TU DORTMUND UNIVERSITY

Introductory Case Studies

Project 3: Contingency Tables

Aspirin, the Miracle Drug?

Lecturers:

Prof. Dr. Paul Bürkner
Dr. Daniel Habermann
Lars Kühmichel
Svenja Jedhoff

Author: Sourav Poddar

Matriculation No: 236625

Group Number: 11

Group Members:
Dia Tasneem Naser
Apu Kumar Saha
Nandita Chakrobortty

November 25, 2025

Contents

1 Introduction

Aspirin is one of the most widely used medications in the world,(Cancer Association of South Africa 2021) renowned for its ability to relieve pain, reduce fever, and lower inflammation. Beyond these common uses, it has also been investigated for its potential to prevent serious cardiovascular events like heart attacks. This project focuses on a key question in public health: Is it worth taking aspirin regularly to prevent a heart attack?

To answer this, we analyze data from an unpublished five-year study that began in 1993 (Bürkner et al. 1993). This large-scale study involved 20,021 male participants who were randomly assigned to take either an aspirin or a placebo tablet every two days. The study was double-blind, meaning neither the participants nor the researchers knew who was receiving which treatment, which helps to prevent bias in the results.

The core of our analysis involves comparing the frequency of heart attacks between the aspirin and placebo groups. We begin with a descriptive analysis, calculating key metrics like relative frequencies, the risk ratio, and the odds ratio to quantify the relationship between aspirin intake and heart attack risk. We also calculate a confidence interval to understand the precision of our estimates. Following this, we create a contingency table and perform a formal statistical test (the Chi-Square test) to determine whether the observed difference in heart attack rates is statistically significant.

Our initial analysis strongly suggests that aspirin has a protective effect. The risk of a heart attack appears to be substantially lower in the group taking aspirin. Section 2 states the objectives and briefly describes the data. Summarises the statistical concepts used in the analysis explained in Section 3. In Section 4, presents the empirical results, including overall and stratified effect estimates as well as hypothesis tests. Sections 5 concludes with a short discussion of the findings and possible limitations.

2 Problem Definition

2.1 Project Objectives

The primary goal of this project is to evaluate the effectiveness of regular aspirin intake in preventing heart attacks. Based on the provided dataset, we aim to answer the question: "Does taking aspirin regularly significantly reduce the risk of having a heart attack?" To achieve this, we have defined the following statistical objectives:

First, we will conduct a descriptive analysis to summarize the distribution of participants across the treatment (Bayesball 2023) groups and heart attack outcomes, providing an initial overview of the data. Following this, we will quantify the relationship between aspirin intake and heart attack risk by calculating and interpreting key epidemiological measures, including the relative frequencies, risk ratio, and odds ratio. To assess the precision of our estimated risk ratio, we will specify a suitable confidence interval. Finally, we will formally test for a statistically significant association by creating a contingency table and applying the Chi-Square test of independence, which will allow us to determine if the observed difference in heart attack rates between the aspirin and placebo groups is unlikely to be due to chance alone.(Moore et al. 2009)

2.2 Data Description

For our analysis, we worked with the dataset **Aspirin**. The data comes from a randomized controlled trial (RCT), which is considered the gold standard for clinical research. A total of 20,021 men were randomly assigned to one of two groups:

Placebo Group: 10,034 participants
Aspirin Group: 9,987 participants

The study was conducted over five years and was double-blind. The dataset includes the following variables relevant to our analysis:

- `Group`: A categorical variable indicating the treatment assignment (Placebo / Aspirin).
- `Heart`: A categorical variable indicating whether the participant had a heart attack during the study (No / Yes).
- `Smoker`: Indicates the smoking status of the participant (Yes/No).
- `Age`: A numerical variable representing the participant's age in years.

There are no missing values for these key variables or no duplicate observations, and the random assignment ensures that the groups are comparable at the start of the study, which allows us to make causal inferences about the effect of aspirin.

3 Statistical Methods

All analyses were carried out using the R programming language (R Core Team 2024) in RStudio. Standard packages such as `readr`, `dplyr`, `ggplot2`, and `janitor` were used for data import, cleaning, summaries, and graphics. Effect measures and confidence intervals for 2×2 tables were computed with functions from epidemiological helper packages (e.g. `epiR` / `epitools`), while base R functions were used for chi-square tests and Fisher's exact test.

Below, the main statistical concepts used in the project are briefly recalled.

3.1 Contingency Tables and Relative Frequencies

A contingency table (or cross-tabulation) is a fundamental tool for analyzing the relationship between two categorical variables. It displays the frequency distribution of the variables .(Agresti 2009) In our case, we use a 2 × 2 table to show the counts of heart attacks (Yes/No) for each treatment group (Aspirin/Placebo).

From this table, we calculate *relative frequencies*, which express the proportion of participants in each group who experienced a heart attack. For example, the relative frequency of a heart attack in the Aspirin group is calculated as:

$$\text{Relative Frequency}_{\text{Aspirin}} = \frac{\text{Number of heart attacks in Aspirin group}}{\text{Total number in Aspirin group}}.$$

Relative frequencies are often expressed as percentages.

3.2 Risk and Risk Ratio

The risk of an event in a given group is simply the probability of observing this event in that group. To compare the risks between two groups we use the *risk ratio* (relative risk).(Altman 1990) If $\hat{R}_{\text{asp}}$ is the estimated risk in the aspirin (exposed) group and $\hat{R}_{\text{pla}}$ is the risk in the placebo (unexposed) group, the risk ratio is

$$\widehat{RR} = \frac{\hat{R}_{\text{asp}}}{\hat{R}_{\text{pla}}}.$$

Interpretation:

- $\widehat{RR} = 1$: no difference in risk.
- $\widehat{RR} < 1$: aspirin group has a *lower* risk of heart attack.
- $\widehat{RR} > 1$: aspirin group has a *higher* risk.

We will also use the *risk difference*

$$\widehat{RD} = \hat{R}_{\text{asp}} - \hat{R}_{\text{pla}},$$

which shows the absolute change in risk. The inverse of the risk difference gives the *number needed to treat* (NNT) (Porta 2014):

$$\widehat{\text{NNT}} = \frac{1}{|\widehat{RD}|},$$

interpreted as the approximate number of people that would need to take aspirin to prevent one additional heart attack, compared with placebo.

3.3 Odds and odds ratio

The odds of an event in a group are defined as

$$\text{Odds} = \frac{P(\text{event})}{1 - P(\text{event})}.$$

In our context, the odds of a heart attack in the aspirin group equal

$$\text{Odds}_{\text{asp}} = \frac{\hat{R}_{\text{asp}}}{1 - \hat{R}_{\text{asp}}},$$

and analogously for the placebo group.

The *odds ratio (OR)* compares these odds:

$$\widehat{OR} = \frac{\text{Odds}_{\text{asp}}}{\text{Odds}_{\text{pla}}}.$$

Alternatively, for a 2 × 2 table with cell counts

	Heart attack	No heart attack
Aspirin	a	b
Placebo	c	d

the odds ratio can be computed as

$$\widehat{OR} = \frac{a \cdot d}{b \cdot c}.$$

Again, values below 1 indicate a protective effect of aspirin, values above 1 indicate a harmful effect, and 1 corresponds to no association.

3.4 Confidence intervals

A *confidence interval* (CI) provides a range of values that is plausible for an unknown population parameter, such as the true risk ratio (Altman 1990). A 95% confidence interval for a parameter θ has the form

$$CI_{95\%} = \big(\hat{\theta} - z_{0.975} \cdot SE(\hat{\theta}), \hat{\theta} + z_{0.975} \cdot SE(\hat{\theta})\big),$$

where $\hat{\theta}$ is an estimate from the sample, $SE(\hat{\theta})$ is its standard error, and $z_{0.975} \approx 1.96$ is the critical value from the standard normal distribution.

For risk ratios and odds ratios, it is more convenient to work on the log scale. For example, a CI for the risk ratio is obtained via

$$\log(\widehat{RR}) \pm 1.96 \cdot SE\big(\log(\widehat{RR})\big),$$

and then exponentiating the bounds. If the CI for RR or OR does not include 1, this indicates a statistically significant difference at the 5% level.

3.5 Hypothesis Testing

Hypothesis testing (Lehmann and Romano 2005) is used to make inferences about the population based on sample data.

- **Null Hypothesis** (H_0): There is no association between the treatment group (Aspirin vs. Placebo) and the incidence of heart attacks ($RR = 1$ or $OR = 1$ or $RD = 0$).
- **Alternative Hypothesis** (H_1): There is a significant association between the treatment group and the incidence of heart attacks.

To test this, the following tests are used:

- **Chi-Square (χ^2) Test**: Measures the discrepancy between observed and expected frequencies in a contingency table (Pearson 1900). A large χ^2 value (and small p-value) suggests a significant association, leading to the rejection of H_0.

  $$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i},$$

  where O_i are the observed cell counts and E_i the expected counts under H_0. For a 2×2 table, the test has 1 degree of freedom. A large chi-square statistic with a small p-value provides evidence against H_0.
- **Fisher's Exact Test**: An alternative used when sample sizes are small or expected cell frequencies are low (though often used as a conservative check in larger samples).(Fisher 1922)

4 Statistical Analysis

4.1 Data quality and descriptive analysis

4.1.1 Group Distribution

We began by examining the overall distribution of the variables. As shown in Table 1, the participants were almost evenly split between the Placebo (50.12%) and Aspirin (49.88%) groups.

Table 1: Treatment Group Distribution

Group	Count	Percent
Placebo	10034	50.12%
Aspirin	9987	49.88%

Regarding the outcome, the vast majority of participants did not have a heart attack during the study.As illustrated in Figure 1 and summarized in Table 2, only 791 out of 20,021 participants (3.95%) experienced a heart attack.

Table 2: Heart Attack Distribution

Heart Attack	Count	Percent
No	19230	96.05%
Yes	791	3.95%

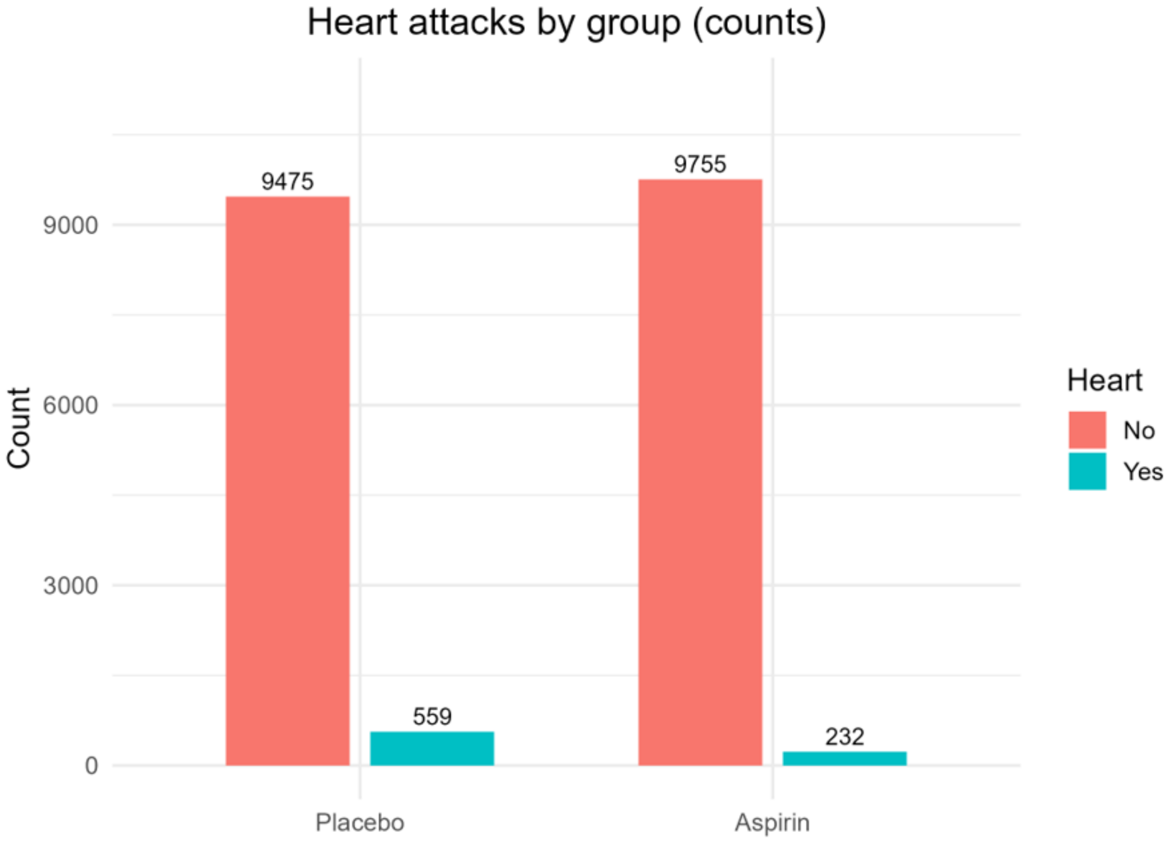

Figure 1: Bar plot of heart attack incidence across all participants.

4.1.2 Age Distribution

Table 3 provides the summary statistics for the Age variable, offering insights into the age distribution of the study participants.

Participants are all in a narrow age range around 63 years. This focus on an older male population is reasonable because the risk of heart attack increases with age, so enough events can be observed to detect a treatment effect.

Table 3: Summary Description of *Age*

Statistic	Count	Mean	Std Dev	Min	25%	Median	75%	Max
Overall	20021	63.0086	1.4105	61.0	62.0	63.0	64.0	65.0

4.1.3 Smoking status

Smoking is an important cardiovascular risk factor and is therefore included in the analysis. Table 4 shows the distribution of smokers and non-smokers

Table 4: Smoking Status Distribution

Smoker	Count	Proportion	Percentage
No	15 992	0.799	79.9%
Yes	4 029	0.201	20.12%

Around 20% of the participants are smokers and 80% are non-smokers. This provides enough observations in both groups to investigate whether the effect of aspirin differs by smoking status.

4.2 Relative frequencies of heart attacks

We first conducted an overall analysis ignoring smoking status. The core of this analysis is the 2x2 contingency table. From the Figure 1,At first glance, participants in the aspirin group have fewer heart attacks than those in the placebo group: 232 vs. 559 events, despite almost equal group sizes.

Table 5: Contingency Table of Heart Attacks by Treatment Group

Group	No heart attack	Heart attack	Total
Aspirin	9 755	232	9 987
Placebo	9 475	559	10 034
Total	19 230	791	20 021

From the Table 5, we calculated the *empirical risks* (relative frequencies) of heart attack in each group:

- Risk in aspirin group:

$$\hat{R}_{\text{asp}} = \frac{232}{9\,987} \approx 0.0232 \quad (2.32\%),$$

- Risk in placebo group:

$$\hat{R}_{\text{pla}} = \frac{559}{10\,034} \approx 0.0557 \quad (5.57\%).$$

The proportion without heart attack is higher in the aspirin group (97.7% vs. 94.4%), while the proportion with heart attack is less than half as large (2.3% vs. 5.6%).So in this sample, men taking aspirin had a heart attack in about 2.3% of cases, while in the placebo group the incidence was about 5.6%. This already hints at a strong preventive effect of aspirin.

4.3 Effect Measures

4.3.1 Risk Ratio and Risk Difference

Using the relative frequencies above, the *risk ratio* for heart attack comparing aspirin to placebo is

$$\widehat{RR} = \frac{\hat{R}_{\text{asp}}}{\hat{R}_{\text{pla}}} = \frac{0.0232}{0.0557} \approx 0.42.$$

This means that the risk of heart attack in the aspirin group is roughly 42% of the risk in the placebo group. Equivalently, aspirin users have around 58% lower risk of heart attack than those taking placebo.

The *risk difference* is

$$\widehat{RD} = \hat{R}_{\text{asp}} - \hat{R}_{\text{pla}} \approx 0.0232 - 0.0557 = -0.0325.$$

So aspirin reduces the absolute heart attack risk by about 3.3 percentage points over five years.

From this we can compute the *number needed to treat*:

$$\widehat{\text{NNT}} = \frac{1}{|\widehat{RD}|} \approx \frac{1}{0.0325} \approx 31.$$

We can Interpret this, about 31 men would need to take aspirin for five years to prevent one additional heart attack compared with placebo, according to this dataset.

4.3.2 Odds Ratio

To compute the odds, we first determine the probability of having a heart attack in each group. In the Aspirin group, the probability of having a heart attack is 0.0232, while in the Placebo group the probability is 0.0557. The odds are then calculated by dividing the probability of the event occurring by the probability of the event not occurring. For example, in the Aspirin group, the odds of having a heart attack are

$$\text{Odds}_{\text{asp}} = \frac{0.0232}{1 - 0.0232} = 0.0238,$$

and in the Placebo group, the odds are

$$\text{Odds}_{\text{pla}} = \frac{0.0557}{1 - 0.0557} = 0.0590.$$

The odds ratio is calculated by dividing the odds in the Placebo group by the odds in the Aspirin group.

$$\widehat{OR} = \frac{\text{Odds}_{\text{asp}}}{\text{Odds}_{\text{pla}}} \approx \frac{0.0238}{0.0590} \approx 0.4031.$$

The results indicate that the odds of having a heart attack are approximately 40.31% lower in the Aspirin group compared to the Placebo group. This suggests that regular aspirin intake may have a protective effect against heart attacks.

Table 6: Odds Ratio Calculation

Group	**Odds**
Aspirin	0.0238
Placebo	0.0590

4.3.3 Confidence Interval

Using standard log-scale methods for 2 × 2 tables, a 95% confidence interval for the risk ratio is obtained. For this ”Aspirin” dataset the approximate interval is

$$RR_{95\%\,\text{CI}} \approx [0.36, 0.49].$$

This interval is clearly below 1, which supports the conclusion that aspirin truly lowers the risk of heart attack in the underlying population, not just in this particular sample. The values in the interval suggest that the true risk in the aspirin group is somewhere between 36% and 49% of the risk in the placebo group which is a substantial reduction.

4.4 Stratified analysis by smoking status

We know Smoking is a major cardiovascular risk factor (Centers for Disease Control and Prevention 2024). So we have explored whether aspirin works similarly well for smokers and non-smokers, or whether the effect is different in size.

4.4.1 Non-Smokers

For the 15,992 non-smokers, the contingency table is shown in Table 7

Table 7: Heart attack by group among non-smokers

Group	No heart attack	Heart attack	Total
Aspirin	8 361	157	8 518
Placebo	7 134	340	7 474

The risks are $\hat{R}_{\text{asp, non}} = \frac{157}{8518} \approx 0.0184$ (1.84%) and $\hat{R}_{\text{pla, non}} = \frac{340}{7474} \approx 0.0455$ (4.55%).

The resulting risk ratio is $\widehat{RR}_{\text{non-smokers}} = \frac{0.0184}{0.0455} \approx 0.404$.

A 95% confidence interval is approximately

$$RR_{\text{non-smokers}, 95\%\,\text{CI}} \approx [0.34, 0.49].$$

The risk difference in non-smokers is about -0.027, corresponding to an NNT of roughly 37. That is, among non-smokers, about 37 men would need to take aspirin to prevent one additional heart attack.(Porta 2014)

From the Figure 2,Both the Risk Ratio and Odds Ratio are located to the left of 1, and their confidence intervals do not cross the line of no effect. This provides strong visual evidence that aspirin has a protective effect against heart attacks in non-smokers.

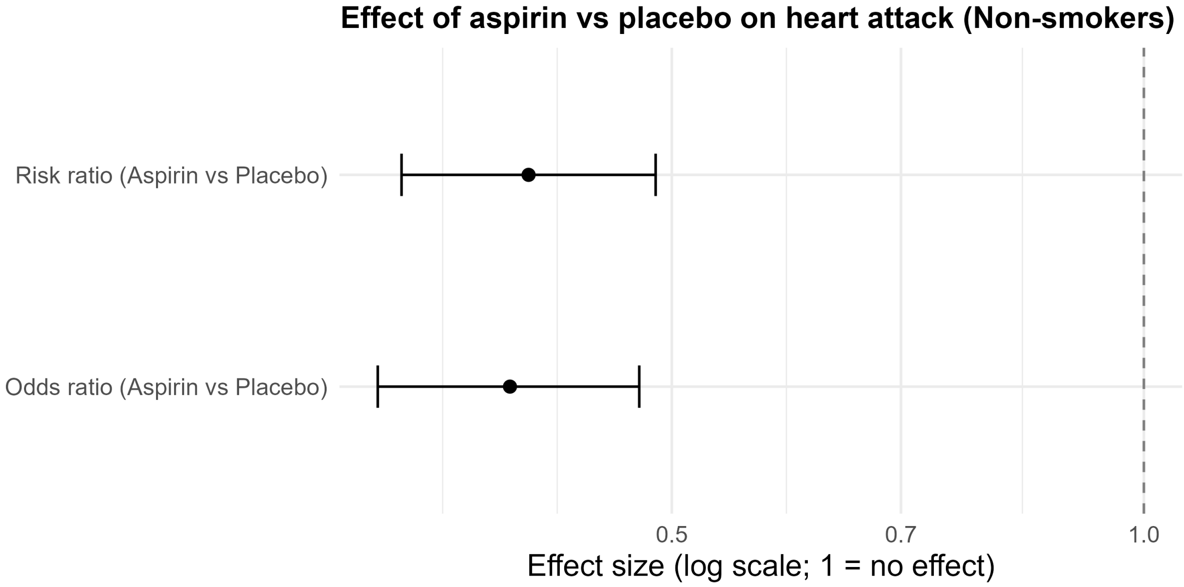

Figure 2: Forest plot of Risk Ratio and Odds Ratio for non-smokers.

4.4.2 Smokers

For the 4,029 smokers, the contingency table is shown in Table 8.

Table 8: Heart attack by group among smokers

Group	No heart attack	Heart attack	Total
Aspirin	1 394	75	1 469
Placebo	2 341	219	2 560

The risk of heart attack was higher in both groups compared to non-smokers: 8.55% in the placebo group and 5.11% in the aspirin group. Aspirin still showed a protective effect:

Risk Ratio (RR): 0.597 (95% CI: 0.463, 0.770), Odds Ratio (OR): 0.576 (95% CI: 0.437, 0.751), Number Needed to Treat (NNT): 29 (95% CI: 20, 53).

The stratified analysis reveals that while aspirin provides a protective effect in both groups, the effect is stronger in non-smokers ($RR = 0.41$) than in smokers ($RR = 0.60$). This suggests that smoking may attenuate, but not eliminate, the beneficial effect of aspirin in preventing heart attacks.

4.5 Hypothesis tests for association

Smokers: We perform the hypothesis tests separately for smokers to confirm the findings.

- **Chi-Square (χ^2) Test:** Statistic = 16.4154, p-value = 1e-04.
- **Fisher's Exact Test:** p-value < 0.0001 .

With p-values well below the standard $\alpha = 0.05$, we reject the null hypothesis H_0 and conclude that there is a statistically significant difference in heart attack frequency between the Aspirin and Placebo groups for smokers.

Non-smokers: We also look at the hypothesis tests for non-smokers.

- **Chi-Square (χ^2) Test:** Statistic = 96.802, p-value < 0.0001 (reported as 0).
- **Fisher's Exact Test:** p-value < 0.0001.

With an extremely small p-value ($p < 0.0001$), we reject the null hypothesis (Wasserstein and Lazar 2016) H_0 and since all p-values are far below our chosen significance level $\alpha = 0.05$, we reject the null hypothesis H_0 in both subgroups and conclude that there is a statistically significant difference in heart attack frequency between the Aspirin and Placebo groups for both smokers and non-smokers.

5 Summary

This project set out to investigate the effectiveness of regular aspirin intake in preventing heart attacks, utilizing data from a large, double-blind, randomized controlled trial. Our analysis supports the statement that: aspirin significantly reduces the risk of heart attacks.

The initial overall analysis revealed a substantial protective effect. The risk of a heart attack was 5.57% in the placebo group but only 2.32% in the aspirin group. This translated to a risk ratio of 0.42, indicating that individuals taking aspirin had a 58% lower risk compared to those on the placebo. The extremely small p-value ($p < 0.001$) from the Chi-Square test allowed us to confidently reject the null hypothesis and conclude that this association was not due to chance.

To gain a more nuanced understanding, we conducted a stratified analysis based on smoking status. This revealed that the beneficial effect of aspirin is present in both smokers and non-smokers, but its strength varies:

- Among **non-smokers**, the protective effect was very strong, with a risk ratio of 0.41. This means non-smokers taking aspirin reduced their relative risk by about 59%.
- Among **smokers**, the effect, while still significant and clinically important, was somewhat attenuated, with a risk ratio of 0.60, corresponding to a 40% reduction in relative risk.

In both subgroups, the confidence intervals for the risk and odds ratios did not cross 1, and hypothesis tests confirmed statistical significance.

Our findings confirm that aspirin significantly reduces heart attack risk; however, this conclusion is subject to limitations. Potential confounding factors not addressed in our analysis necessitate a cautious interpretation. We recommend further research to validate these results over the long term and to thoroughly investigate adverse effects and drug interactions. Future work should also focus on understanding how aspirin's efficacy and safety vary across different populations and dosages, which is crucial for developing targeted clinical recommendations.

References

Agresti, Alan (2009). *An Introduction to Categorical Data Analysis.* 2nd ed. Hoboken, NJ: John Wiley & Sons.

Altman, Douglas G. (1990). *Practical Statistics for Medical Research.* London: Chapman & Hall / CRC.

Bayesball (2023). *A Course in Exploratory Data Analysis.* `https://bayesball.github.io/EDA`. Retrieved from `https://bayesball.github.io/EDA`.

Bürkner, Paul, Daniel Habermann, and Lars Kühmichel (1993). *Aspirin and Heart Attacks: Teaching Dataset for Introductory Case Studies.* Unpublished dataset used in the course "Introductory Case Studies" at TU Dortmund University. Randomized study on the influence of regular aspirin intake on heart attack risk.

Cancer Association of South Africa (2021). *Fact Sheet on Aspirin.* `https://cansa.org.za/`. States that aspirin is one of the most widely used medications in the world, with an estimated 40,000 tons consumed each year.

Centers for Disease Control and Prevention (2024). *Health Effects of Cigarettes: Cardiovascular Disease.* `https://www.cdc.gov/tobacco/about/cigarettes-and-cardiovascular-disease.html`. States that smoking is a major cause of cardiovascular disease and is responsible for about one in every four deaths from CVD.

Fisher, Ronald A. (1922). "On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P". In: *Journal of the Royal Statistical Society* 85.1, pp. 87–94.

Lehmann, Erich L. and Joseph P. Romano (2005). *Testing Statistical Hypotheses.* 3rd ed. New York: Springer.

Moore, David S., William I. Notz, and Michael A. Fligner (2009). *Introduction to the Practice of Statistics.* 6th ed. New York: W. H. Freeman.

Pearson, Karl (1900). "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50.302, pp. 157–175.

Porta, Miquel (2014). *A Dictionary of Epidemiology.* 6th ed. Oxford: Oxford University Press.

R Core Team (2024). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Wasserstein, Ronald L. and Nicole A. Lazar (2016). "The ASA's Statement on p-Values: Context, Process, and Purpose". In: *The American Statistician* 70.2, pp. 129–133.

A Appendix

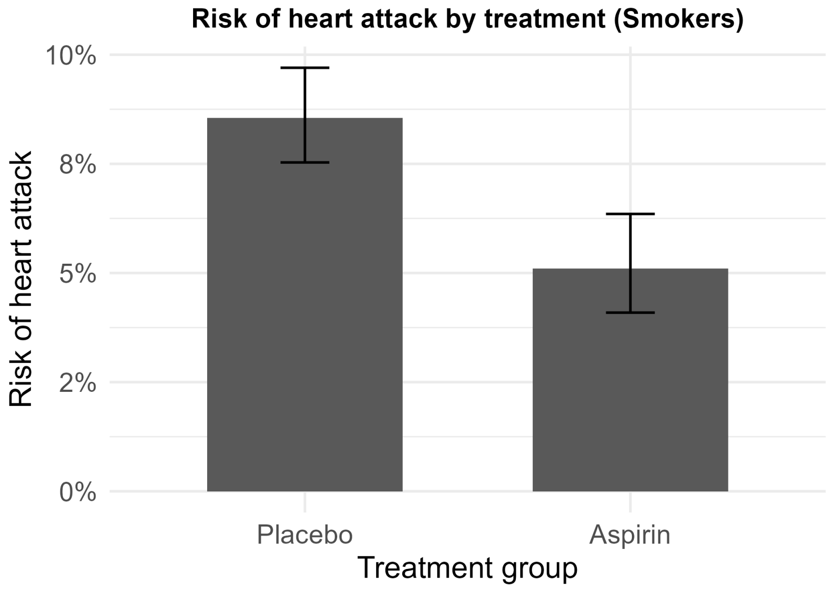

Figure 3: Heart attack risk by treatment group for smokers

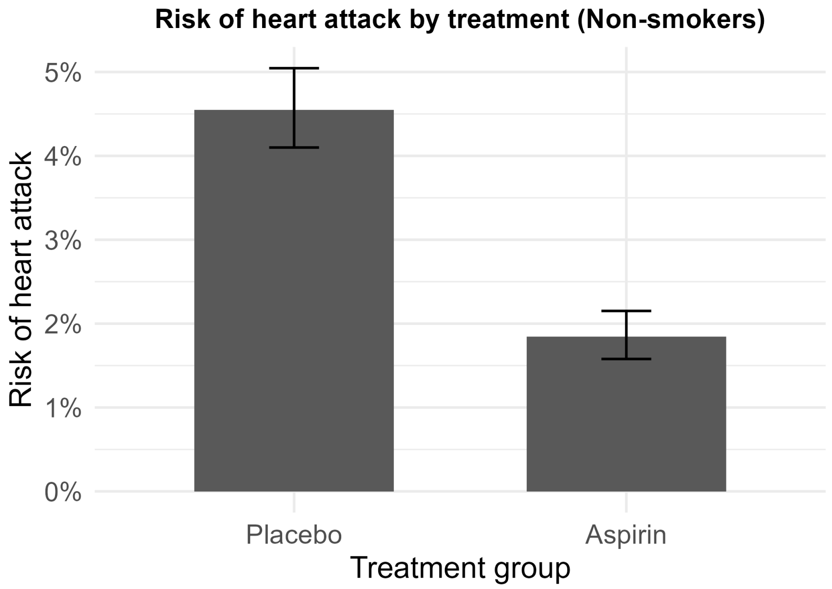

Figure 4: Heart attack risk by treatment group for non-smokers