# Caffeine Consumption and Its Association with Key Health Biomarkers

Danielle, Beckett, and Alexia

## Use of AI tools

We used ChatGPT to help conduct a test on the linearity of the logit assumption for logistic regressions and to create cleaner tables for our logistic regression models.

## Introduction

Caffeine is the most commonly consumed drug worldwide. Despite its effectiveness as a stimulant, caffeine can cause both psychological and physiological effects on the body (Samoggia & Rezzaghi, 2021). The drug is associated with antagonistic effects on adenosine receptors, which regulate stress, arousal, and fatigue (Reichert et al., 2022). As a result, scientific inquiry has increasingly focused on the potential health impacts of caffeine.

Previous research has found increases in heart rate and blood pressure following frequent daily consumption (Green & Suls, 1996). Additionally, some studies have suggested an association between chronic caffeine use and other addictions, such as nicotine and alcohol. However, other scholars argue that no causal relationship exists between caffeine consumption and the use of substances like alcohol, nicotine, or cannabis. Evidence indicates that any observed causal relationships may not reflect direct phenotypical associations but instead stem from shared underlying liabilities or disadvantages (Verweij et al., 2018). These findings suggest that relationships between caffeine, alcohol use, and smoking may extend beyond purely genetic factors.

Although caffeine consumption is widespread, differences in intake and physiological effects have been documented across demographic groups. Consumption habits vary based on geographical origin, culture, and socioeconomic status (Reyes & Cornelis, 2018). These factors influence preferred brewing methods and beverage choices, resulting in substantial variation in caffeine content per standard drink across countries. For example, an individual would need to consume approximately five cups of black tea to match the caffeine content of a standard

American coffee ("Caffeine in Tea vs. Coffee: How Do They Compare?", n.d.). Western countries tend to consume more caffeine overall than many Eastern nations. Further discrepancies are seen across age and sex, with women aged 50–64 consuming the highest levels of caffeine (Supplements et al., 2014; Temple & Ziegler, 2011).

However, few studies have simultaneously examined caffeine's effects on BMI, heart rate, stress, and sleep while accounting for demographic differences. The goal of this research is to further examine the effects of high caffeine intake on key biomarkers: BMI, heart rate, sleep quality, and stress levels. It also aims to identify differences in consumption and health outcomes across demographic variables such as gender, age, and nationality. A clearer understanding of caffeine's impact on health may inform public health practices and guidelines. Identifying which groups are most at risk could help target interventions and encourage individuals to monitor their consumption before negative effects occur.

The present study uses the "Global Coffee Health Dataset," which contains 10,000 synthetic records. The dataset includes "patients" from 20 countries and provides information on daily coffee intake, caffeine levels, sleep duration and quality, BMI, heart rate, stress, physical activity, health issues, occupation, smoking, and alcohol consumption. Its aim is to simulate real-world data on caffeine consumption, sleep, and health outcomes to support wellness research using logistic analysis and predictive modeling.

## Methodology

To show the differences in patient demographics across caffeine consumption habits, distributions of participants by country, gender, and age were considered. For the categorical variables, country and gender, a bar chart was used to visualize counts. For age, a continuous variable, a histogram was used to plot the overall numeric values of different variables, showing a frequency distribution of values, which groups data into continuous intervals. On the other hand, a bar chart was used for the categorical variables, as their length/height represents the quantity of each variable, rather than a type of distribution. Therefore, each respective visualization represents both the continuous and categorical patient demographics in the dataset.
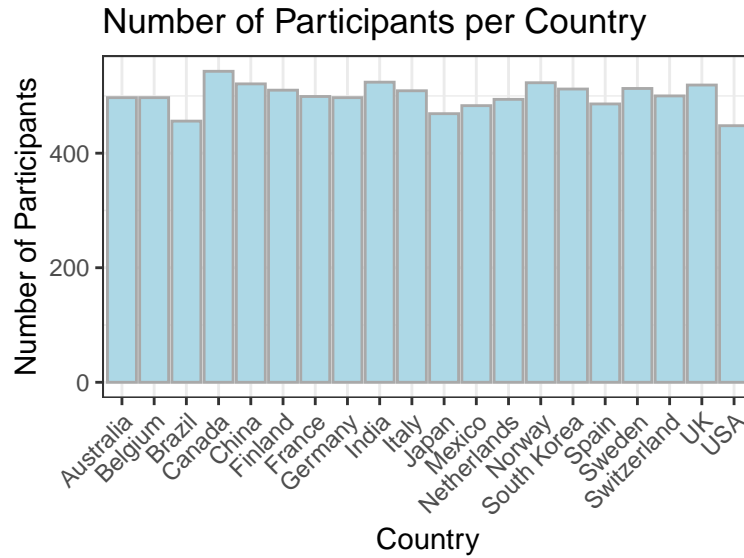
Figure 1: Distribution of Participants by Country

Overall, Figure 1 shows that the distribution is fairly balanced, with most countries contributing a similar number of participants, approximately between the mid-400s to low-500s. A few countries show slightly higher participation, while others are slightly below the average, but no extreme outliers are visible. This suggests that the sample is relatively evenly distributed across the included countries.
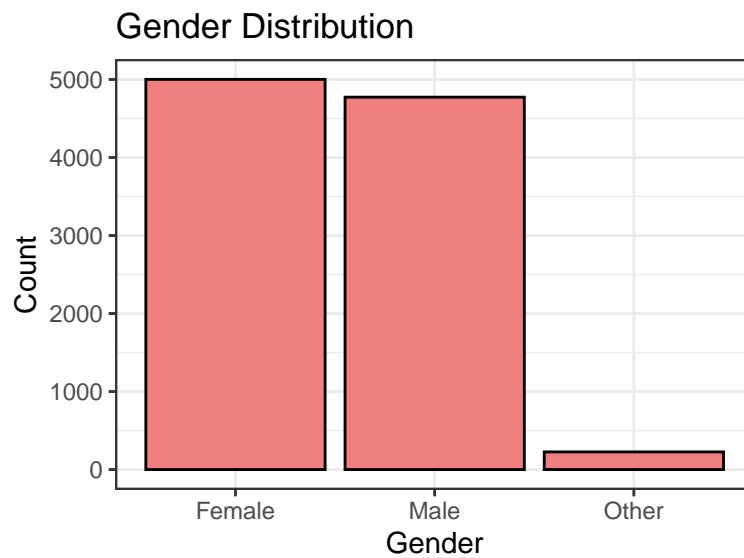


Figure 2: Distribution of Participants' Gender

Figure 2 shows the number of participants identifying as Female, Male, or Other. The majority of participants identify as either Female or Male, with both groups contributing roughly similar counts (around 5,000 each). A much smaller portion of the sample identifies as Other (~250), indicating limited representation in that category. Overall, the figure highlights a largely binary gender distribution within the participant pool.
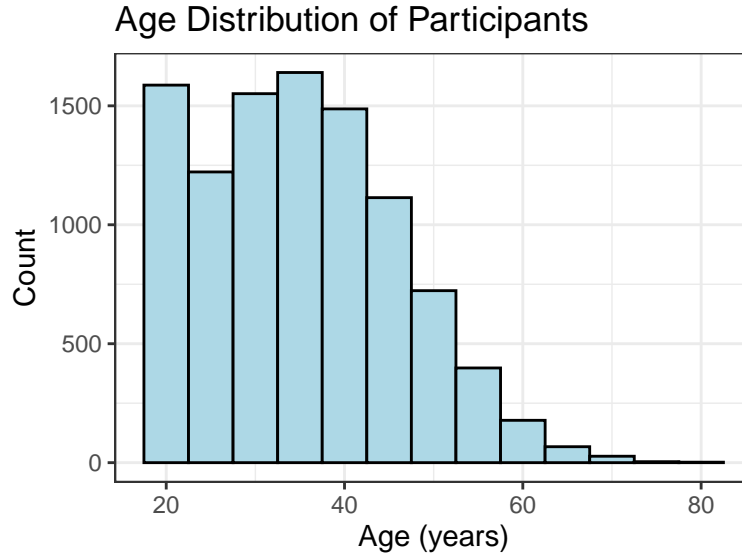


Figure 3: Distribution of Participants' Age

Figure 3 displays the distribution of participants' ages, showing that most individuals fall between their early 20s and mid-40s. The largest concentrations appear around the late 20s to late 30s, while the participation declines steadily after about age 45, with very few individuals above age 60. Overall, the age distribution is right-skewed, with its tail extending into older ages.

Logistic regression was chosen because the outcome variable is binary (high vs. low caffeine intake). To build our logistic regression model, we found our binary outcome variable based on FDA guidance indicating that daily intake up to 400 mg is associated with minimal negative effects (Commissioner, 2024). This study defined high caffeine intake as above 400 mg and low caffeine intake as below 400 mg, which then set the threshold for the binary variable accordingly.

–> Insert Logsitic Models

This study fit logistic regression models to examine the association between high caffeine consumption and four predictors: BMI, heart rate, sleep quality, and stress level. Logistic regression was appropriate because the outcome variable—high caffeine intake—was binary, defined using the FDA threshold of 400 mg/day. All variables used in the models contained no

missing values. The baseline for outcome was considered "low caffeine" intake <400 mg/day. The baseline for stress level was "high," and the baseline for sleep quality was "Excellent"

When an initial model including both sleep quality and stress level was fit, two stress-level coefficients were not estimable due to perfect or near-perfect collinearity between the two categorical predictors (Appendix A). Because This violated the assumption of no perfect multicollinearity, we fit two separate logistic regression models: Model 1 included BMI, heart rate, and sleep quality, while Model 2 included BMI, heart rate, and stress level. This approach allowed each variable to be evaluated independently without violating assumptions of logistic regression.

Our null hypothesis when considering our regression was, when holding the non target variables (BMI, sleep quality, stress level, or heart rate) constant, the slope parameter is 0; therefore, there is no relationship between the specified targeted variable and outcome. Our alternative hypothesis when considering our regression was, when holding the non target variables (BMI, sleep quality, stress level, or heart rate) constant, the slope parameter is not 0; therefore, there is relationship between the specified targeted variable and outcome. We are considering statistical significance at $\alpha$ =0.05.

–> Insert table of Null and Alternative

It is also important to analyze the assumptions that accompany these tests: linearity of log odds and independence of observations. Although the data is synthetic, this doesn't prove that either assumptions are valid. For independence to be met, each observation must provide unique data that is not influenced by another observation. In the real world of clinical trials, this kind of data would come from taking biomarkers of two patients who have no overlapping lifestyles, where neither of these patients has the opportunity to affect the data of the other. However, an example of non-independent data would be if the data was taken from siblings. In this study, we assume the former.Linearity of the log-odds was tested using the Box–Tidwell procedure, which showed no evidence of violation (Appendix B).The methodology provided is replicable and assumes that the requirement of a binary outcome variable, followed by a logistic regression model, can predict (with limitations due to uncertainty of assumptions verification) how levels of caffeine consumption relate to various health markers.

**Results**

Table 1 presents a logistic regression examining the association between high caffeine intake and sleep quality, including BMI and heart rate as additional predictors.

| Variable | Estimate | P-value |
|---|---|---|
| (Intercept) | -3.31 | $1.85 \times 10^{-26}$ |
| BMI | -0.000289 | 0.971 |
| Heart Rate | 0.00910 | 0.00333 |
| Sleep Quality: Fair | 0.981 | $8.75 \times 10^{-15}$ |
| Sleep Quality: Good | 0.612 | $2.47 \times 10^{-7}$ |
| Sleep Quality: Poor | 1.33 | $1.67 \times 10^{-22}$ |

Table 1: Logistic Regression of High Caffeine Intake (with Sleep Quality Predictor)

Table 2 shows a logistic regression examining the association between high caffeine intake and stress level, with BMI and heart rate as additional predictors.

| Variable | Estimate | P-value |
|---|---|---|
| (Intercept) | -1.98 | $5.22 \times 10^{-11}$ |
| BMI | -0.00124 | 0.873 |
| Heart Rate | 0.00944 | 0.00230 |
| Stress Level: Low | -0.816 | $< 2 \times 10^{-16}$ |
| Stress Level: Medium | -0.351 | 0.000459 |

Table 2: Logistic Regression of High Caffeine Intake (with Stress Level Predictor)

–> Add the Odds ratio and CL stuff (fix the odds ratio) –> Add baseline references Table 1's and Table 2's results indicate that there is sufficient evidence that several physiological and behavioral factors have a significant relationship with high caffeine consumption, defined as intake above 400 mg per day. In both Table 1 and 2, heart rate emerged as a significant positive predictor (p=0.002), with each unit increase in heart rate multiplying the participant's predicted odds of high caffeine intake by $e^{-0.00944} = 0.99$, while holding all other variables constant. Sleep quality in Table 1 also showed strong associations, where individuals reporting fair, good, or poor sleep quality had statistically significantly higher odds of high caffeine consumption compared to the reference group, as reflected by the log odds coefficients and highly significant p-values (all greatly less than 0.05), holding all other variables constant. Stress level had a similar relationship in Table 2; when we hold all other variables constant, both low and medium stress categories significantly predict the odds of high caffeine intake relative to the high-stress reference group (both p < 0.001 with patient's predicted odds of high caffeine intake multiplied by $e^{(-0.81)} = 2.26$ and $e^{(-0.351)} = 1.42$, respectively). In contrast, BMI was not a significant predictor (p = 0.97) in either Table 1 or Table 2, indicating there was insufficient evidence to support a statistically significant relationship with high caffeine consumption in this dataset. Overall, the results show that caffeine consumption has a statistically significant relationship with sleep quality, stress, and heart rate, while BMI had insufficient evidence to support such relationship.

<—In Discussion you say you were unable to do this, but do not fully explain why or tie it to the original goal. ## Discussion The primary goal of this study was to evaluate how caffeine consumption relates to key health markers—BMI, heart rate, sleep quality, and stress level—and to determine whether these relationships differ across demographic groups. Our logistic regression results partially address these goals. We successfully identified several physiological and behavioral variables that predict high caffeine intake, but we were unable to evaluate demographic differences due to issues with multicollinearity and category imbalance.

Our logistic regression models effectively identify the health and lifestyle factors associated with high caffeine consumption. Heart rate, sleep quality, and stress level consistently emerged as significant predictors of consuming more than 400 mg of caffeine per day. These results suggest that caffeine intake is not uniform across individuals but is linked to physiological and psychological states, particularly stress and sleep patterns. In contrast, there was insufficient evidence of BMI being a significant predictor, indicating no significant association with caffeine consumption in this dataset. The GLM outputs show that higher resting heart rates are associated with increased odds of high caffeine intake, consistent with known physiological responses to stimulants. Additionally, the significant effects of sleep quality and stress level reinforce the conclusion that caffeine consumption is closely related to sleep disturbances and heightened stress.

However, our models do not fully address our secondary goal of identifying demographic patterns. Variables such as age, gender, and nationality could not be included due to multicollinearity and the structure of the synthetic dataset. As a result, we were unable to determine whether certain demographic groups are at greater risk of consuming excessive caffeine or experiencing related health consequences.

Logistic regression was an appropriate starting point for predicting a binary threshold of caffeine intake. However, several limitations emerged. Category imbalance and redundancy between variables caused singularities in the model, preventing certain levels of sleep quality and stress from being included simultaneously. Alternative approaches, such as regularized logistic regression (LASSO or ridge), decision trees, or random forest models, could better handle correlated predictors and complex interactions (Wohlwend, 2023). Additionally, future work done in assessing multicollinearity formally using variance inflation factors would strengthen the reliability of model selection.

Data quality also introduces constraints on validity. Because the dataset consists of synthetic health records, the associations identified may not fully reflect real-world variability in caffeine use or health outcomes. Synthetic data often smooths over noise, reducing natural variance between demographic groups and producing artificial linear relationships (Melton, 2025). In addition, due to being synthetic data, contextual information about the caffeine intake, like type of caffeinated beverage, timing of consumption can not be found though they may influence caffeine's physiological effects.

If this study were conducted again in the future, several improvements would strengthen the analysis. First, incorporating real-world observational data would significantly improve exter-

nal validity and correlations. Second, including additional variables, such as socioeconomic status, occupation, genetics, and chronic health conditions, would offer a more comprehensive understanding of caffeine's impact as well as allowing us to conclude stronger relationships.

Overall, our findings suggest that high caffeine consumption is strongly associated with higher heart rate, lower sleep quality, and heightened stress, but not BMI. While the results partially meet the goals of the study, limitations in data quality and model constrain the conclusions that can be drawn. Future studies employing larger, real datasets and more flexible modeling techniques would better capture the behavioral and physiological mechanisms underlying high caffeine intake.

## Sources

Caffeine in Tea vs. Coffee: How Do They Compare? (n.d.). Retrieved November 20, 2025, from https://www.healthline.com/nutrition/caffeine-in-tea-vs-coffee#caffeine-concerns

Commissioner, O. of the. (2024). Spilling the Beans: How Much Caffeine is Too Much? FDA. https://www.fda.gov/consumers/consumer-updates/spilling-beans-how-much-caffeine-too-much

Green, P. J., & Suls, J. (1996). The effects of caffeine on ambulatory blood pressure, heart rate, and mood in coffee drinkers. Journal of Behavioral Medicine, 19(2), 111–128. https://doi.org/10.1007/BF01857602

Melton, A. (2025). Synthetic Data in Healthcare: When It Works & When It Fails. Invene.com. https://www.invene.com/blog/synthetic-data-healthcare

Reichert, C. F., Deboer, T., & Landolt, H. (2022). Adenosine, caffeine, and sleep–wake regulation: State of the science and perspectives. Journal of Sleep Research, 31(4), e13597. https://doi.org/10.1111/jsr.13597

Reyes, C. M., & Cornelis, M. C. (2018). Caffeine in the Diet: Country-Level Consumption and Guidelines. Nutrients, 10(11), 1772. https://doi.org/10.3390/nu10111772

Samoggia, A., & Rezzaghi, T. (2021). The Consumption of Caffeine-Containing Products to Enhance Sports Performance: An Application of an Extended Model of the Theory of Planned Behavior. Nutrients, 13(2), 344. https://doi.org/10.3390/nu13020344

Supplements, P. C. for a W. on P. H. H. A. with C. of C. in F. and D., Board, F. and N., Policy, B. on H. S., & Medicine, I. of. (2014). Intake and Exposure to Caffeine. In Caffeine in Food and Dietary Supplements: Examining Safety: Workshop Summary. National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK202226/

Temple, J. L., & Ziegler, A. M. (2011). Gender Differences in Subjective and Physiological Responses to Caffeine and the Role of Steroid Hormones. Journal of Caffeine Research, 1(1), 41–48. https://doi.org/10.1089/jcr.2011.0005

Verweij, K. J. H., Treur, J. L., & Vink, J. M. (2018). Investigating causal associations between use of nicotine, alcohol, caffeine and cannabis: A two-sample bidirectional Mendelian randomization study. Addiction (Abingdon, England), 113(7), 1333–1338. https://doi.org/10.1111/add.14154

Wohlwend, B. (2023, July 14). Three Regression Models for Data Science: Linear Regression, Lasso Regression, and Ridge Regression. Medium. https://medium.com/@brandon93.w/three-regression-models-for-data-science-linear-regression-lasso-regression-and-ridge-regression-6aac73c0d7a5

## Appendix A

This study fit two logistic regression models examining the associations between high caffeine consumption and four predictors: BMI, sleep quality, stress level, and heart rate. All variables had no missing values and therefore required no imputation. An initial model including all four predictors was attempted:

| Variable | Estimate | P-value |
|---|---|---|
| (Intercept) | -3.31 | $< 2 \times 10^{-16}$ |
| BMI | -0.0003 | 0.971 |
| Heart Rate | 0.009 | 0.00333 |
| Sleep Quality: Fair | 0.981 | $8.75 \times 10^{-15}$ |
| Sleep Quality: Good | 0.612 | $2.47 \times 10^{-7}$ |
| Sleep Quality: Poor | 1.332 | $< 2 \times 10^{-16}$ |
| Stress Level: Low | NA | NA |
| Stress Level: Medium | NA | NA |

Table 3: Logistic Regression of High Caffeine Intake (with All Predictors)

However, Table 3 indicates that two coefficient estimates were not defined due to singularities, specifically for the Low and Medium categories of stress level. This suggests perfect or near-perfect collinearity between the categorical predictors (sleep quality and stress level), preventing the model from estimating all parameters. Because the full model could not be reliably estimated, the analysis was separated into two logistic regression models: one including sleep quality (but excluding stress level), and one including stress level (but excluding sleep quality). This approach resolved the singularity issue and allowed each predictor to be evaluated without collinearity constraints.

## Appendix B

The test for the linearity of log odds can only be conducted on continuous variables, (e.g., BMI, Heart Rate). Categorical variables (Sleep_Quality, Stress_Level) can't be tested. Categorical

variables are converted to dummy variables automatically in R if they are factors, which we confirmed they were using the factor function. The model estimates log-odds differences relative to the reference, so there is no linearity assumption because each category is treated separately. To test the linearity of the logit assumption in logistic regression in R, we ran a Box–Tidwell test (logit linearity test).

| Variable | Estimate | P-value |
|---|---|---|
| (Intercept) | -4.27 | 0.148 |
| BMI | 0.061 | 0.837 |
| BMI_log | -0.015 | 0.836 |
| Heart Rate | 0.055 | 0.759 |
| HR_log | -0.009 | 0.798 |
| Sleep Quality: Fair | 0.981 | $9.0 \times 10^{-15}$ |
| Sleep Quality: Good | 0.612 | $2.5 \times 10^{-7}$ |
| Sleep Quality: Poor | 1.33 | $< 2 \times 10^{-16}$ |

Table 4: Box–Tidwell Test for Logit Linearity Test

Table 4 shows that the p-values for the BMI_log and HR_log terms (0.836 and 0.798, respectively) both exceed 0.05. This indicates insufficient evidence of non-linearity, and therefore the linearity assumption is considered to be satisfied.