Advanced Workbook

DLMAIIAC01 – Inference and Causality

Maximilian Stahl

Matriculation Number: 102305465

June 2nd 2024

Course of Study: Applied Artificial Intelligence M.Sc.

Tutor: Dr. Bertram Taetz

INTERNATIONAL
UNIVERSITY OF
APPLIED SCIENCES

Table of Contents

List of Figures

List of Tables

List of Equations

List of Abbreviations

Abbreviation	Definition
ADF	Augmented Dickey-Fuller
DAG	Directed Acyclic Graph
LLM	Large Language Model
MAR	Missing at Random
MCAR	Missing Completely at Random
MICE	Multiple Imputation by Chained Equations
MNAR	Missing not at Random
PSM	Propensity Score Matching
VAR	Vector Auto Regression

Preface

This advanced workbook, for the DLMAIIAC01 – Inference and Causality module, includes six tasks that explore essential concepts to the field. To support these tasks, a comprehensive code repository is available. The repository contains all necessary code and examples to ensure reproducibility.

Access the code repository here (https://github.com/MaxiStahl1992/advanced-workbook-code).

Task 1 – Bayesian Statistics: Comparison of Conjugate and Jeffreys Prior Distributions

Bayesian Statistics is a powerful framework for updating beliefs with new data using Bayes' theorem, combining prior knowledge with observed data to produce a posterior distribution reflecting updated beliefs. Choosing an appropriate prior distribution is crucial for accurate posterior distribution results, especially when data is scarce. Here, conjugate and Jefferys priors are compared, examining their influence on the posterior computation and their implications on real-world examples.

Conjugate priors simplify updating beliefs by ensuring the posterior distribution remains in the same family as the prior, which is particularly useful in analytical solutions. Examples include the Beta distribution for modeling probabilities like the possibility of a user clicking a button on a website, Normal distribution with a known variance for estimating mean effects, i.e. drug efficiency, and the Gamma distribution for modeling rates such as hospital admissions per day. This approach reduces the computational complexity and is particularly advantageous for real-time applications or large datasets. However, relying heavily on computational convenience without capturing the data complexities sufficiently, or inaccurate prior information can compromise the accuracy (Gutierrez-Pena & Muliere, 2004, p. 243).

Jefferys prior, derived from the Fisher information, is non-informative and designed to influence the posterior distribution minimally. They are useful in objective settings without clear prior knowledge, such as regulatory or legal contexts. By introducing minimal bias, Jeffreys prior helps obtain an objective posterior distribution, but may not utilize available strong prior knowledge effectively (Gutierrez-Pena & Muliere, 2004, pp. 235-244).

A real-world example could be the development of a diabetes prediction tool. To achieve the highest possible accuracy in predictions the posterior distributions from Beta conjugate priors and Jeffreys priors are compared, using datasets of 10 and 500 people. Figure 1 shows the resulting posterior distributions calculated with different assumptions for the prior.

Using an informative prior like conjugate necessitates domain knowledge that could be obtained through expert consultation. In this fictitious example, the expert analysis could lead to a prior probability of 50% ($\alpha=2$, $\beta=2$), or possibly to an 80% probability ($\alpha=5$, $\beta=1$). The resulting posterior distributions in Figure 1 display that with more data points available the posterior distributions always converge towards the true probability of 15%. With fewer available data points the calculated posterior hinges heavier on the selected prior parameters for the conjugate prior. Jeffreys prior is closer to the correct probability even with only a few available data points.

Another way to use prior information in the real world would be in a drug efficiency study, examining the effect of a new drug on reducing blood pressure, where the readings follow a normal distribution with a known variance. Figure 2 shows the resulting posterior distributions using several conjugate

priors and the Jeffreys prior. Three versions of the conjugate prior are used in the posterior calculation, each with different initial assumptions on the mean and precision. When the data is scarce, wrong initial assumptions for the conjugate prior have a larger effect on the convergence of the posterior distribution on the true mean, amplified through a high precision value. However, good initial assumptions lead to a quick convergence of the posterior distribution towards the true mean. Jeffreys prior grasps the data structures well even in small datasets, and quickly converges toward the true mean.

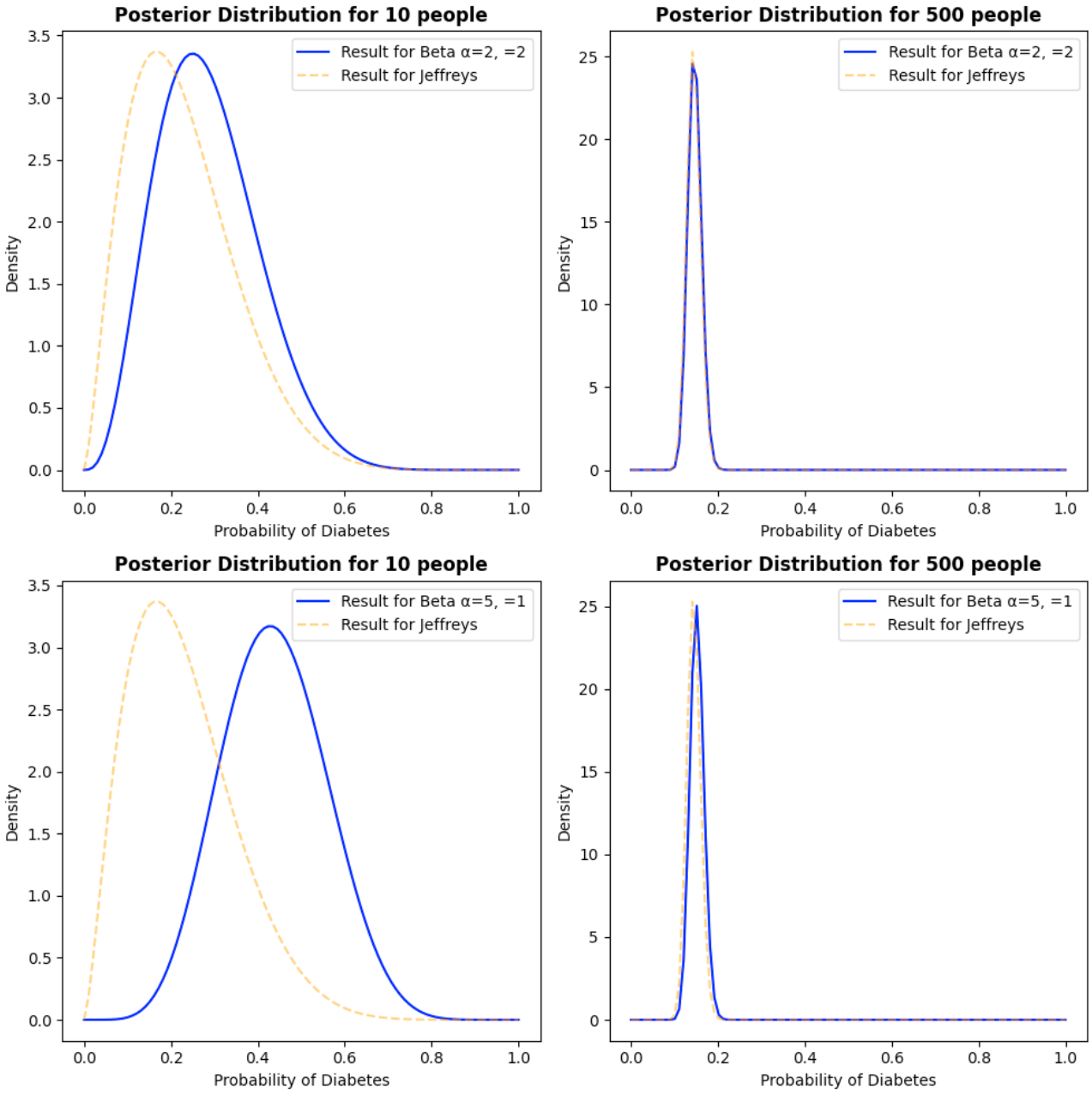

Figure 1: Comparison of Beta Conjugate Priors and Jeffreys Prior in the Diabetes Prediction example (Own Figure) (Task 1)

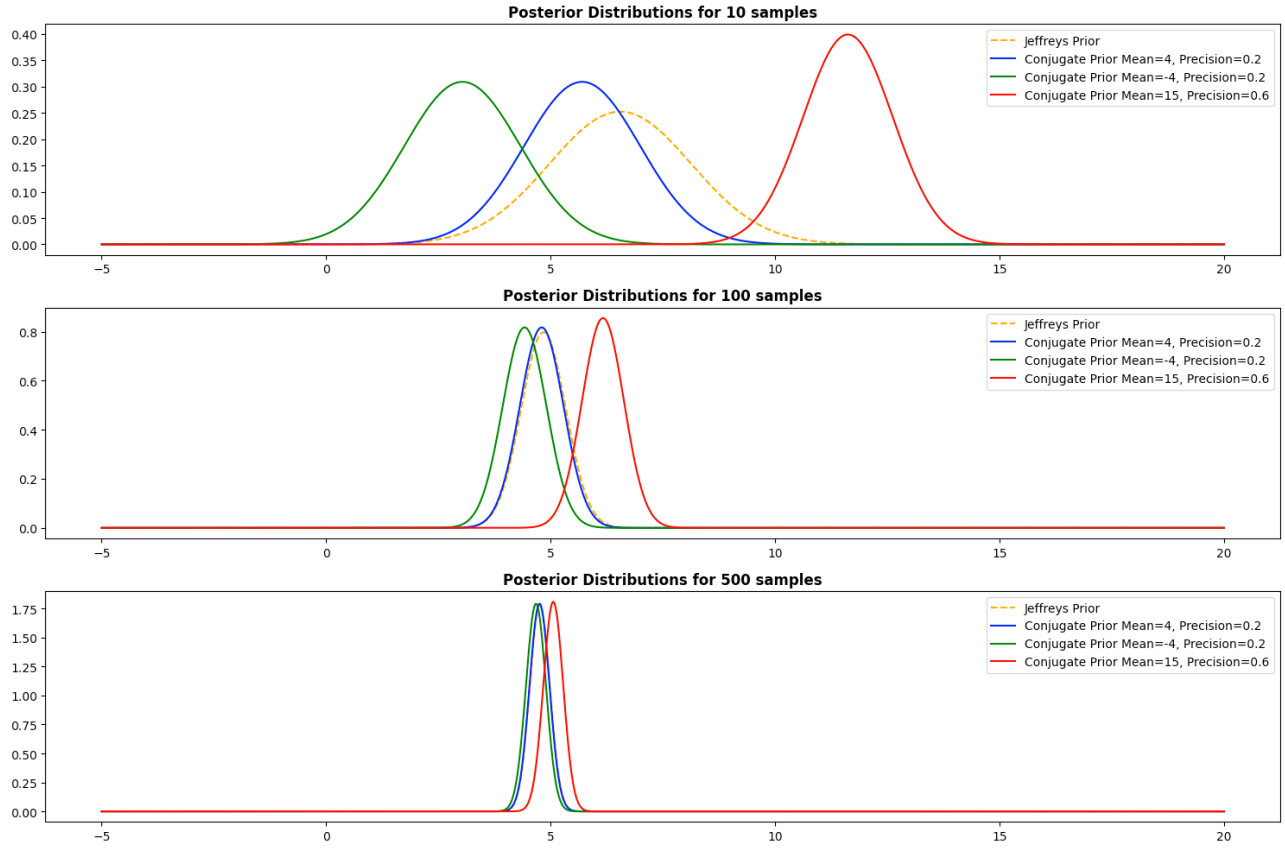

Figure 2: Comparison of Normal Conjugate Priors and Jeffreys Prior in the Drug Effect example (Own Figure) (Task 1)

Choosing the appropriate prior is essential in Bayesian statistics (Tokdar, 2013, pp. 9-10). Jeffreys prior is beneficial when prior knowledge is uncertain, providing more accurate posterior distributions. Conjugate priors are advantageous when the data is limited and prior information is correct, as shown in the drug efficiency example. Poor prior assumptions in medical decision-making could lead to incorrect diagnosis or treatment plans, in public health inappropriate priors could lead to over- or underestimating a threat. A poor prior selection in clinical trials could reject effective drugs or approve ineffective ones.

Task 2 – Granger Causality on a Stock Market Example

Granger causality tests whether past values of one time series can predict the future values of another time series.

For two different time series to be tested for Granger causality, the data must be stationary, and the dataset should be sufficiently large to capture the relationship dynamics. The test involves constructing econometric models, such as Vector Auto Regression (VAR), and performing statistical tests like the F-test to determine if one series provides predictive power over the other (Shojaie & Fox, 2022, pp. 1-4).

The example case examines the daily closing prices of Germany's DAX and MDAX stock indices over three timeframes: 5 years, 1 year, and 3 months. The goal is to determine if movements in mid-cap stocks (MDAX) can indicate larger market trends (DAX), or the other way around.

First, the stationarity is confirmed using the Augmented Dickey-Fuller (ADF) test (Guo, 2023, p. 101). Next, it is tested for Granger causality by constructing VAR models and performing F-tests to identify significant lags (Shojaie & Fox, 2022, pp. 5-6). Both directions are tested: MDAX to DAX and DAX to MDAX.

Key parameters for interpretation include:

- **Lag:** The difference in days between the series.
- **F-Test:** Determines if lagged values add statistically significant value to the forecasting model.
- **P-value:** Measures the probability that the null hypothesis (e.g. DAX does not cause MDAX) is true. Fisher (1928, p. 45), states that significant values are below 0.05.

Over 5 years, no Granger causality is detected in either direction. However, Tables 1 and 2 show significant lags in the 1-year and 3-month periods, indicating that larger market trends (DAX) influence the mid-cap market (MDAX).

Table 1: Significant Granger Causality Results for DAX causing MDAX – 1 year (Own Table) (Task 2)

Lag	F-Test	p-value
10	1.889421	0.047768

Table 2: Significant Granger Causality Results for DAX causing MDAX – 3 months (Own Table) (Task 2)

Lag	F-Test	p-value
5	3.249241	0.013925
6	2.538469	0.034996
10	2.208703	0.046941

Discovering Granger causality in financial indices can impact market perceptions. Indices predicting others might be seen as leading indicators. Traders could time trades based on predictions of market movements. For example, using the 1-year data, traders might look at the DAX to predict MDAX behavior 10 days later and adjust their strategies. Risk managers could adjust portfolios, and economists could integrate these relationships into broader models to forecast economic trends.

Larger datasets tend to provide more stable predictive power. Smaller datasets showed statistical Granger causality between DAX and MDAX, while the larger dataset did not. This could be due to short-term anomalies captured by smaller datasets, leading to spurious correlations. Figure 3 compares the DAX and MDAX indices over the three timeframes with no lag of either in the left panels, MDAX lagging DAX by 10 days in the middle panels, and DAX lagging MDAX by 10 days in the right panels. This visual doesn't show a clear clue that either index influences the other, as they all seem to correlate the most in the not-lagged visualization for all three timeframes. This can be an additional indicator that the significant correlations observed in Tables 1 and 2 are due to short-term anomalies, that could be gone when new data is added. Introducing new data can make predictions stronger in smaller datasets, making regular model updates necessary to maintain the model's accuracy.

Based on the data it cannot be concluded that mid-cap market trends indicate larger market trends. Larger market trends, on the other hand, do influence mid-cap market trends in smaller timeframes based on the Granger causality tests performed on these datasets. However, Granger causality cannot imply true causal inference, as it does not account for underlying mechanisms. Therefore, Granger Causality should always be used with caution.

Finding a true Granger Causal relationship can offer value to Traders, Risk Managers, and Economists in making reliable predictions about future market trends and adjusting their strategies accordingly. However, reliable models need enough data to overcome spurious correlations.

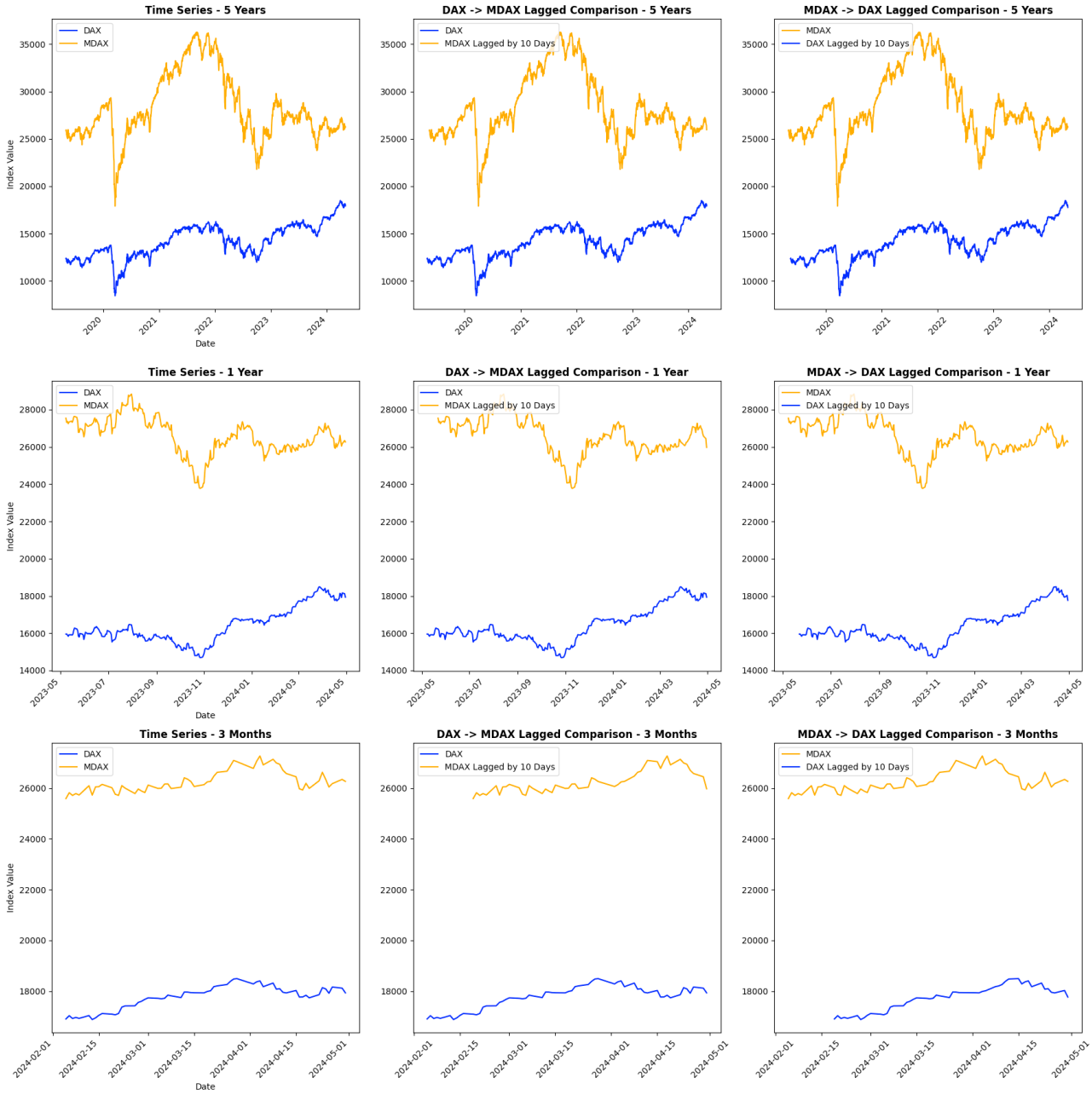

Figure 3: Dax and MDAX Indices in Different Timeframes – 5 years, 1 year, 3 months (Own Figure) (Task 2)

Task 3 – Confounders in the Context of Artificial Intelligence

Confounders are variables that influence both the dependent and independent variables, potentially biasing the estimation of the effect of the independent variable. In machine learning, unidentified confounders can lead to models that generalize poorly on new data. For example, image recognition algorithms might learn to detect cows based on typical landscapes, when presented with cows in unusual contexts, like a beach, the algorithm fails to identify them (Beery et al., 2018, pp. 1-2).

(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

Figure 4 Unidentified Confounders in Image Recognition (Beery et al. 2018, p. 2) (Task 3)

Another example is the spurious correlation between *shark attacks* and *ice cream sales*, driven by the confounder *temperature*. Training a model on *ice cream sales* to predict *shark attacks* might perform well on test data but will almost certainly fail with new data due to the unaccounted confounder. The true relationships between the variables are shown in the Directed Acyclic Graph (DAG) in Figure 5.

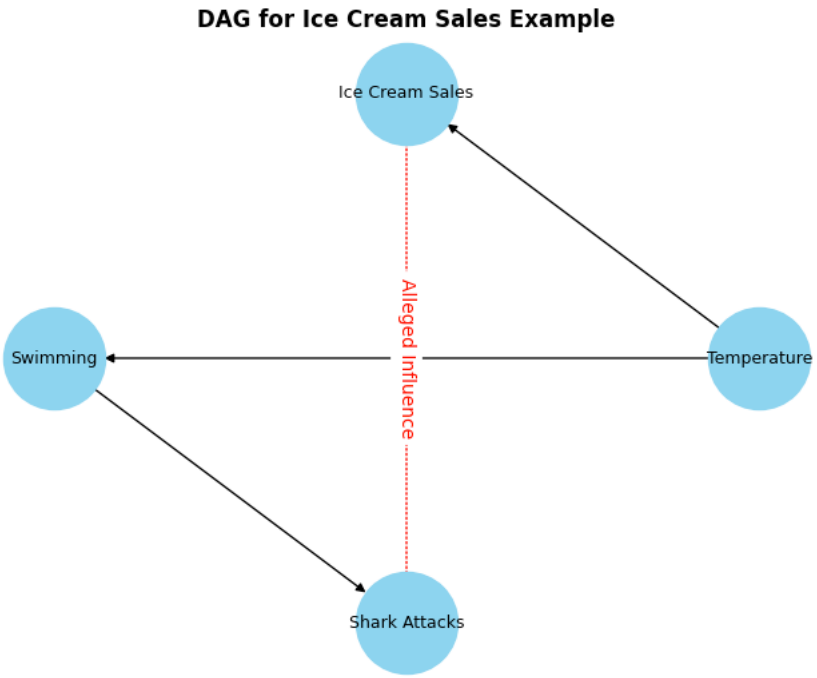

Figure 5: DAG for Ice Cream Sales and Shark Attacks (Own Figure) (Task 3)

There are various options to adjust for confounders including stratification, matching, multiple regression, and more sophisticated techniques like propensity score matching (PSM). PSM pairs units with similar propensity scores, balancing the distribution of confounders across treatment groups. However, it can be challenging in higher dimensions and may leave residual confounding if the propensity score is wrongly specified, or the dataset gets too small (Arbogast & VanderWeele, 2013, p. 139).

A powerful tool for addressing confounders can be a causal inference framework like DoWhy. DoWhy explicitly models and adjusts for confounders, providing a clear understanding of the causal relationships between variables. However, this approach can be computationally expensive and requires detailed and accurate data on potential confounders (Blöbaum et al., 2022, pp. 2-4; Sharma & Kiciman, 2020, p. 4).

In this case study, a dataset is simulated to study the impact of *exercise level* and *calorie intake* on *muscle mass*, with *exercise level* as a confounder influencing both *calorie intake* and *muscle mass*. The DAG in Figure 6 shows the true connection for the variables and is the basis on which the data for this case study was created.

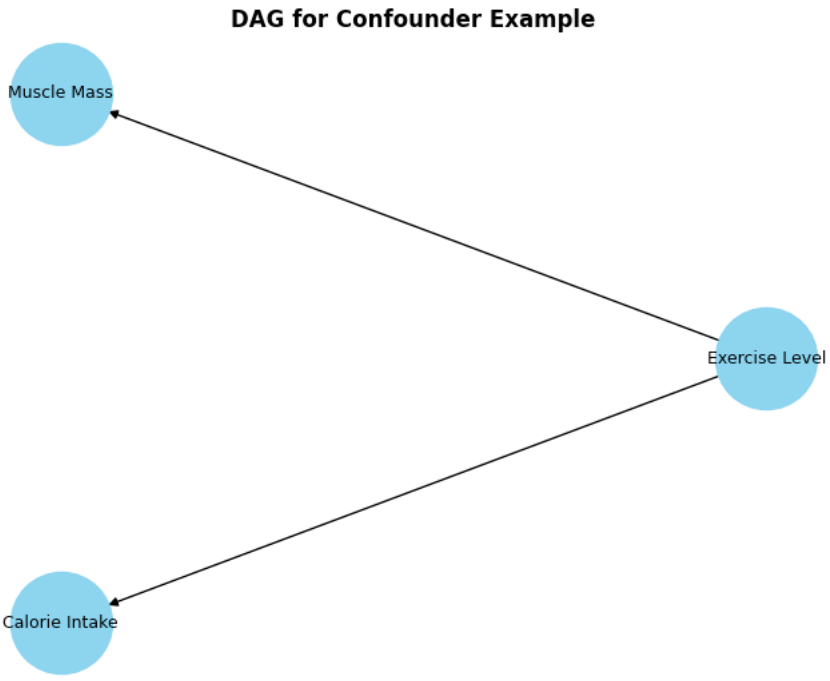

Figure 6: DAG for Calorie Intake, Exercise Level, and Muscle Mass (Own Figure) (Task 3).

Figure 7, displaying the initial variable relationships, shows a strong correlation between *exercise level* and *muscle mass*, and an imbalance between *calorie intake* and *muscle mass*. Both independent variables (*exercise level* and *calorie intake*) could predict *muscle mass* when only looking at the initial relationship.

To inspect the impact of the confounder two linear regression models are trained:

1. **Model 1** uses *calorie intake* alone to predict *muscle mass*.
2. **Model 2** includes both *calorie intake* and *exercise level*.

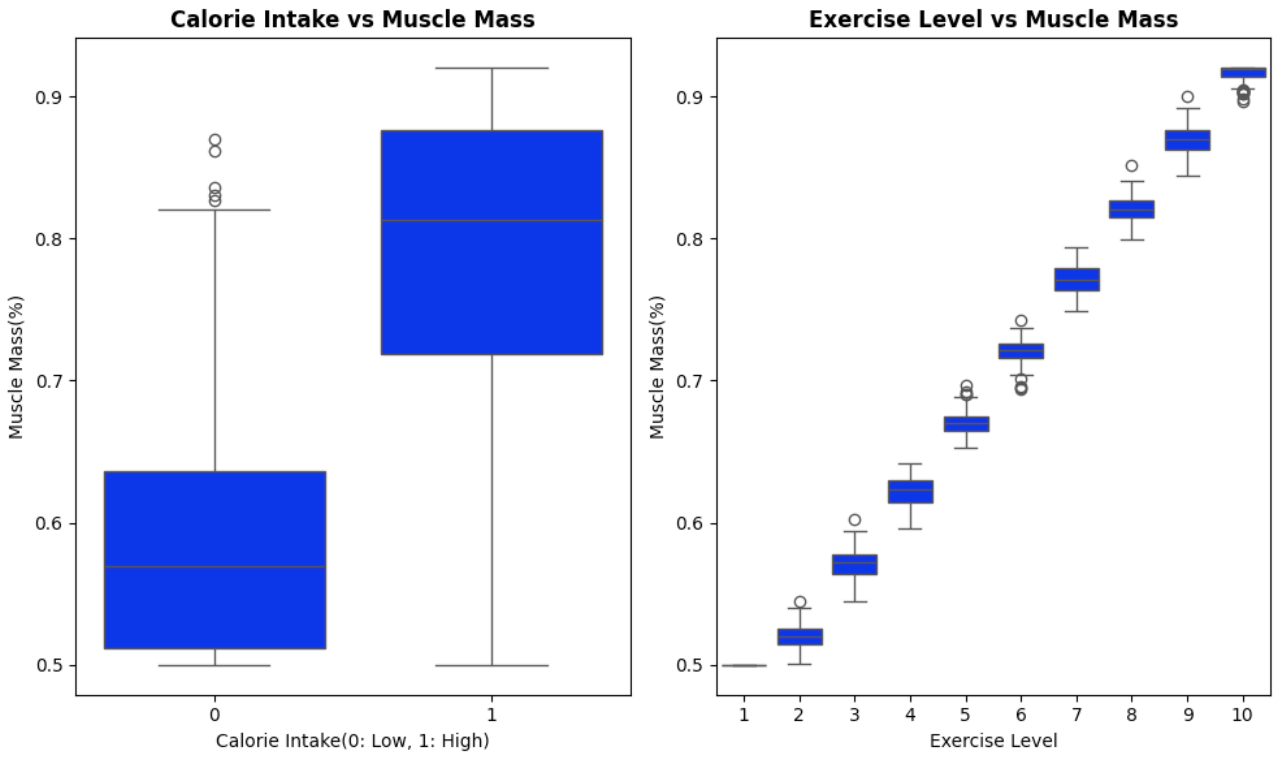

Figure 7: Relationships between Exercise Level and Muscle Mass and Calorie Intake and Muscle Mass (Own Figure) (Task 3)

Table 3 shows that including the confounder as an independent variable improves the training and the test accuracy of the model measured by the R-squared metric. Additionally, in Model 2 the coefficient for *calorie intake* was estimated to be much lower than in Model 1. This leads to the conclusion that including the confounder can be enough for a linear regression model to pick up on the underlying relationships in the data. However, according to Bellman (2010, p. IX). it becomes difficult to find hidden structures with many variables. Therefore, multiple regression is only useful for datasets with few variables.

Table 3: Linear Regression Results for Confounder Example (Own Table) (Task 3)

Model	Coefficients	Intercept	R-Squared Score Training Data	R-Squared Score Test Data
1	*Calorie Intake*: 0.2057	0.5867	0.5238	0.4991
2	*Calorie Intake*: -0.0007 *Exercise Level*: 0.0482	0.4333	0.9934	0.9930

Applying PSM to the data creates a balanced dataset by matching individuals with similar *exercise levels*. The data shows a retained pattern between *exercise level* and *muscle mass*, however, the variability and imbalance between *calorie intake* and *muscle mass* has decreased (Figure 8). This more balanced and tighter distribution of *muscle mass* between the two *calorie intake* groups suggests that PSM has created a more comparable and less biased dataset.

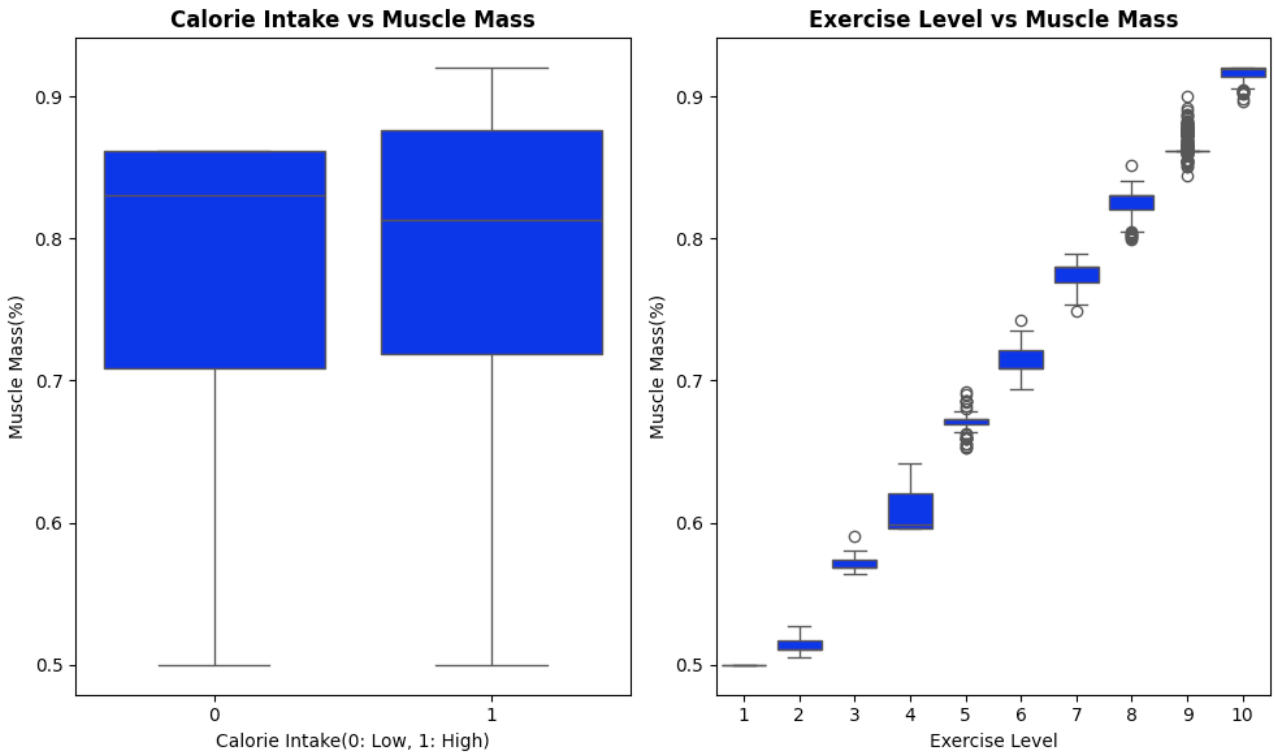

Figure 8: Relationships between Exercise Level and Muscle Mass and Calorie Intake and Muscle Mass after PSM (Own Figure) (Task 3)

The results for performing a linear regression on Model 1 and Model 2, when using the matched dataset produced by PSM are shown in Table 4. Model 1 performs even worse than it did when simply fitting a regression model to the entire dataset, for the R-squared test on the test data it has less predictive power than fitting a horizontal line. Model 2, however, produces similar results as before matching the data.

Table 4: Linear Regression Result for Confounder Example after PSM (Own Table) (Task 3)

Model	**Coefficients**	**Intercept**	**R-Squared Score Training Data**	**R-Squared Score Test Data**
Matched 1	*Calorie Intake*: 0.008	0.7772	0.0015	-0.0178
Matched 2	*Calorie Intake*: 0.0027 *Exercise Level*: 0.0498	0.4187	0.99	0.9906

Another approach to identifying a possible confounder is using DoWhy a causal inference framework described by Sharma & Kiciman (2020, pp. 4-5). In this example, the causal effect of *calorie intake* on *muscle mass* can be identified by adding *exercise level* as a common cause. This approach uses causal graphs and the backdoor criterion to adjust for confounders. The estimated causal effect size is negative (-1.925), with a high p-value (1.00), indicating that the impact of *calorie intake* on *muscle mass* is practically nonexistent.

Confounder adjustment is crucial for accurate AI model interpretation and reliability. This case study uses different approaches to identify and adjust for confounders. Using a causal inference framework like DoWhy, and utilizing methods like the backdoor criterion, provides a profound and straightforward way to identify confounders. Addressing a confounder can be as simple as including it in model creation, as shown in this example to improve the training and testing accuracy. The second example using PSM effectively identifies the confounder impact, by creating a balanced and comparable dataset.

Studies like Recognition in Terra Incognita by Beery et al. (2018, pp. 1-2) show that not accounting for confounders can lead to poor model performance when applying it to new data. This can have severe results in cases like medical imaging. Missing potential serious health conditions or diagnosing non-existent diseases can lead to harmful decisions for the patient.

Task 4 – Educational Platform and the Front Door Criterion

The causal effect of an intervention through a mediator, even in the presence of an unmeasured confounder, can be estimated using the front door criterion and the front door adjustment formula.

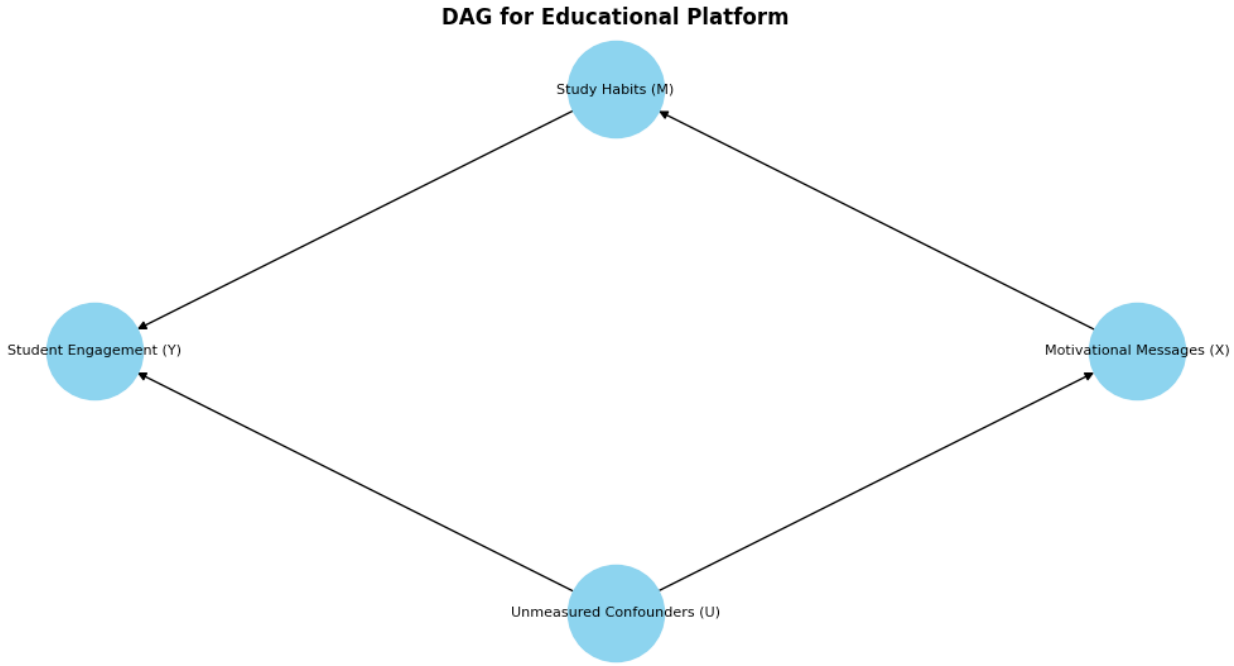

Figure 9: DAG for the Front Door Criterion Educational Platform Example (Own Figure) (Task 4)

Consider an educational platform detecting student disengagement and sending motivational messages to improve engagement. The variables involved are:

- *X (Intervention or Treatment)*: Motivational message (0 or 1).
- *M (Mediator)*: Study habits, influenced by the motivational message, measured from 0 to 1.
- *Y (Outcome)*: Student engagement after receiving the message, measured from 0 to 1.
- *U (Unmeasured Confounder)*: This could be the student's socioeconomic status or other personal circumstances.

The relationships between the variables are visualized through DAG shown in Figure 9. For the front-door criterion to hold three conditions must be met:

1. ***M (Mediator)* intercepts all direct paths from *X (intervention or treatment)* to *Y (outcome)*:**
   The primary causal pathway should be $X \rightarrow M \rightarrow Y$, with no direct edge from *X* to *Y*.
   In the example case, the only direct path from *X (motivational message)* to *Y (student engagement)* goes through *M*.
   The condition is met.
2. **No unblocked backdoor path from *X* to *M*:**
   There should not be any confounders that open a backdoor path from *X* to *M*.

The example case has no backdoor path from *X (motivational message)* to *M (study habits)*. The condition is met.

3. **X blocks all backdoor paths from *M* to *Y*:**
   X should block potential confounders affecting both *M* and *Y*.
   In the example, the unmeasured confounder *U (unmeasured confounder)* does not directly affect *M (study habits)*.
   The condition is met.

Equation 1: Front-Door Adjustment Formula (Task 4)

$$P(Y \mid do(x)) = \sum m \, P(m \mid x) \sum x' \, P(y \mid x', m) \, P(x')$$

The front-door adjustment formula can be used to estimate the causal effect of *X* on *Y*. It is typically expressed for the probability distribution $P(Y \mid do(x))$, however in this case it is sufficient to calculate the expected value which is a single summary measure of the effect. To calculate the expected value the front-door adjustment formula needs to be tweaked. In the case of continuous variables *M* and *X* the summations for $\sum m$ and $\sum x'$ would have to be Integrated instead. For this example, however, the calculation uses discrete values which leads to the following equation (Pearl, 2009, pp. 103-111):

Equation 2: Front-Door Adjustment Formula for the Expected Value (Pearl, 2009, pp. 103-111) (Task 4)

$$E[Y \mid do(X=x)] = \sum m \, P(M=m \mid X=x) \sum x' \, P(Y \mid X=x', M=m) \, P(X=x')$$

These are the steps to calculate this formula for this example (Pearl, 2009, pp. 103-111):

1. Estimate $P(M \mid X=x)$:
   $P(M=m \mid X=x) = \beta 0,M + \beta 1,M * x$
2. Estimate $P(Y \mid X0x', M=m)$:
   $P(Y=y \mid X=x', M=m) = \beta 0,Y + \beta 1,Y * x' + \beta 2,Y * m$
3. Calculate Marginal Probability:
   $P(X=x') = empirical\ distribution$
4. Compute $E[Y \mid do(X=x)]$:
   $E[Y \mid do(X=x)] = \sum m \, P(M=m \mid X=x) \sum x' \, P(Y \mid X=x', M=m) \, P(X=x')$
5. Estimate the causal effect:
   $Causal\ Effect = E[Y \mid do(X=1)] - E[Y \mid do\ (X=0)]$

Following these steps leads to an estimated total effect of *X* on *Y* of 0.045, indicating that sending a *motivational message* increases *student engagement* by 4.5%. The interpretation of this effect depends on the intervention's goal. For instance, if a 2% drop in *student engagement* triggers the message to be sent, a 5% increase in *student engagement* would be successful. However, the measure might not justify the means if *motivational messages* are sent after a 20% drop in *student engagement*.

The front-door criterion provides a robust method for estimating causal effects when unobserved confounders are present, making it invaluable in contexts where other methods fall short. In the example of the educational platform, applying the front-door criterion clarified the impact of *motivational messages* on *student engagement*, highlighting the practical significance of considering indirect effects through mediators. This approach improves model accuracy and ensures interventions are based on sound causal relationships, ultimately leading to better-informed decisions and strategies.

Task 5 – The Impact of Missing Data on the Example of Weight Loss and Diet

Analyzing the relationship between diet and weight loss using observational data can be challenging, particularly if data is missing. Missing values can significantly bias results and affect the validity of conclusions. Addressing missing data is crucial for making accurate causal inferences (Pedersen et al., 2017, pp. 164-165). The following discusses how missing values impact the analysis, comparing imputation techniques, and proposing a causal approach for handling missing data.

Missing data can occur in three primary forms (Pedersen et al., 2017, pp. 157-158):

- **Missing Completely at Random (MCAR):** The probability of data being missing is independent of both observed and unobserved variables (e.g. survey responses lost due to a technical error)
- **Missing at Random (MAR):** The probability of missing data depends on the observed data (e.g., the likelihood of reporting *diet adherence* depends on *weight*)
- **Missing not at random (MNAR):** The probability of missing data depends on the unobserved data (e.g. participants with lower *diet adherence* are less likely to report their *diet adherence*)

Missing data can lead to bias. For example, if low *diet adherence* is systematically underreported, an analysis without proper imputation will likely overestimate or underestimate diet effectiveness. Additionally, missing data reduces the study's power to detect true effects (Pedersen et al., 2017, pp. 157-158).

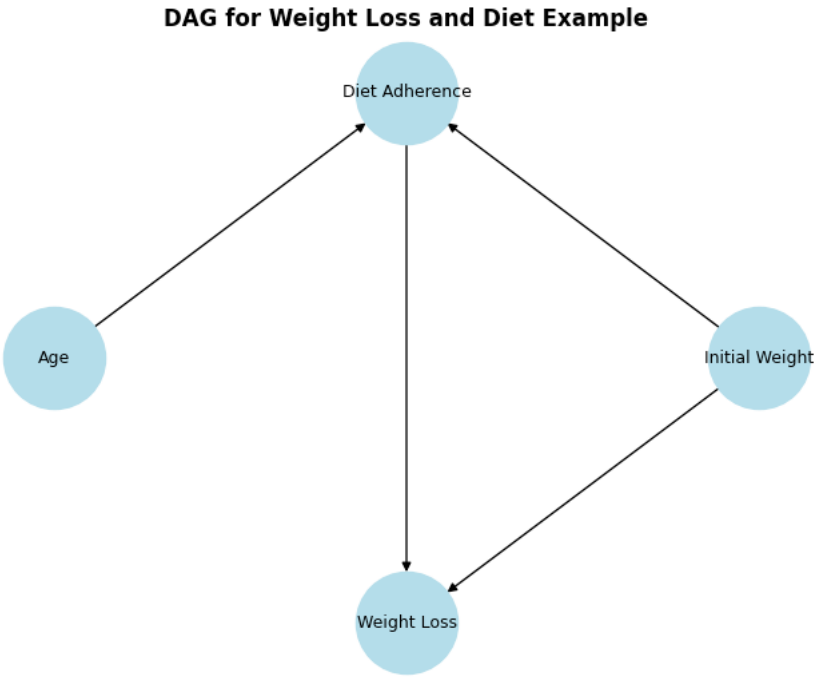

Figure 10: DAG for Weight Loss and Diet Example (Own Figure) (Task 5)

Traditional imputation techniques often assume MCAR or at best MAR cases, using techniques like mean substitution. In contrast, causal imputation methods incorporate additional structural

information from DAGs, modeling dependencies among variables more accurately and maintaining the integrity of the dataset (Pedersen et al., 2017, pp. 159-161).

A causal imputation approach can be used based on regression models informed by the causal relationships identified in the DAG to handle the missing data effectively. This approach leverages the known causal structure to predict values, ensuring that the imputation process respects and uses the true correlations within the variables (Ding & Li, 2018, pp. 218-221).

For this example, four datasets are simulated by the variable relationships represented in the DAG in Figure 10. The first dataset is complete and serves as a baseline for the others, in the second dataset random instances of *diet adherence* were removed (MCAR). In dataset three *diet adherence* instances were removed with a higher likelihood if the initial weight was above 80kg (MAR). In the fourth dataset instances of *diet adherence* were more likely to be removed if *diet adherence* itself was below 30% (MNAR).

The left panel of Figure 10 shows the distribution of *diet adherence* for each dataset. All three datasets with missing data differ from the original distribution. The resulting coefficients calculated by a regression model for *diet adherence* differ from the original baseline calculation. Depending on the case the calculated coefficient is weighted higher (MAR) or lower (MCAR and MNAR) as for the original dataset.

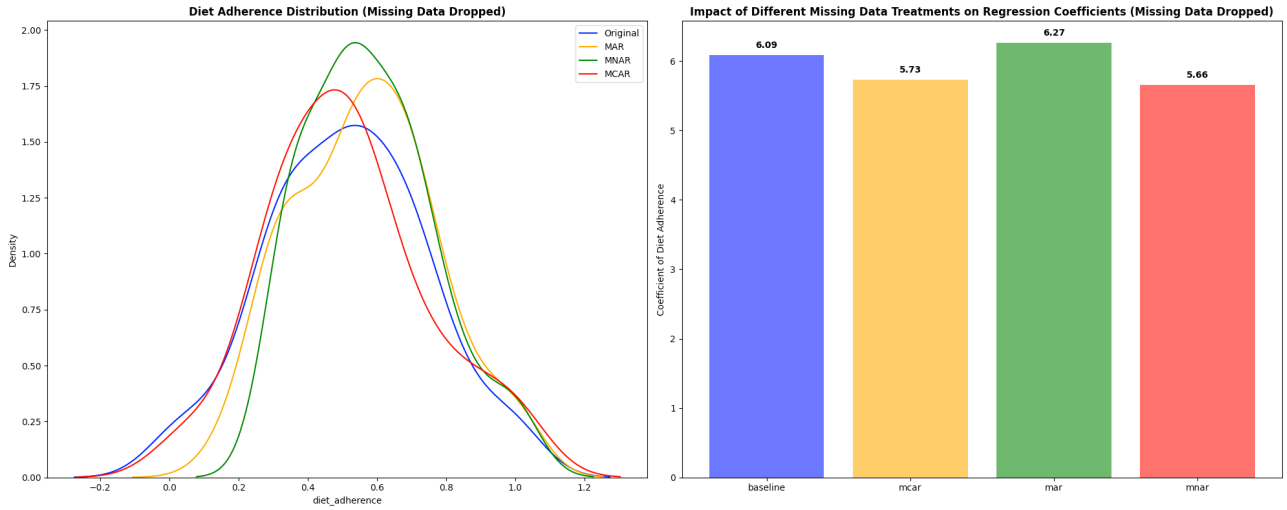

Figure 11: Impact of Missing Data on Data Distribution and Regression Results (Own Figure) (Task 5)

To impute the missing data, two approaches are compared. First, a simple imputation algorithm using the mean values, and second a causal approach using a linear regression model based on causal information retrieved from the DAG.

Figure 11 shows the data distributions for the *diet adherence* column after imputation for both cases. The left panel visualizes the data distributions after using simple imputation. Noticeably, all three simple imputed dataset distributions are narrower and have a higher peak around the mean. The

most extreme deviation can be observed for the MCAR case. The causal approach, shown in the right panel, follows the original distribution almost exactly for all three cases.

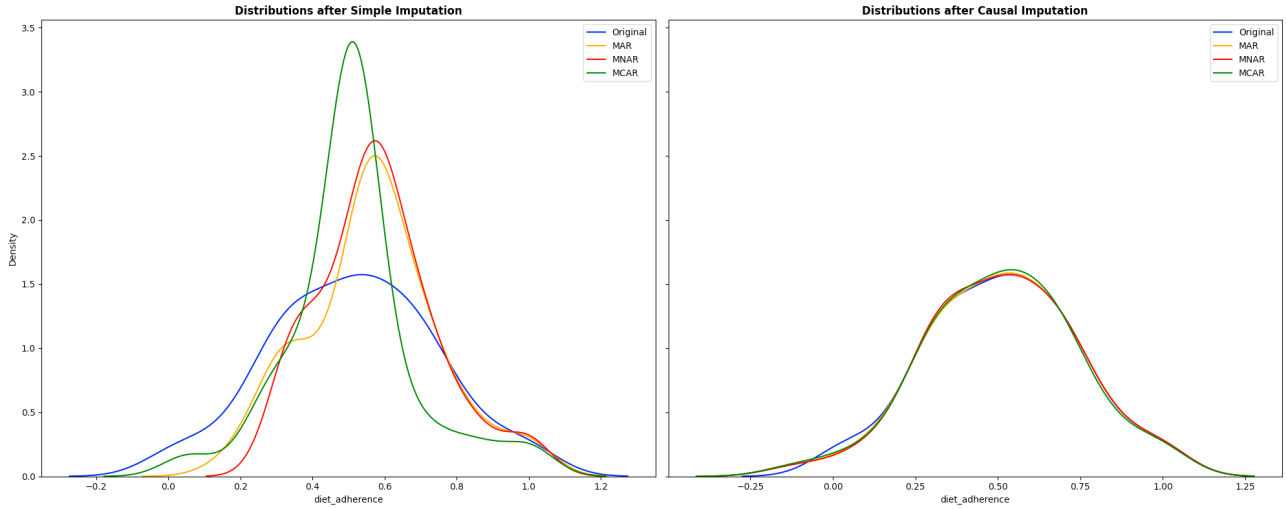

Figure 12: Data Distribution for Simple Imputation vs. Causal Imputation (Own Figure) (Task 5)

The resulting regression coefficients (Figure 12) reflect the results observed in the data distributions. Simple imputation, shown in the left panel, leads to a significant underestimation of the impact of *diet adherence* to the model in all cases. Causal imputation, shown in the right panel, on the other hand, leads to coefficient calculations that are not the same as in the baseline case but are very close to the original across all three cases.

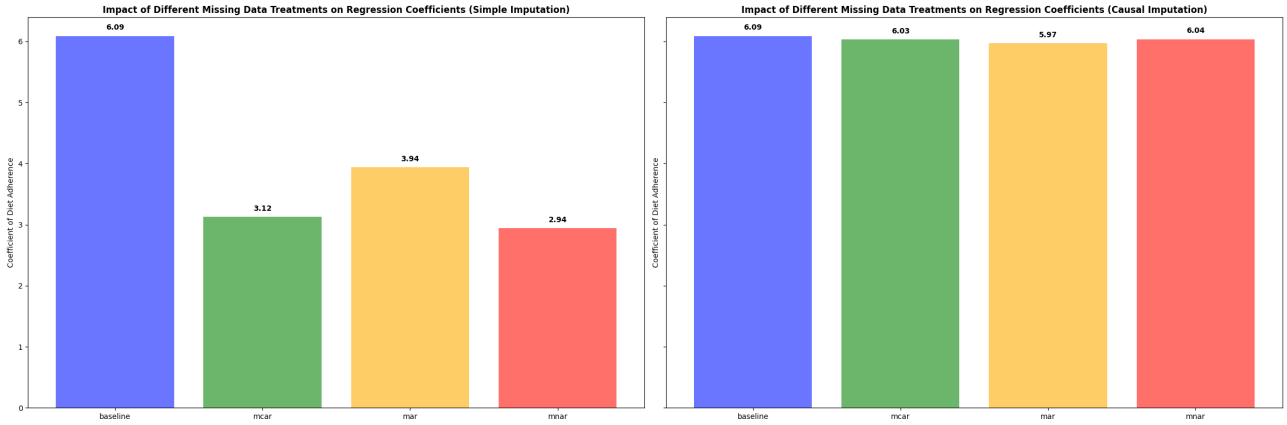

Figure 13: Calculated Regression Coefficients for Imputed Datasets (Own Figure) (Task 5)

Causal imputation is beneficial in observational studies as it helps preserve the causal structure of the data. By leveraging known causal relationships among variables, causal imputation ensures that missing values are filled in a way that respects these dependencies, leading to more accurate predictions and maintaining the data's integrity. This approach is particularly effective in MAR and MNAR cases, reducing bias that simpler approaches like mean imputation might introduce (Pedersen et al., 2017, pp. 159).

However, implementing a causal method requires a thorough understanding of causal relationships, accurate DAG construction, and substantial domain expertise (Rodrigues et al., 2022, pp. 1339).

The analysis shows that causal imputation estimates of regression coefficients are closer to the baseline models. They better replicate the data's distribution compared to simple or no imputation. Another approach that handles complex data well is the Iterative Imputer from scikit-learn (Pedregosa et al., 2011, n.p.). It is based on Multiple Imputation by Chained Equations (MICE) and, for this example, yields the same results as the causal approach.

Consistency in the results is key, as they ensure the validity and reproducibility of the findings. In the context of this example, consistent results mean that the estimated effects of *diet adherence* on *weight loss* are reliable and can be generalized to similar populations. By employing a causal imputation approach the bias can be minimized, the accuracy of models enhanced, and more robust evidence obtained for decision-making.

Recent research indicates that large language models (LLMs) like ChatGPT show promise in imputing and generating new data, achieving superior results to traditional approaches. Nazir et al. (2023, p. 1) were able to impute data, using ChatGPT with a human in the loop, and achieve enhanced accuracy. They see the possibility that ChatGPT may directly be able to generate numerical biological data in the future with high accuracy. Chen et. al (2023, p. 7) developed an innovative framework using graph attention networks with LLMs to impute spatiotemporal data. They see the potential for LLMs as a powerful tool in spatiotemporal data imputation.

Task 6 – Collider Bias in a Hypothetical Study on Stress, Smoking, and Heart Disease

Collider bias occurs when a variable, influenced by two other variables, is controlled, creating a spurious association between them. This bias is significant in causal inference as it can lead to incorrect conclusions about the causal relationships.

In this hypothetical scenario, three binary variables are considered: *stress (0: no stress, 1: stress), smoking (0: non-smoker, 1: smoker), heart disease (0: no heart disease, 1: heart disease)*, and a fourth not observed *outside factor*, i.e. a genetic predisposition, influencing both *smoking* and *heart disease*. The relationships between the variables are shown in Figure 14.

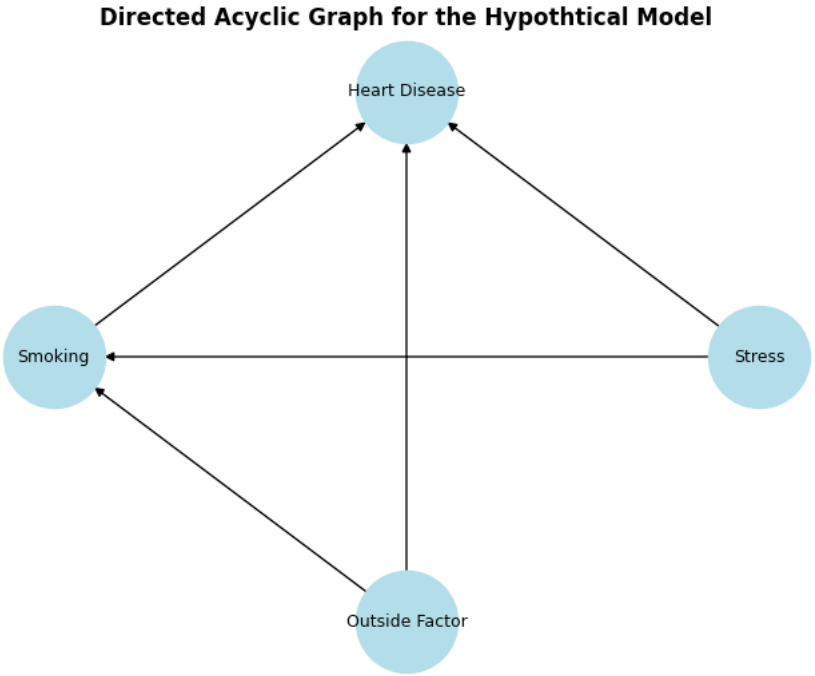

Figure 14: DAG for the Relationship of Smoking and Stress on Heart Disease in the Presence of an Unobserved Outside Factor (Own Figure) (Task 6)

Controlling for *smoking* in this example introduces bias, as it is a collider influenced by both *stress* and an *outside factor*. When conditioning on *smoking* it inadvertently creates a spurious correlation between *stress* and *heart disease* by opening a backdoor path through *smoking*. Controlling for *smoking* mixes the effects of *stress* and the *outside factor* on *heart disease* through *smoking*, creating a spurious association between *stress* and *heart disease*.

Although *stress* has a direct causal link to *heart disease*, the presence of a backdoor path still matters as it introduces additional spurious associations, which can distort the estimated effect of *stress* on *heart disease* (Rohrer, 2018, p. 31).

Figure 15 illustrates the impact of collider bias by showing the relationships between *stress* and *heart disease* in the general population and after controlling for *smoking*.

In the general population (left panel) the bar for *stressed* individuals with *heart disease* is a lot higher than for *non-stressed* individuals. Also, the difference in the gaps between people with and without *heart disease* is noticeably larger for *stressed* individuals. Only looking at *smokers* (right panel) the gaps in *heart disease* cases for both *stressed* and *non-stressed* individuals widen. However, there are relatively fewer individuals without *heart disease* compared to the general population.

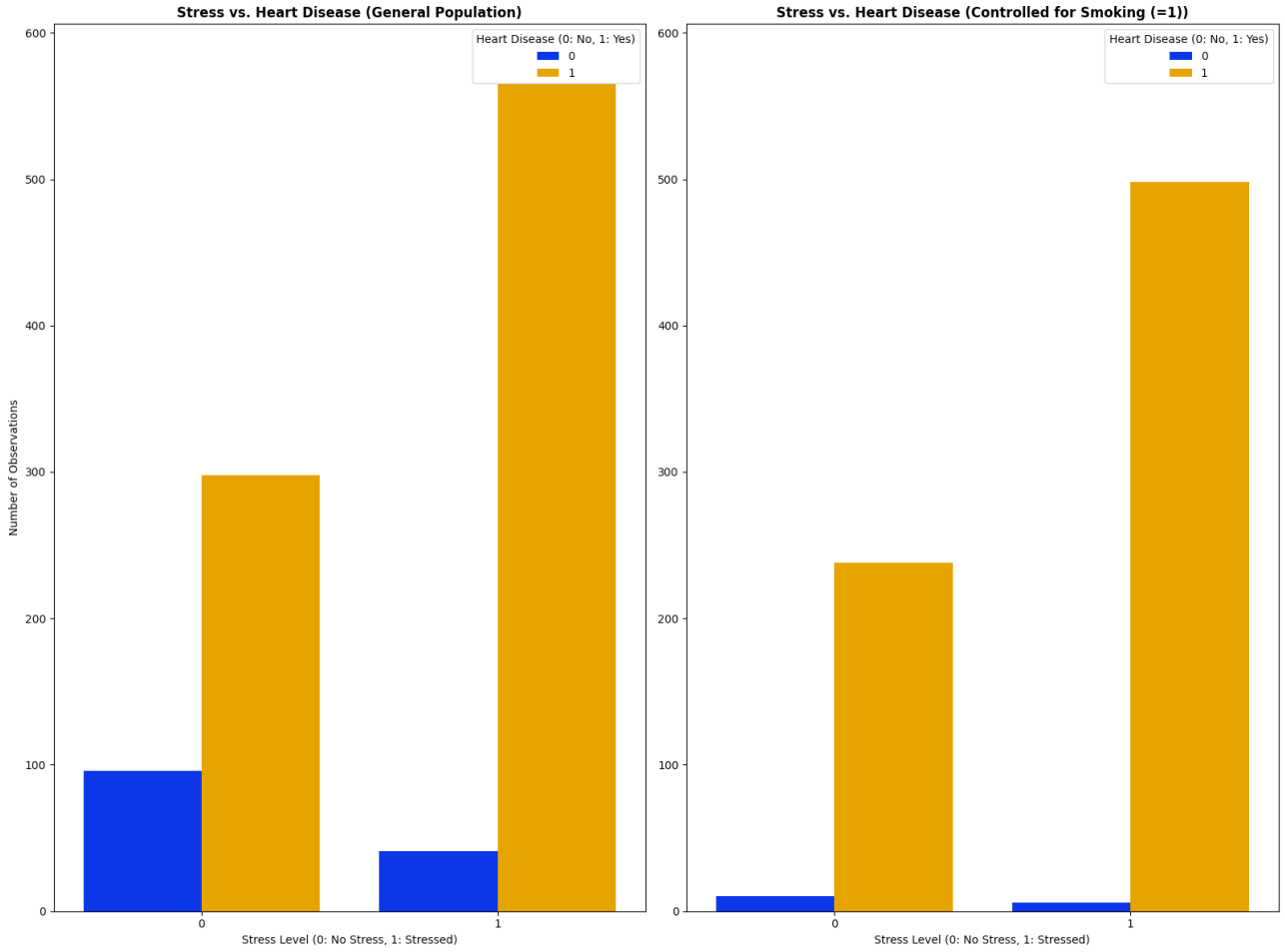

Figure 15: Distribution of Heart Disease Occurrence Based on Stress Level when Controlling and Not Controlling for Smoking (Own Figure) (Task 6)

The observed change suggests that a third factor influences the relationship between *stress* and *heart disease* when controlling for *smoking*. Including the DAG (Figure 14) in the analysis, shows that a backdoor path has been opened mixing those who *smoke* due to *stress* with those influenced by the *outside factor*, overstating the effect of *stress* on *heart disease*.

Table 5: Chi2_Contingency Results for the Relationship between Stress and Heart Disease (Own Table) (Task 6)

	General Population	**Smokers**
Chi-square	61.074	5.1534
p-value	5.49e-15	0.0232
Correlation	0.2501	0.0926

The Chi-squared test results in Table 5, demonstrate significant associations (Turney, 2022, n.p.). The general population results indicate a strong positive relationship between *stress* and *heart disease*. When controlling for *smoking* the correlation between *stress* and *heart disease* is reduced, but still statistically significant.

The significant reduction in Chi-square values when controlling for *smoking* indicates that *smoking* acts as a collider, introducing an external factor that influences *heart disease*. The *outside factor* is more prevalent in *smokers*, explaining the reduced correlation.

Identifying colliders includes examining the DAG for collider structure and performing statistical tests like the Chi-square tests and correlation analysis to observe relationship changes when controlling for potential colliders. Constructing a comprehensive DAG offers several advantages, such as clarity in the causal relationships, informed variable selection, and an enhanced understanding of the variable interaction. However, constructing a comprehensive DAG can be challenging. It can be complex, missing expertise might lead to missing variables or misinterpretation or subjectivity can introduce biases and erroneous assumptions (Rodrigues et al., 2022, p. 1345-1347).

Mitigating collider bias can be done during the study design and analysis, by using Causal Diagrams (DAGs) researchers can identify colliders and leave them uncontrolled (Lee H, 2019, n.p.). In this example not controlling for *smoking* but controlling for true confounders like the *outside factor* can help isolate the direct effect of *stress* on *heart disease*.

In conclusion, controlling for *smoking* introduces collider bias, distorting the true relationship between *stress* and *heart disease*. In this case, controlling for the collider will overestimate the effect of *stress* on *heart disease* (for *smoking* = 1). By carefully mapping out the causal pathways potential colliders can be identified and controlling for them can be avoided, thus preserving the true relationship between variables. Proper application of DAGs in causal inference is crucial in obtaining valid and reliable results, highlighting the importance of thorough planning, and understanding the study design.

Bibliography

Arbogast, P. G., & VanderWeele, T. J. (2013). Considerations for Statistical Analysis. In *Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide.* Rockville (MD): Agency for Healthcare Research and Quality (US). https://www.ncbi.nlm.nih.gov/books/NBK126192/

Beery, S., van Horn, G., & Perona, P. (2018). *Recognition in Terra Incognita* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1807.04975

Bellman, R. E. (2010). *Dynamic Programming*. Princeton University Press. https://doi.org/10.1515/9781400835386

Blöbaum, P., Götz, P., Budhathoki, K., Mastakouri, A. A., & Janzing, D. (2022). *DoWhy-GCM: An extension of DoWhy for causal inference in graphical causal models* (arXiv:2206.06821). arXiv. http://arxiv.org/abs/2206.06821

Chen, Y., Wang, X., & Xu, G. (2023). *GATGPT: A Pre-trained Large Language Model with Graph Attention Network for Spatiotemporal Imputation* (arXiv:2311.14332). arXiv. http://arxiv.org/abs/2311.14332

Ding, P., & Li, F. (2018). Causal Inference: A Missing Data Perspective. *Statistical Science*, *33*(2). https://doi.org/10.1214/18-STS645

Fisher, R. A. (1928). *Statistical methods for research workers* (2nd ed.). Oliver and Boyd.

Guo, Z. (2023). Research on the Augmented Dickey-Fuller Test for Predicting Stock Prices and Returns. *Advances in Economics, Management and Political Sciences*, *44*(1), 101–106. https://doi.org/10.54254/2754-1169/44/20232198

Gutiérrez-Peña, E., & Muliere, P. (2004). Conjugate Priors Represent Strong Pre-Experimental Assumptions. *Scandinavian Journal of Statistics*, *31*(2), 235–246. https://doi.org/10.1111/j.1467-9469.2004.02-019.x

Lee, H., Aronson, J., & Nunan, D. (2019). *Catalogue of bias collaboration*. Collider Bias. In Catalogue of Bias. https://catalogofbias.org/biases/collider-bias/

Nazir, A., Cheeema, M. N., & Wang, Z. (2023). ChatGPT-based biological and psychological data imputation. *Meta-Radiology*, *1*(3), 100034. https://doi.org/10.1016/j.metrad.2023.100034

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, *3*(none). https://doi.org/10.1214/09-SS057

Pedersen, A., Mikkelsen, E., Cronin-Fenton, D., Kristensen, N., Pham, T. M., Pedersen, L., & Petersen, I. (2017). Missing data and multiple imputation in clinical epidemiological research. *Clinical Epidemiology*, *Volume 9*, 157–166. https://doi.org/10.2147/CLEP.S129785

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Rodrigues, D., Kreif, N., Lawrence-Jones, A., Barahona, M., & Mayer, E. (2022). Reflection on modern methods: Constructing directed acyclic graphs (DAGs) with domain experts for health services research. *International Journal of Epidemiology*, *51*(4), 1339–1348. https://doi.org/10.1093/ije/dyac135

Rohrer, J. M. (2018). Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27–42. https://doi.org/10.1177/2515245917745629

Sharma, A., & Kiciman, E. (2020). *DoWhy: An End-to-End Library for Causal Inference* (arXiv:2011.04216). arXiv. http://arxiv.org/abs/2011.04216

Shojaie, A., & Fox, E. B. (2022). Granger Causality: A Review and Recent Advances. *Annual Review of Statistics and Its Application*, *9*(1), 289–319. https://doi.org/10.1146/annurev-statistics-040120-010930

Tokdar, S. (2014). *Choosing a Prior Distribution*. [Duke University]. https://www2.stat.duke.edu/~st118/sta732/Prior.pdf

Turney, S. (2022). *Chi-Square (X^2) Tests | Types, Formula & Examples*. Statistics. https://www.scribbr.com/statistics/chi-square-tests/