast the author's choices. This is used to evaluate whether their analytical approach satisfies guidelines in applied ethics frameworks. We aim to effectively showcase the need for transparency and highlight the challenges of maintaining ethical fairness, accuracy, and replicability in observational research.

Background & Literature Review

In this section, we establish the theoretical foundations discussed in our report and the problems and concepts we use to consider the ethical consequences of our findings.

The Role of Causality

In its most basic form, causality determines the effects of different actions or treatments and helps calculate the extent to which an action induces an observed event or pattern (Barocas et al, 2018), which in the context of health research, can inform decisions on treatment, health interventions, policies and more. Causal models such as the Directed Acyclic Graph (DAG) act as a "mechanism to incorporate scientific domain knowledge and exchange plausible assumptions for plausible conclusions" (Barocas et al, 2018). These models let us know which variables to adjust for by defining the flow of information in different categories, which we now discuss.

Confounders, Colliders, and Mediators

DAGS graphically represent the assignment structure of a given causal model, visually distinguishing between classes of variables, namely confounders, colliders, and mediators (Barocas et al, 2018). We can define these classes using a simple DAG identifying the causal effect of X on Y with an additional third variable Z.

When a variable Z has two or more directed edges going outwards, that is to say Z is a common cause of both X and Y, then Z has a confounding effect. So when analysing the effect of X on Y, Z, a confounding variable, may create false associations, so it must be adjusted to remove bias in order to estimate the true effect. However, this becomes increasingly difficult in the presence of unobserved confounders, particularly in a real-life scenario where data is noisy, correlated, and high-dimensional. Then, unrecorded or inaccurately recorded variables that we do not adjust still confound X and Y, giving less accurate results.

When Z has two direct edges going inward representing causal influence by both X and Y, Z is a collider. Alternatively, we can say X and Y are confounded if there exists a backdoor path between the two, defined as non-causal paths that allow information to pass from X to Y. This pathway shows no association between X and Y, but if we were to adjust Z we would be conditioning on a variable influenced by X and Y, thus creating collider bias; a false correlation.

If instead Z has one edge directed inward and one directed outward, we say that Z is a mediator, so the path from X to Z to Y contributes to the total effect that X has on Y. Thus, adjusting for mediating variables can hide a proportion of the true effect of X on Y.

Backdoor Criterion

To remove the bias introduced by backdoor pathways, we must select a set of variables that block every backdoor between X and Y (Barocas et al, 2018).

Formally, a set Z satisfies the backdoor criterion if:

1. There is no element or variable z in Z that is a descendant of X.
2. Z blocks all backdoor pathways from X to Y.

This is done by adjusting solely for confounders.

Replication Crisis

The replication crisis refers to the “widespread failures to reproduce published scientific results” (“Replication Crisis”, 2026), despite replication being "the cornerstone of science" (Moonesinghe et al., 2007). There are three types of replication: direct (repeating procedures as closely as possible), systematic (repeating with intentional changes), and conceptual (testing hypotheses using different procedures to assess generalizability) (“Replication Crisis”, 2026). Furthermore, reproducibility refers to the re-examination of a paper based on its data to validate its outcomes, whilst replication is the process of repeating the study with new, independent data (Replication Crisis”, 2026).

The Open Science Collaboration conducted 100 replications of studies published in 3 psychology journals and found that only 36% of replications achieved statistically significant results in the same direction as the original studies, compared to 97% of the original studies showing significant effects (Open Science Collaboration, 2015). These failures of replication occurred despite the rigorous replication of design and materials provided by the original authors, and advanced review for “methodological fidelity”, showing that the lack of replication quality could not have been the reason for the low reproducibility rate. They instead attribute the low reproducibility rate to factors including: lack of transparency and accountability, publication/outcome reporting/citation bias, etc (Open Science Collaboration, 2015).

The reduced replicability represents a systematic issue rooted in the incentive structure in science, named the “publish or perish” culture (Tran, 2024), where scientific incentives reward innovation over verification. Obviously, financial motives also play a part, with studies finding that the “pharmaceutical industry has a financial motive for suppressing unfavorable results" (Cristea et al, 2018). However, there are similar motives in areas with no profit motives such as psychotherapy (Cristea et al, 2018), showing that the problem extends beyond just monetary incentives, but within the focus on reputation, novelty, and the structure of scientific publication.

Reproducing the Paper

Within this culture of incentive-driven publications and limited reproducibility, replication is central to evaluating the reliability of empirical findings. In this section, we directly reproduce the paper's results which allows us to holistically assess the significance of the reported causal relations and provide a concrete basis for our methodological and ethical analysis.

Data & Variable Construction

Data Introduction

First, in order to reproduce the paper, we must start from the same base dataset, which we compile by pulling NHANES data from Kaggle by accessing a complete database, and then the necessary xpt files from the CDC that are available online. All the coding details and output can be found in the submitted DATA file for transparency, while in this report, we will only show the relevant sections. The report we're emulating is not very clear on the exact variables it uses and the NHANES has many different variables measuring the same thing. In parts of the report, only a few covariates are listed while indicating that more are used without specification. Because of this, we cannot be certain that our variables are an exact match. To make our own process reproducible and transparent, we have explicitly listed all variables we use, all of which are left joined on **SEQN**, an id variable.

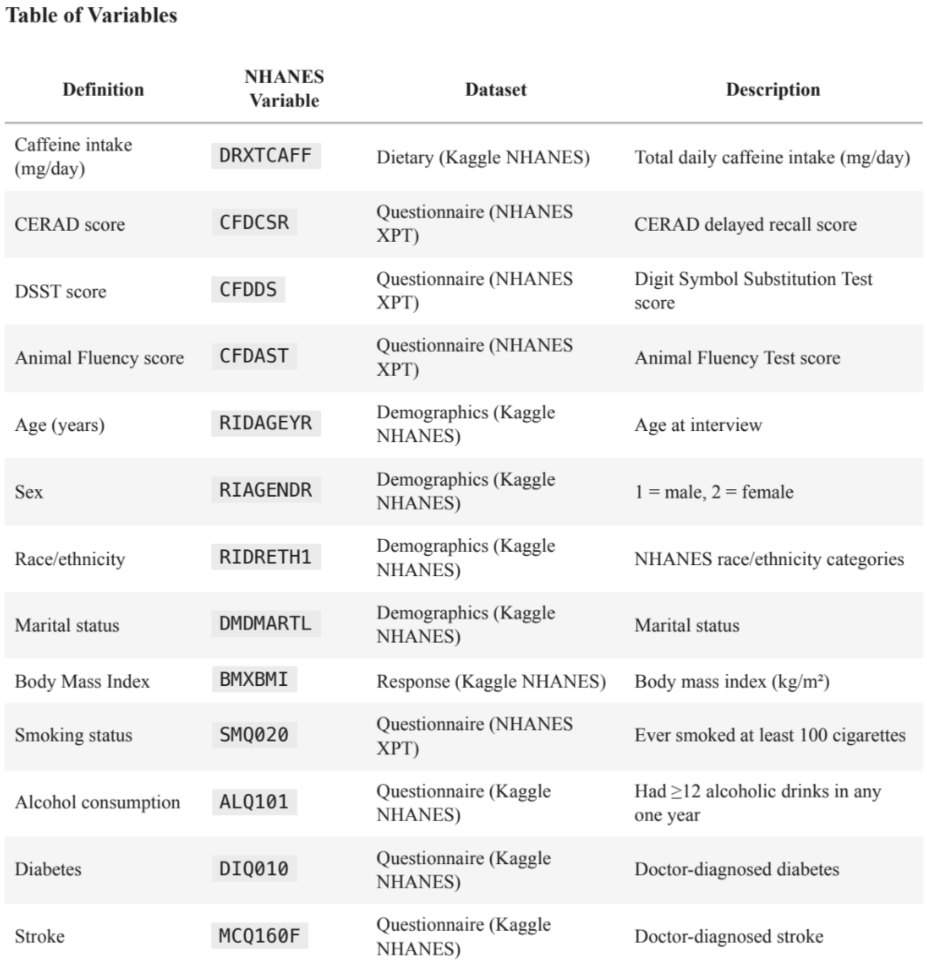
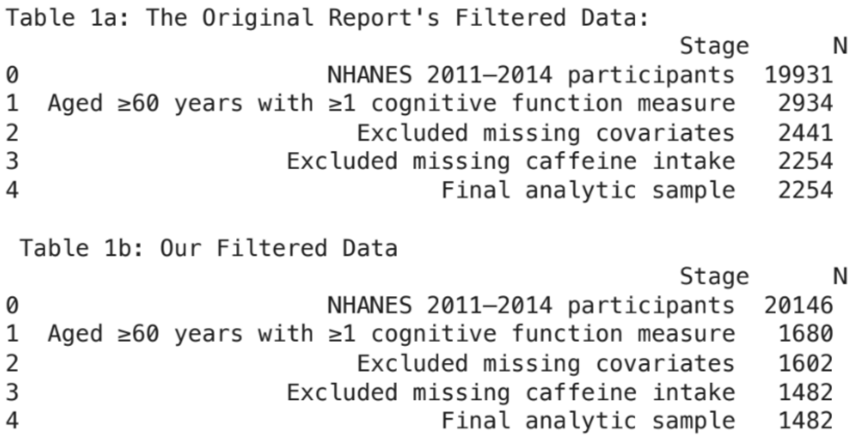


Table of Variables

Definition	NHANES Variable	Dataset	Description
Caffeine intake (mg/day)	`DRXTCAFF`	Dietary (Kaggle NHANES)	Total daily caffeine intake (mg/day)
CERAD score	`CFDCSR`	Questionnaire (NHANES XPT)	CERAD delayed recall score
DSST score	`CFDDS`	Questionnaire (NHANES XPT)	Digit Symbol Substitution Test score
Animal Fluency score	`CFDAST`	Questionnaire (NHANES XPT)	Animal Fluency Test score
Age (years)	`RIDAGEYR`	Demographics (Kaggle NHANES)	Age at interview
Sex	`RIAGENDR`	Demographics (Kaggle NHANES)	1 = male, 2 = female
Race/ethnicity	`RIDRETH1`	Demographics (Kaggle NHANES)	NHANES race/ethnicity categories
Marital status	`DMDMARTL`	Demographics (Kaggle NHANES)	Marital status
Body Mass Index	`BMXBMI`	Response (Kaggle NHANES)	Body mass index (kg/m²)
Smoking status	`SMQ020`	Questionnaire (NHANES XPT)	Ever smoked at least 100 cigarettes
Alcohol consumption	`ALQ101`	Questionnaire (Kaggle NHANES)	Had ≥12 alcoholic drinks in any one year
Diabetes	`DIQ010`	Questionnaire (Kaggle NHANES)	Doctor-diagnosed diabetes
Stroke	`MCQ160F`	Questionnaire (Kaggle NHANES)	Doctor-diagnosed stroke

Data Cleaning

The original paper uses data from 2011-2014, so we restrict the data to stay within this timeframe. Further, we collapse every table to one row per person for the chosen variables, as NHANES often has multiple observations. For some, such as caffeine intake and BMI, we use the mean of the multiple observations, following what is done in the report. Then we filter the data using the same order as the original; the results follow below.



```
Table 1a: The Original Report's Filtered Data:
                                               Stage      N
0                      NHANES 2011–2014 participants  19931
1  Aged ≥60 years with ≥1 cognitive function measure   2934
2                         Excluded missing covariates   2441
3                    Excluded missing caffeine intake   2254
4                               Final analytic sample   2254

 Table 1b: Our Filtered Data
                                               Stage      N
0                      NHANES 2011–2014 participants  20146
1  Aged ≥60 years with ≥1 cognitive function measure   1680
2                         Excluded missing covariates   1602
3                    Excluded missing caffeine intake   1482
4                               Final analytic sample   1482
```

Data Limitations

Comparing our final dataset numbers with the report shows that our data doesn't match up exactly. Especially given the number of participants from 2011-2014, one wouldn't expect there to be any discrepancy, yet we have 215 more participants, and because of this, all further filtering is not aligned. Differences in data imply that the outcome will be different as well.

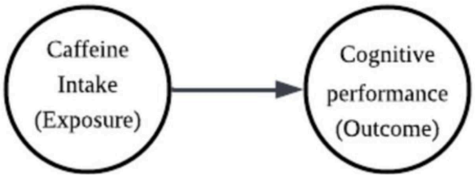
Model Replication

Model Introduction

We then apply the data to the models. There are three outcome variables, but instead of having three identical graphs for each model with different labels, we represent **CERAD, DSST,** and **Animal Fluency** tests with the label **Cognitive Performance**. We have four models, the first three are replicated from the original study, and the fourth is our own proposed model. We argue in Model 4 that some of these covariates from Model 3 may actually be colliders or mediators, leading to inappropriate adjustment. Through our argument we aim to show that there is no one true causal graph and that the assumed relationships and the choice of mediators, colliders, and confounders do in fact affect the outcome of the result. The progression in these nested models allows us to analyse how estimated effects of caffeine intake change as some variables are reclassified from confounders to mediators or excluded from having no effect.

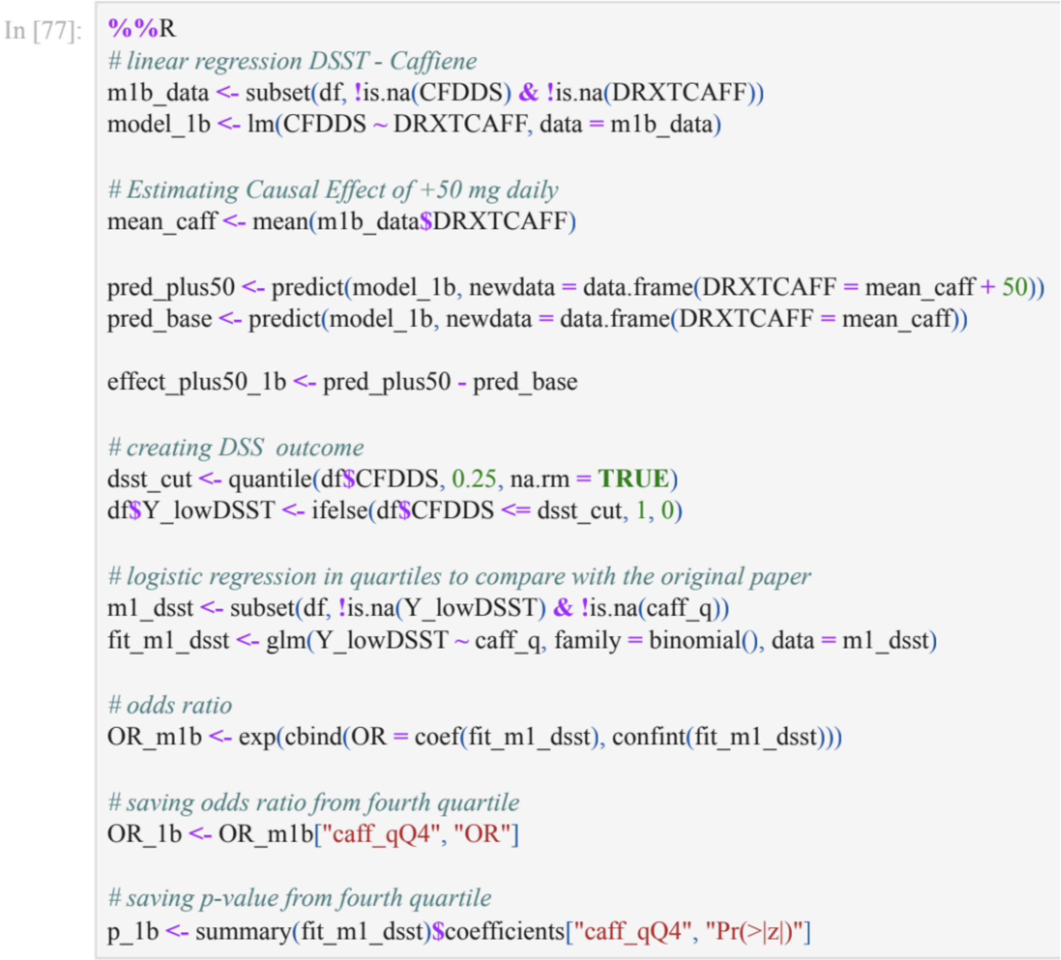
To quantify this, we apply linear regression to estimate the causal effect of increasing caffeine consumption by 50 mg daily for each cognitive performance outcome. We chose 50 mg because it is approximately half a cup of coffee, so it allows us to interpret results on a realistic marginal change. To reproduce the paper as closely as possible, we copy their technique for all models: through logistic regression, we calculate the odds ratio by labelling the bottom 25% (Quartile 1) as low cognitive performance (cognitive impairment) and compare with the top 25% (Quartile 4), as well as calculating the p-value and significance with this method. By using the same benchmarks and statistical indicators, we can compare our results with the results listed in the paper. The full code can be found in the SUPPORTING CODE Notebook submitted alongside this report. In this, we have shown the code for the first model for one outcome variable, the Digit Symbol Substitution Test Score. The rest of the code structure is the same, just with different variables for the models or a different outcome variable for cognitive function. So the regression formula and the corresponding variable subset change to reflect the different adjustment sets of each DAG. All the values are stored and displayed in a table at the end of this section.

Model 1



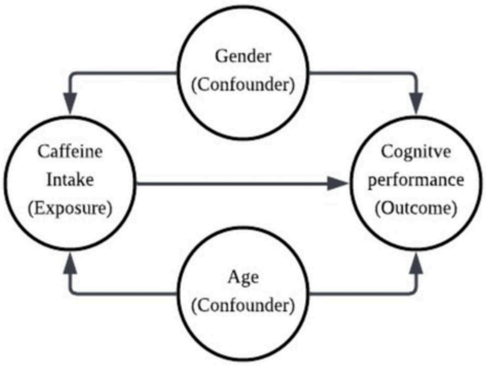
Model 1 (Unadjusted / Naive)

This model only estimates the unadjusted causal effect of caffeine intake on cognitive performance. The code below implements the first model (as shown in the DAG diagram) for the outcome variable of DSST Score. First, we remove individuals with missing DSST or caffeine intake data and then fit a linear regression. Using the fitted model, we predict the average DSST score at the mean observed caffeine intake and then compare it to the difference in DSST score if caffeine intake was to increase by 50 mg daily. This gives the estimated causal effect of caffeine intake in this model, allowing us to compare it across various models. Then, as with the original study, caffeine intake is separated into 4 quartiles with the highest and lowest quartile used to calculate the regression outcome, odds ratio and the associated p-value used to judge significance. The same code is then repeated for the two other cognitive function outcomes of Animal Fluency and the CERAD test.



```
In [77]:  %%R
          # linear regression DSST - Caffiene
          m1b_data <- subset(df, !is.na(CFDDS) & !is.na(DRXTCAFF))
          model_1b <- lm(CFDDS ~ DRXTCAFF, data = m1b_data)

          # Estimating Causal Effect of +50 mg daily
          mean_caff <- mean(m1b_data$DRXTCAFF)

          pred_plus50 <- predict(model_1b, newdata = data.frame(DRXTCAFF = mean_caff + 50))
          pred_base <- predict(model_1b, newdata = data.frame(DRXTCAFF = mean_caff))

          effect_plus50_1b <- pred_plus50 - pred_base

          # creating DSS  outcome
          dsst_cut <- quantile(df$CFDDS, 0.25, na.rm = TRUE)
          df$Y_lowDSST <- ifelse(df$CFDDS <= dsst_cut, 1, 0)

          # logistic regression in quartiles to compare with the original paper
          m1_dsst <- subset(df, !is.na(Y_lowDSST) & !is.na(caff_q))
          fit_m1_dsst <- glm(Y_lowDSST ~ caff_q, family = binomial(), data = m1_dsst)

          # odds ratio
          OR_m1b <- exp(cbind(OR = coef(fit_m1_dsst), confint(fit_m1_dsst)))

          # saving odds ratio from fourth quartile
          OR_1b <- OR_m1b["caff_qQ4", "OR"]

          # saving p-value from fourth quartile
          p_1b <- summary(fit_m1_dsst)$coefficients["caff_qQ4", "Pr(>|z|)"]
```

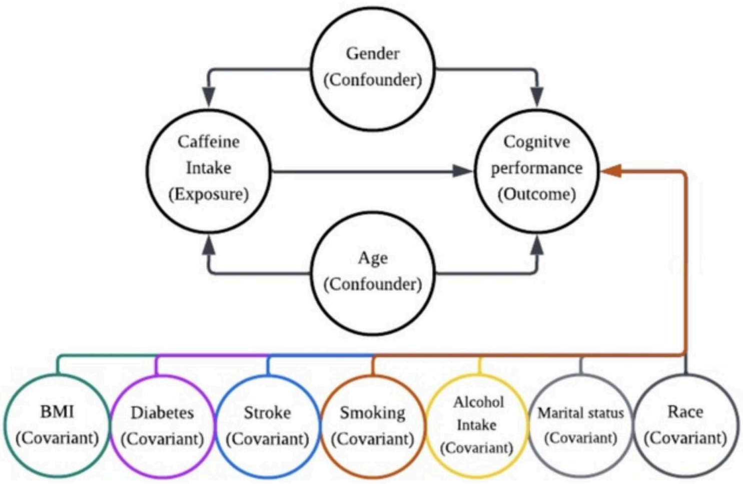
Model 2



Model 2 (Adjusted for Sex and Age)

In this model, we adjust for age and sex, treating these variables as confounders affecting both caffeine consumption and cognitive performance. In the code, we now also check that Sex and Age information is available and add these variables as additional predictors in the regression.

Model 3



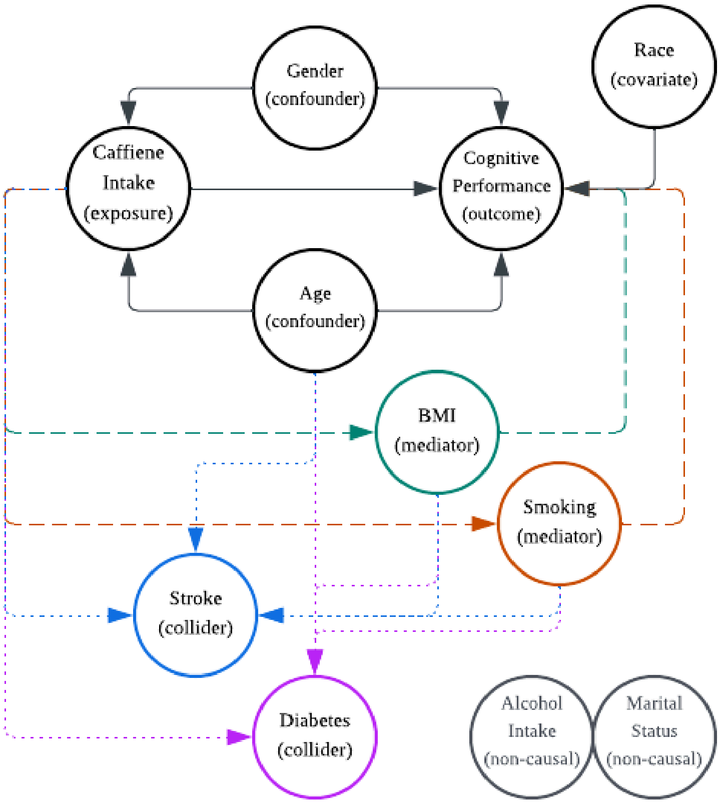
Model 3 (Adjusted for Further Covariates)

We further adjust for BMI, race, marital status, alcohol intake, smoking, stroke, and diabetes, which the original study includes to reduce residual confounding. Again, the regression formula changes to accommodate these additional variables while the rest of the code follows the same implementation as before.

Model 4 (Our Proposed Alternate Model)

In our proposed model, we deliberately change the adjustment set from the previous models to highlight our point on causal inference; there is no one true regression case, and one can commit to and justify any causal inference graph within a variety of options. Instead of using the statistical criteria, which can demonstrate local associations within specific datasets, we primarily use causal structure logic to create our DAG Model. In our case, we believe that age and sex are important confounders of caffeine intake and cognitive performance as there exists extensive literature showing the causal impact of age and gender on cognitive function due to the "differences in cognitive processing strategies" (Jockwitz et al, 2021; Dodig et al, 2020). We chose not to consider marital status and alcohol intake as confounders as we found that age and gender already act as proxies for these variables (White, 2020). We also adjust for race as a baseline covariate because we think it may confound social or behavioural patterns linked to our model (Williams et al, 2016).

Model 4



The main difference is that we chose both BMI and smoking as potential mediators because we assume that they lie on the pathway from caffeine intake to cognitive function. Increased caffeine intake could imply less concern for health so a higher BMI and more cigarettes smoked.

Similarly, we think both diabetes and stroke should be treated as colliders. Using the logic that a higher caffeine intake may mean higher chances of diabetes or stroke, and that these chances are also affected by age, BMI, and smoking. So if we treat diabetes and stroke as collider variables, we shouldn't adjust for them to avoid opening non-causal paths and inducing bias.

Our DAG model 4 shows the dashed lines for mediators and dotted lines for colliders. So the variables our code adjusts for and includes in the regression model are caffeine intake, gender, age, and race, as including the other ones would violate the backdoor criterion. With all these changes, our proposed model demonstrates how causal effect estimates can change when we reclassify variables using different plausible causal assumptions.

Reproduced Results: Comparison

The table below compares the original paper's odds ratios and significance levels with our replicated model's results. We have also included our proposed models' outcome results for completeness.

Table: Original Paper vs. Replicated Model

	Odds Ratio M1	Repli Odds Ratio M1	Sig M1	Repli Sig M1	Odds Ratio M2	Repli Odds Ratio M2	Sig M2	Repli Sig M2	Odds Ratio M3	Repli Odds Ratio M3	Sig M3	Repli Sig M3	Prop Odds Ratio M4	Prop Sig M4
Animal Fluency	**0.58**	**0.53**	**≥ 0.05**	**< 0.001**	**0.73**	**0.55**	**≥ 0.05**	**< 0.001**	**1.00**	**0.63**	**≥ 0.05**	**< 0.001**	**0.58**	**< 0.001**
CERAD	**0.40**	**0.62**	**< 0.001**	**< 0.01**	**0.49**	**0.62**	**< 0.001**	**< 0.01**	**0.58**	**0.61**	**< 0.01**	**< 0.01**	**0.58**	**< 0.01**
DSST	**0.50**	**0.42**	**< 0.001**	**< 0.001**	**0.66**	**0.42**	**< 0.05**	**< 0.001**	**1.18**	**0.41**	**≥ 0.05**	**< 0.01**	**0.37**	**< 0.001**

Table: Estimated Causal Effect of +50 mg/day Caffeine on Cognitive Performance Variables

	Model 1	Model 2	Model 3	Model 4
Animal Fluency	**0.39**	**0.27**	**0.23**	**0.26**
CERAD	**0.08**	**0.08**	**0.07**	**0.09**
DSST	**0.76**	**0.71**	**0.72**	**0.81**

For each model, we exactly followed the original statistical procedures so our replicated Models 1-3's odds ratios and significance levels should be comparable with the original report. However, these do not fully align. We were only able to validate 2 of the significance values reported in the paper (DSST for Model 1, CERAD for Model 3) and the other 7 had different results. The odds ratios reported in the paper were also significantly higher 6 out of 9 times, indicating the paper reports a stronger statistical association than our replicated result. It is uncertain whether this is due to misreporting or some other data manipulation the paper fails to mention or if it stems from having slightly different foundational data. Some slight difference is unavoidable due to the original paper's lack of detailed transparency in their methodology.

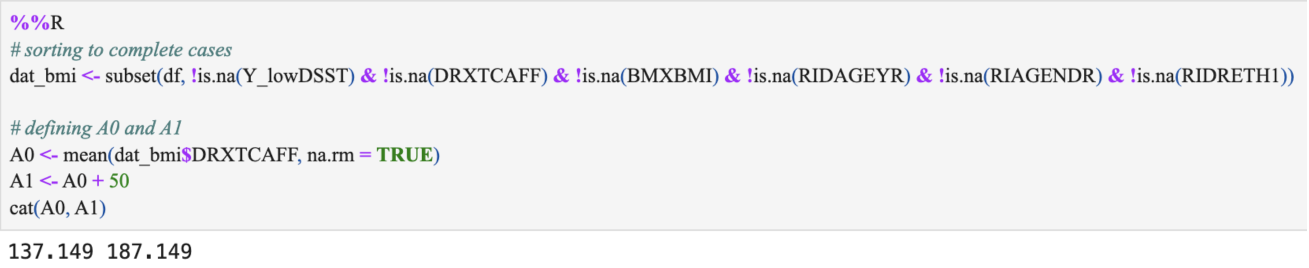
For Animal Fluency, our results show statistically significant associations while the original study reports no such effects. This divergence highlights the sensitivity of conclusions to modelling and data processing decisions. The DSST results are also easily swayed by covariate adjustment, with weakening significance as more variables are added to the model. This again draws importance to the fact that regression-based conclusions depend very strongly on the chosen adjustment set.

The estimated causal effects also reinforce this consequence. As seen in the table, the magnitude of the causal effect varies slightly across models as different variables are adjusted for. These results do not distinguish between direct and indirect pathways which makes it difficult to determine whether variables are confounders or mediators and whether adjustment of these variables distort the true causal effect. To analyse this, we can use counterfactual decomposition.

Model Evaluation with Counterfactual Causality Effects

We made assumptions in our proposed model that are not backed up by statistics or data, but we can check whether our variable reclassifications are valid within our working dataset. We can formalise the effect of some variables using the counterfactual notation for Total Effect, Controlled Direct Effect, Natural Direct Effect, and Natural Indirect Effect. These measures can help us classify certain variables by recognising their specific effects. We present the case of the variable BMI with the outcome of the Digit Symbol Substitution Test.

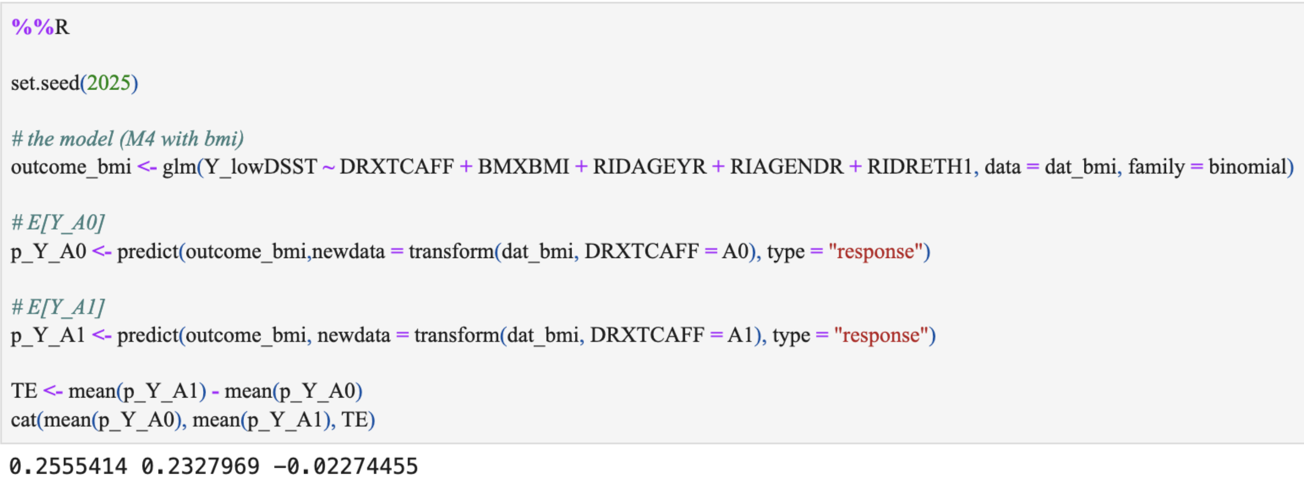
In our proposed model, we consider BMI to be a mediator in the relationship of caffeine intake to DSST score. Let X be the mediator BMI, Y the outcome DSST test, and A the caffeine intake. This gives the causal pathway of A → X → Y. Further, define A0 as the mean caffeine intake and A1 as the mean plus 50 mg of caffeine daily. Using common causal notation, Y_a is the outcome we would observe if caffeine was set to A = a and X_a is the mediator value again with caffeine intake as A = a. Let Y_ax be the outcome if caffeine was set to A = a and the mediator was forced to be x.



```
%%R
# sorting to complete cases
dat_bmi <- subset(df, !is.na(Y_lowDSST) & !is.na(DRXTCAFF) & !is.na(BMXBMI) & !is.na(RIDAGEYR) & !is.na(RIAGENDR) & !is.na(RIDRETH1))

# defining A0 and A1
A0 <- mean(dat_bmi$DRXTCAFF, na.rm = TRUE)
A1 <- A0 + 50
cat(A0, A1)
```

```
137.149 187.149
```

The ***Total Effect (TE)*** computes the expected risk if caffeine was to increase by 50 mg to the average Where TE = E[Y_A1] - E[Y_A0]. The code below gives E[Y_A1] = 0.233. E[Y_A0] = 0.256. With the total effect being -0.023. Implying that by increasing caffeine by 50mg a day, there is a 2.3 percentage point lower risk of low DSST test performance.

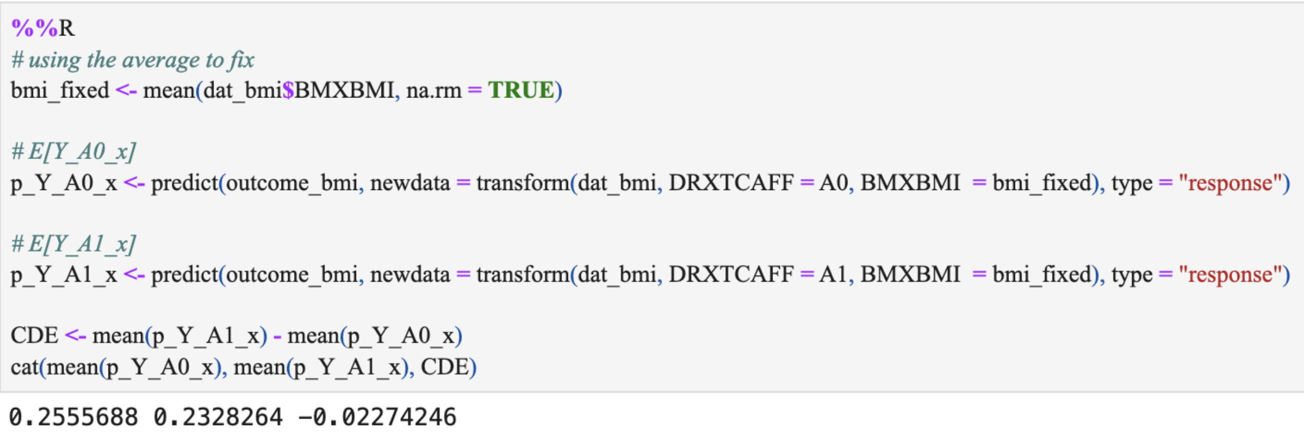


```
%%R

set.seed(2025)

# the model (M4 with bmi)
outcome_bmi <- glm(Y_lowDSST ~ DRXTCAFF + BMXBMI + RIDAGEYR + RIAGENDR + RIDRETH1, data = dat_bmi, family = binomial)

# E[Y_A0]
p_Y_A0 <- predict(outcome_bmi,newdata = transform(dat_bmi, DRXTCAFF = A0), type = "response")

# E[Y_A1]
p_Y_A1 <- predict(outcome_bmi, newdata = transform(dat_bmi, DRXTCAFF = A1), type = "response")

TE <- mean(p_Y_A1) - mean(p_Y_A0)
cat(mean(p_Y_A0), mean(p_Y_A1), TE)
```

```
0.2555414 0.2327969 -0.02274455
```

To calculate the ***Controlled Direct Effect (CDE)***, we have to fix X to a constant value, which we use as the sample mean for BMI. This allows us to isolate the direct effect of caffeine.

CDE(x) = E[Y_A1_x] - E[Y_A0_x] = -0.0227

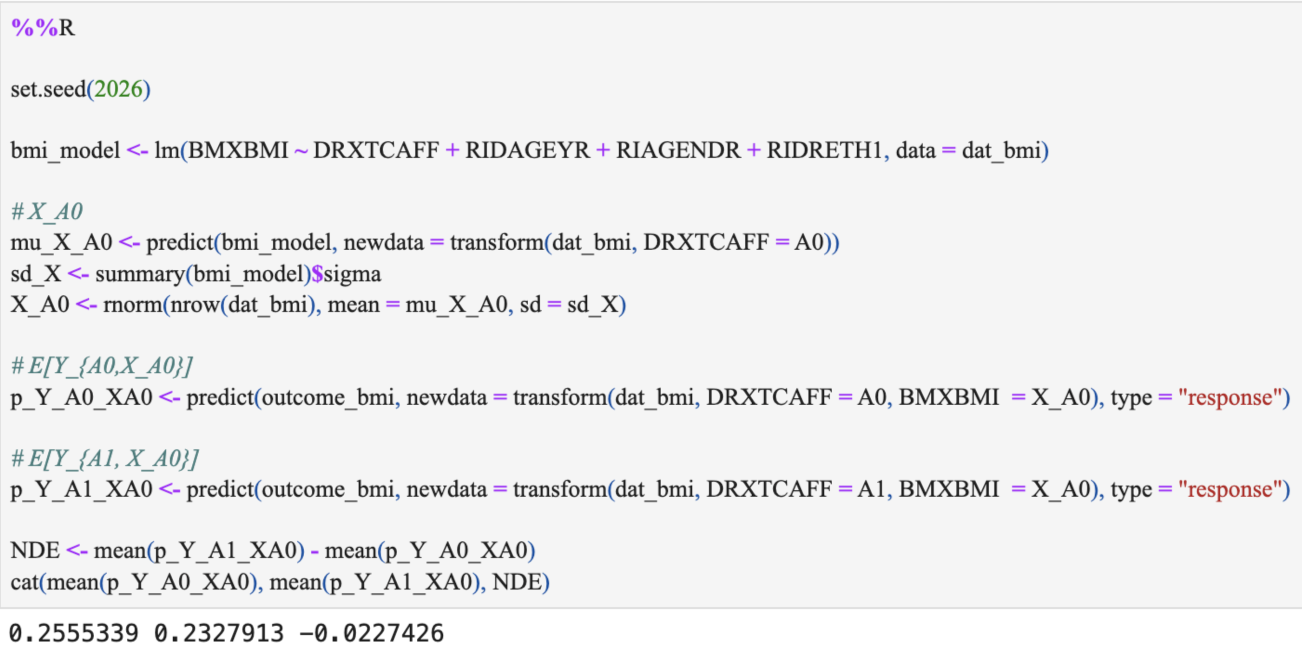


```
%%R
# using the average to fix
bmi_fixed <- mean(dat_bmi$BMXBMI, na.rm = TRUE)

# E[Y_A0_x]
p_Y_A0_x <- predict(outcome_bmi, newdata = transform(dat_bmi, DRXTCAFF = A0, BMXBMI  = bmi_fixed), type = "response")

# E[Y_A1_x]
p_Y_A1_x <- predict(outcome_bmi, newdata = transform(dat_bmi, DRXTCAFF = A1, BMXBMI  = bmi_fixed), type = "response")

CDE <- mean(p_Y_A1_x) - mean(p_Y_A0_x)
cat(mean(p_Y_A0_x), mean(p_Y_A1_x), CDE)
```

```
0.2555688 0.2328264 -0.02274246
```

The ***Natural Direct Effect (NDE)*** compares the outcome when caffeine changes but BMI is held at the natural value when A = A0 to give X_A0.

NDE = E[Y_{A1, XA0}] − E[Y{A0,X_A0}] = -0.023



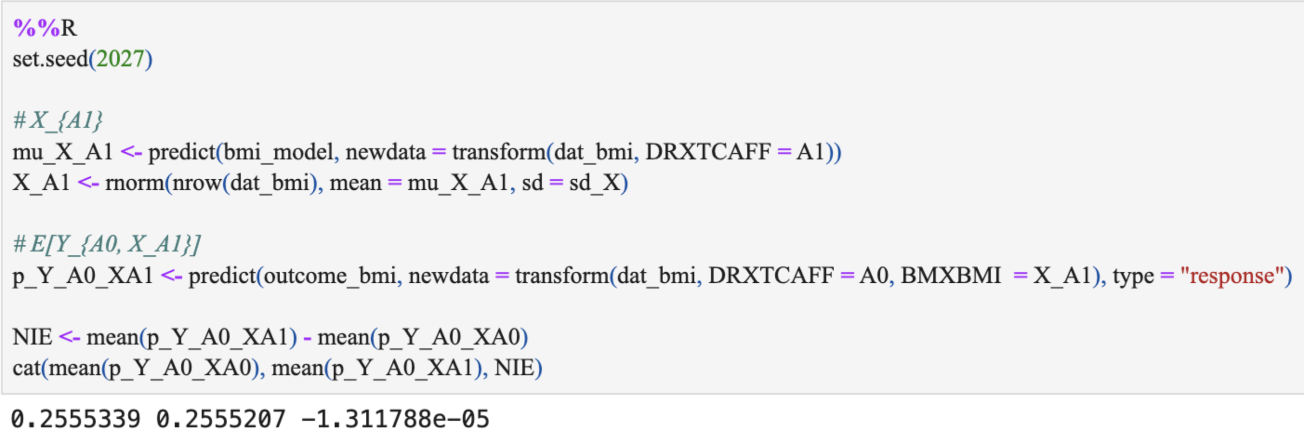
```
%%R

set.seed(2026)

bmi_model <- lm(BMXBMI ~ DRXTCAFF + RIDAGEYR + RIAGENDR + RIDRETH1, data = dat_bmi)

# X_A0
mu_X_A0 <- predict(bmi_model, newdata = transform(dat_bmi, DRXTCAFF = A0))
sd_X <- summary(bmi_model)$sigma
X_A0 <- rnorm(nrow(dat_bmi), mean = mu_X_A0, sd = sd_X)

# E[Y_{A0,X_A0}]
p_Y_A0_XA0 <- predict(outcome_bmi, newdata = transform(dat_bmi, DRXTCAFF = A0, BMXBMI  = X_A0), type = "response")

# E[Y_{A1, X_A0}]
p_Y_A1_XA0 <- predict(outcome_bmi, newdata = transform(dat_bmi, DRXTCAFF = A1, BMXBMI  = X_A0), type = "response")

NDE <- mean(p_Y_A1_XA0) - mean(p_Y_A0_XA0)
cat(mean(p_Y_A0_XA0), mean(p_Y_A1_XA0), NDE)
```

```
0.2555339 0.2327913 -0.0227426
```

The ***Natural Indirect Effect (NIE)*** fixes caffeine at A = A0, and instead compares how the output would vary when BMI shifts from using X_A0 to X_A1.

NIE = E[Y_{A0, XA1}] − E[Y{A0,X_A0}] = - 0.0013

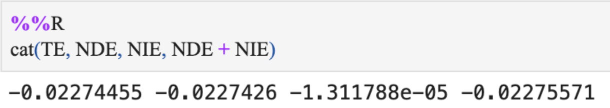
As this NIE is essentially zero, it implies that the partial effect of caffeine on DSST mediated through changes in BMI is negligible.



```
%%R
set.seed(2027)

#X_{A1}
mu_X_A1 <- predict(bmi_model, newdata = transform(dat_bmi, DRXTCAFF = A1))
X_A1 <- rnorm(nrow(dat_bmi), mean = mu_X_A1, sd = sd_X)

#E[Y_{A0, X_A1}]
p_Y_A0_XA1 <- predict(outcome_bmi, newdata = transform(dat_bmi, DRXTCAFF = A0, BMXBMI  = X_A1), type = "response")

NIE <- mean(p_Y_A0_XA1) - mean(p_Y_A0_XA0)
cat(mean(p_Y_A0_XA0), mean(p_Y_A0_XA1), NIE)
```

```
0.2555339 0.2555207 -1.311788e-05
```

The ***Decomposition Identity*** says that the total effect can be written as the sum of natural indirect and direct effect, TE = NDE + NIE. However, as we can see from the results below, the numbers don't get along :(as we only use estimates in code. However, the numbers are similar enough to tell us that we computed the values correctly.



```
%%R
cat(TE, NDE, NIE, NDE + NIE)
```

```
-0.02274455 -0.0227426 -1.311788e-05 -0.02275571
```

Overall, this code and the results show that BMI is not a meaningful mediator between caffeine consumption and DSST. For completeness, we repeat similar code for the outcomes: Animal Fluency Test and CERAD Test Score. Then we calculate the process for the variable of smoking as well, which we also classified as a mediator. The code is largely the same, just different variable names, so it is not included in this report but submitted alongside it in the SUPPORTING CODE document. The results are as follows in the table below.

Table: Counterfactual Causality Effect Results

Outcome	Mediator	TE	CDE	NDE	NIE
Animal Fluency	BMI	**-0.01949**	**-0.01950**	**-0.01949**	**-0.00001383**
CERAD	BMI	**-0.01451**	**-0.01451**	**-0.01451**	**-0.0000004074**
DSST	BMI	**-0.02274**	**-0.02274**	**-0.02274**	**-0.000001312**
Animal Fluency	Smoking	**-0.01994**	**-0.01964**	**-0.01994**	**0.00004775**
CERAD	Smoking	**-0.01563**	**-0.01525**	**-0.01565**	**0.0002514**
DSST	Smoking	**-0.02399**	**-0.02228**	**-0.02402**	**0.0001296**

For BMI, the majority of the total effect in all three outcome variables are explained by the NDE, with the NIE being a few orders of magnitude smaller than the total effect, making it virtually negligible. This is similar for smoking too, with the natural direct effect being far greater than indirect. This implies that BMI and smoking are not really meaningful mediators between caffeine and all cognitive functions. Our Model 4 might be improved if we interpret these as covariates or confounders instead of a mediator, and by mislabelling them, we may have introduced some bias in our model. However the relatively small total effect values indicate that BMI and smoking are not particularly strong confounders either. So classification here is a little bit ambiguous, making it difficult to define exact causal pathways.

This shows that even logical assumptions we make in causal models may not be reflected by the data and may even diverge within different datasets, differing from study to study and thus giving different results for causality. For more complicated models, it is not always feasible to investigate each variable so thoroughly, which makes it easier to avoid justifications of model selection and variable classification.

Ethical Analysis & Discussion

As shown in the results, attempts to reproduce the paper revealed a significant divergence between the replicated results and the published findings. The potential causes for replication failure include choices or biases introduced with the collection and handling of the data itself (pre-processing), the choices made by the researchers to determine causal effect (in-processing), and in the representation or interpretation of the results (post-processing).

We will be basing our evaluation around the American Statistical Association guidelines as this paper involves the study of participants from the US. These ethical guidelines aim to push accountability by "informing those who rely on any aspects of statistical practice of the standards they should expect" (ASA, 2022). Additionally, we will be utilising the Belmont-Menlo principles to evaluate the ethicality of decisions made in the original study.

Pre-processing

A potential area of bias is from the dataset from NHANES itself. The results for the NHANES are taken every year, from 5,000 randomly selected adults and children across the United States (CDC, 2024).

The data set is collected through a mix of:

- "Interviews about health, diet, and personal, social, and economic characteristics"
- "Visits to mobile exam centres for dental exams and health and body measurements"
- "Laboratory tests by highly trained medical professionals" (CDC, 2024).

The paper highlights that the NHANES documentation "does not give a clear specification of the time interval between the household interview and the health assessment at the Mobile Examination Center (MEC)", and that there is a lack of "standardized timeframe explicitly stated" (Li et al, 2025), meaning that there may be contradicting results from the interview and the MEC.

Moreover, one's sugar and saturated fat intake was estimated from a 24-hour dietary recall interview, bringing in potential recall bias of participants, or unrepresentative data that does not reflect one's usual dietary habits.

Additionally, the paper makes use of the variable *Caffeine Intake per day* from the NHANES data itself in order to measure the impact of coffee consumption on cognitive function specifically. However, this variable likely overstates the impact of coffee consumption as one can intake caffeine through other sources such as sodas, tea and energy drinks, etc. (Mayo Clinic, 2017). Moreover, this data does not distinguish the "coffee brand and brewing parameters" used by the coffee consuming participants, meaning that there may be "differences in the bioavailability of caffeine among different individuals" (Li et al, 2025). This prevents the paper from accurately representing the extent of caffeine intake's impact on cognitive function, which may undermine the results of the paper. It would be more accurate if the paper recorded such discrepancies, considering that caffeine level varies across different types of coffee.

Despite the issues with the data source, the paper does address consent to participation and ethics using data from the NHANES, which "was collected with participant informed consent and has been de-identified" (Li et al, 2025). This demonstrates PRINCIPLE D in the ASA guidelines (Responsibilities to Research Subjects, Data Subjects, or Those Directly Affected by Statistical Practices) (ASA, 2022) and aligns with the Belmont Menlo principle of Respect for Persons, which requires voluntary participation and informed consent (DHS, 2012). However, when applying the contextual integrity framework there may be potential privacy concerns. Whilst the participants consented to public health data collection, they may not have anticipated their information being used to make causal claims that could influence public health decisions, potentially violating the transmission principle. This shows the difficulty of addressing privacy with public data, and how deciding whether researchers have fulfilled their "Responsibilities to Research Subjects" depends on which ethical framework one applies. What is deemed appropriate secondary data use under ASA guidelines becomes more ethically ambiguous when evaluated through a contextual integrity lens of information flow appropriateness.

In-processing

Data Processing

Another source of potential bias is in the data cleaning process. The data removed participants with incomplete cognitive tests ($n = 16{,}997$), unreliable dietary recall data ($n = 187$), and missing information regarding smoking, stroke, alcohol consumption, and other variables ($n = 493$) with little to no justification on this choice. It also groups, eliminates, and simplifies data with incomplete justification, and even after following the exact data and data cleaning process of the paper, our direct reproduction gave a different number of participants. Furthermore, for the more progressed models with a greater variable selection, the number of participants dropped is greater as there's more potential for missing data, reducing the sample size and making the results less reliable than simpler models.

In the process of making their work more accessible, the authors forego details of their method and justification, making reproduction a harder process. This again showcases the value theory conflict between accessibility and accuracy. In an extreme case, the reduced sample size may even point to potential cherry-picking, which is a QRP where only favourable evidence is presented (Andrade, 2021). Moreover, justification of the choices of covariate and confounding factors are lacking (Li et al, 2025). The authors also bring emphasis to adjusting for only "known confounding variables", meaning that there are likely unobserved open backdoors in their models.

Additionally, the paper makes use of sensitive attributes in their models such as gender and race, which may make some of the causal paths unfair. Under a causal fairness perspective, racial disparities in cognitive function could reflect external structural inequality such as differences in healthcare, education etc., making race a proxy variable. Thus, adjusting for race could mask real causal mechanisms through how systemic discrimination impacts cognitive function. However, if researchers do not adjust for race, they risk attributing to caffeine intake what actually reflects biological differences racial disparities in cognitive function (Chen et al, 2021). This illustrates the value theory conflicts of accuracy and fairness.

Methodology

Further concerns arise from a theme of unjustified statistical methods and choices made by the authors throughout the paper. For example, the models use a threshold of 25%, comparing the bottom 25th percentile to the rest of the quartiles in the sample, but there is no mention of why this specific threshold was appropriate or if this threshold was chosen before or after any data analysis. Furthermore, the reason for using and comparing logistic regressions and odds ratios as opposed to other statistical approaches is entirely omitted or unclear. The opaque application of sampling weights, statistical comparison methods, and inference thresholds could produce statistically biased estimates and errors - possibly reflecting the growing cargo-cult statistics culture.

Additionally, several methodological considerations that would have strengthened confidence in the subsequent findings are not implemented in the study. The lack of standardised coffee preparation and inconsistent caffeine intake across individuals and the lack of adjustment for additives (cream, milk, sugar etc) raise meaningful questions regarding the exact causal relationship between caffeine and cognitive ability. Studies have previously found that depending on the method of coffee preparation, including changes in grind, that caffeine content per 6oz can range between 50-143mg (Bell et al, 1996). Thus, without controlling such factors, the extent of the effect that caffeine plays on cognitive ability per mg is obscured; theoretically, researchers could confuse the effects of 50mg with 100mg, producing a more extreme view of caffeine's causal nature than actuality.

Also, without an explicit DAG outlining the researchers proposed set of confounding, mediating, or colliding variables, it is difficult for readers to evaluate whether variables are accurately classified and thus appropriately adjusted. It leaves little room to debate whether any spurious associations have been formed as a result of ambiguous variable treatment. This is reinforced by the authors omitting the list of variables they determined to be confounding; "some confounding variables were accounted for” (Li et al, 2025). However, it should be emphasised that all DAG’s are (importantly) wrong. George Box famously stated “All models are wrong, but some are useful” (Box, 1976), which essentially reflects that models such as DAGs are oversimplifications of complex systems, and thus are subject to limitations and redundancy. It is not possible to account for all variables and their influences on each other within an accessible and measurable model, reflecting Occam’s razor; where accuracy is often sacrificed for accessibility (Breiman, 1986). We can however produce models that predict outcomes, and it is crucial to be explicit and transparent with the foundations said models are built upon to allow for comprehensive analysis of its predictions.

Post-processing

The authors acknowledge several limitations of the paper, but do not take into account or adjust for these issues. For example, the lack of standardised caffeine preparation and intake is noted by the authors, but no mention is made about any actions taken to counter this, or any explanation on how this may cause some variability in the results. They briefly recommend that this should be considered in future studies, but make no attempt to be explicit in how these variables should be controlled among large participant studies. Furthermore, they acknowledge that unobserved confounding variables may have impacted the results and their following interpretation; however, very little discussion is provided on these possible variables. It could be argued that the continued ambiguity across the paper weakly conflicts with principle F of the ASA guideline, responsibility to fellow statisticians (ASA, 2022). Principle F states practitioners should be constructive and advise any fellow researchers with appropriate advice for the purposes of strengthening (and not undermining) any further study, which requires transparency in data, method, and documentation (ASA, 2022).

Beyond this specific paper, it is also important to note the surge in low-quality papers based on publicly available NHANES data, with the editor of Scientific Reports receiving “nearly identical papers” every day (O’Grady, 2025). Data sets such as those made available by NHANES allow for “fresh” findings to be found simply by taking existing research and swapping in new variables (O’Grady, 2025). This allows authors to exploit data manipulation techniques, like variable combination cycling, to artificially reach significant results which are more likely to be published due to pre-existing publication biases; with indications of 190 such papers being published between 2022-2024 (O’Grady, 2025). Thus, when a paper based on large public data sets is published, it often leads to several papers utilising the same research methodology and thus inheriting its unreplicable nature, causing an exponential growth in unreplicable published studies. This is especially problematic for health science studies, as many organisations or individuals may exhibit greater levels of scepticism for published works, exacerbating lifestyle choices or medical diagnostic decisions.

Conclusion

Overall, our attempt to directly reproduce the results of the paper largely failed despite following the same dataset and processes of the paper, and our own assumptions in our proposed model conflicted with empirical results for the variables BMI and smoking. Differences in odds ratios, significance levels, and estimated effect sizes show how observational findings are often sensitive to modelling choices, data processing decisions, and assumptions. This reveals the difficulty of producing a fully accurate causal model in observational research, due to constraints in access to information, bias (both intentional and not intentional), and the inherent trade-off between accessibility and accuracy. This is particularly the case for health research, where variables impacting lifestyle habits and health outcomes are often multidimensional and collinear.

To mitigate these fairness and replication issues, several approaches merit consideration, such as adopting a stricter adherence to ethical guidelines. This would allow for greater success when reproducing results as many ethical guidelines (in this case the ASA guidelines) explicitly advocate for increased transparency of data handling processes and statistical methodology. This is especially significant since ethical considerations regarding and among studies using NHANES data is severely lacking, with a study finding just 11% of papers reporting a submission to an institutional review board, and over 75% having no mention of an ethics review at all (Brock, 2021). Additionally, distinct differences across institutional review boards warrant the possible need for standardised criteria and ethical review processes to ensure all papers are held to the same standard.

Improving replication success could also be achieved by removing the existing pressure and publication incentives for researchers to produce studies with significant findings, which may be possibly done by data manipulation techniques. This can be combatted by requiring the preregistration of studies and their intended methodology, and incentivising journals to approve papers for publication before results are curated. This would disincentivise researchers from participating in practices that would otherwise produce artificially significant but unreplicable results.

In combination, these approaches would help to alleviate the conflict between the intrinsic and instrumental values that exist within the field of health studies. Ultimately, this case study illustrates a larger problem within academia. From a pragmatic ethics stand point, progress towards reliable health research requires a fundamental paradigm shift, where we value transparency over novelty, and the scientist over the shoe clerk.

References

About NHANES | National Health and Nutrition Examination Survey | CDC. https://www.cdc.gov/nchs/nhanes/about/index.html. Accessed 30 Jan. 2026.

'All Models Are Wrong'. *Wikipedia*, 26 Jan. 2026. *Wikipedia*, https://en.wikipedia.org/w/index.php?title=All_models_are_wrong&oldid=1334865642.

Andrade. *HARKing, Cherry-Picking, P-Hacking, Fishing Expeditions, and Data Dredging and Mining as Questionable Research Practices - PubMed*. 2021, https://pubmed.ncbi.nlm.nih.gov/33999541/.

Bajwa, Sukhminder Jit Singh, et al. 'The Increasing Trend of Observational Studies in Clinical Research: Have We Forgotten and Started Defying the Hierarchy?' *Indian Journal of Anaesthesia*, vol. 65, no. 3, Mar. 2021, pp. 186–90. *PubMed Central*, https://doi.org/10.4103/ija.IJA_176_21.

Barocas et al. *Causality*. 2018, https://fairmlbook.org/causal.html.

Bell, Leonard N., et al. 'Caffeine Content in Coffee as Influenced by Grinding and Brewing Techniques'. *Food Research International*, vol. 29, no. 8, Dec. 1996, pp. 785–89. *ScienceDirect*, https://doi.org/10.1016/S0963-9969(97)00002-1.

Breiman. *Statistical Modeling: The Two Cultures* -. 2001, https://www-jstor-org.lse.idm.oclc.org/stable/2676681?sid=primo.

Brock, Lydia, et al. 'Ethical Approval among Studies Using the National Health and Nutrition and Examination Survey (NHANES): A Cross-Sectional Analysis: Oklahoma State University Center for Health Sciences Research Days 2021'. 2021, p. 26.

Chen, Ruijia, et al. 'Racial Disparities in Cognitive Function Among Middle-Aged and Older Adults: The Roles of Cumulative Stress Exposures Across the Life Course'. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, vol. 77, no. 2, Apr. 2021, pp. 357–64. *PubMed Central*, https://doi.org/10.1093/gerona/glab099.

Comprehensive Data List. https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx. Accessed 30 Jan. 2026.

David, Daniel, et al. 'Why Cognitive Behavioral Therapy Is the Current Gold Standard of Psychotherapy'. *Frontiers in Psychiatry*, vol. 9, 2018, p. 4. *PubMed*, https://doi.org/10.3389/fpsyt.2018.00004.

Di Maso, Matteo, et al. 'Caffeinated Coffee Consumption and Health Outcomes in the US Population: A Dose–Response Meta-Analysis and Estimation of Disease Cases and Deaths Avoided'. *Advances in Nutrition*, vol. 12, no. 4, Jul. 2021, pp. 1160–76. *ScienceDirect*, https://doi.org/10.1093/advances/nmaa177.

Dodig et al. *The Effect of Age and Gender on Cognitive and Psychomotor Abilities Measured by Computerized Series Tests: A Cross-Sectional Study - PMC*. 2024, https://pmc.ncbi.nlm.nih.gov/articles/PMC7230412/.

‘Estimating the Reproducibility of Psychological Science’. *Science*. *www.science.org*, https://www.science.org/doi/10.1126/science.aac4716. Accessed 30 Jan. 2026.

‘Ethical Guidelines for Statistical Practice’. *Default*, https://www.amstat.org/your-career/ethical-guidelines-for-statistical-practice. Accessed 30 Jan. 2026.

Flaherty. *The Vaccine-Autism Connection: A Public Health Crisis Caused by Unethical Medical Practices and Fraudulent Science*. 2011, https://www.researchgate.net/publication/51643118_The_Vaccine-Autism_Connection_A_Public_Health_Crisis_Caused_by_Unethical_Medical_Practices_and_Fraudulent_Science.

‘How Much Caffeine Is in Your Cup?’ *Mayo Clinic*, https://www.mayoclinic.org/healthy-lifestyle/nutrition-and-healthy-eating/in-depth/caffeine/art-20049372. Accessed 30 Jan. 2026.

Jockwitz, C., et al. ‘Cognitive Profiles in Older Males and Females’. *Scientific Reports*, vol. 11, no. 1, Mar. 2021, p. 6524. *www.nature.com*, https://doi.org/10.1038/s41598-021-84134-8.

Li, Jinrui, et al. ‘Exploring the Impact of Coffee Consumption and Caffeine Intake on Cognitive Performance in Older Adults: A Comprehensive Analysis Using NHANES Data and Gene Correlation Analysis’. *Nutrition Journal*, vol. 24, no. 1, Jul. 2025, p. 102. *Springer Link*, https://doi.org/10.1186/s12937-025-01173-x.

Mendoza, Michael F., et al. ‘Impact of Coffee Consumption on Cardiovascular Health’. *The Ochsner Journal*, vol. 23, no. 2, 2023, pp. 152–58. *PubMed Central*, https://doi.org/10.31486/toj.22.0073.

Moonesinghe et al. *Most Published Research Findings Are False—But a Little Replication Goes a Long Way - PMC*. 2007, https://pmc.ncbi.nlm.nih.gov/articles/PMC1808082/.

National Health and Nutrition Examination Survey. https://www.kaggle.com/datasets/cdc/national-health-and-nutrition-examination-survey. Accessed 30 Jan. 2026.

O’Grady. *Low-Quality Papers Are Surging by Exploiting Public Data Sets and AI | Science | AAAS*. 2025, https://www.science.org/content/article/low-quality-papers-are-surging-exploiting-public-data-sets-and-ai.

Replication Crisis - Wikipedia. https://en.wikipedia.org/wiki/Replication_crisis. Accessed 30 Jan. 2026.

Saltelli, Philip B. Stark and Andrea. ‘Cargo-Cult Statistics and Scientific Crisis’. *Significance Magazine*, 5 Jul. 2018, https://significancemagazine.com/cargo-cult-statistics-and-scientific-crisis/.

The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research.

Tran. *The 'Publish or Perish' Mentality Is Fuelling Research Paper Retractions – and Undermining Science - International Science Council*. 2024, https://council.science/blog/publish-or-perish-mentality/.

Wang et al. *Coffee Drinking Timing and Mortality in US Adults | European Heart Journal | Oxford Academic*. 1 Aug. 2025, https://academic.oup.com/eurheartj/article/46/8/749/7928425?login=true#505321042.

Wang, Xuan, et al. 'Coffee Drinking Timing and Mortality in US Adults'. *European Heart Journal*, vol. 46, no. 8, Feb. 2025, pp. 749–59. *Silverchair*, https://doi.org/10.1093/eurheartj/ehae871.

White, Aaron M. 'Gender Differences in the Epidemiology of Alcohol Use and Related Harms in the United States'. *Alcohol Research : Current Reviews*, vol. 40, no. 2, Oct. 2020, p. 01. *PubMed Central*, https://doi.org/10.35946/arcr.v40.2.01.

Williams, David R., et al. 'Understanding Associations between Race, Socioeconomic Status and Health: Patterns and Prospects'. *Health Psychology : Official Journal of the Division of Health Psychology, American Psychological Association*, vol. 35, no. 4, Apr. 2016, pp. 407–11. *PubMed Central*, https://doi.org/10.1037/hea0000242.