Causality in the Can: Diet Coke's Impact on Fatness

Yicheng Qi[1], **Ang Li**[2]

[1]Soochow University, Suzhou, China
[2]Florida State University, Tallahassee, FL, USA
ycqi@stu.suda.edu.cn, angli@cs.fsu.edu

Abstract

Artificially sweetened beverages like Diet Coke are often considered healthier alternatives, but the debate over their impact on obesity persists. Previous research has predominantly relied on observational data or randomized controlled trials (RCTs), which may not accurately capture the causal relationship between Diet Coke consumption and obesity. This study uses causal inference methods, employing data from the National Health and Nutrition Examination Survey (NHANES) to examine this relationship across diverse demographics. Instead of relying on RCT data, we constructed a causal graph and applied the back-door criterion with its adjustment formula to estimate the RCT distributions. We then calculated the counterfactual quantity, the Probability of Necessity and Sufficiency (PNS), using both NHANES data and estimated RCT data. We propose that PNS is the essential metric for assessing the impact of Diet Coke on obesity. Our results indicate that between 20% to 50% of individuals, especially those with poor dietary habits, are more likely to gain weight from Diet Coke. Conversely, in groups like young females with healthier diets, only a small proportion experience weight gain due to Diet Coke. These findings highlight the influence of individual lifestyle and potential hormonal factors on the varied effects of Diet Coke, providing a new framework for understanding its nutritional impacts on health.

Introduction

As global health consciousness rises, low-calorie artificial sweetener beverages like Diet Coke have become popular choices in the beverage market. Since its launch in 1982 (Spren 2023), Diet Coke has quickly captivated a large consumer base with its zero-sugar or low-calorie profile. Statistics (Mashed 2023) show that Diet Coke holds a significant market share in the soft drink markets of many countries, particularly among those pursuing healthier lifestyles, where it has become the preferred daily beverage. Consumers primarily choose Diet Coke for its low-calorie properties and its potential benefits in weight management. Many people consume it as a substitute for high-sugar drinks, aiming to reduce sugar intake and control weight.

In past research concerning the relationship between Diet Coke consumption and fatness, conclusions have been varied and contentious. (Ma et al. 2009) found that although sucralose does not directly stimulate the release of insulin, GLP-1, or GIP, its actions within the gastrointestinal tract may indirectly affect energy intake and metabolism by influencing gastric emptying. According to (Gardener et al. 2012), the consumption of Diet Coke is linked to higher blood sugar levels and cardiovascular diseases, suggesting that Diet Coke may promote obesity and related health issues by affecting metabolic pathways. (Peters et al. 2015) compared the effects of water and non-nutritive sweetened beverages on obesity and insulin sensitivity, discovering that substitute sugar sweeteners in beverages might not deliver the anticipated health benefits. Lastly, (Wu et al. 2023) indicated that excessive consumption of Diet Coke is significantly associated with metabolic dysfunction-related fatty liver disease (MASLD), with BMI possibly playing a mediating role in this relationship.

In previous studies examining the relationship between Diet Coke consumption and obesity, researchers primarily relied on either observational data or randomized controlled trials (RCTs). While these studies have highlighted a correlation between Diet Coke consumption and weight gain, they exhibit significant limitations. Many studies did not adequately account for confounding variables, potentially leading to misunderstandings about the actual effects of Diet Coke.

To overcome the limitations of previous studies and to more accurately decipher the causal relationship between Diet Coke consumption and obesity, this paper employs advanced causal inference techniques. Initially, utilizing the framework of Structural Causal Models (SCM), this paper constructs a causal diagram (Pearl 1995; Spirtes et al. 2000; Pearl 2009; Koller and Friedman 2009) through the IC^* algorithm (Pearl 2009). This diagram clearly illustrates both the direct and indirect relationships between Diet Coke consumption and obesity, along with the effects of potential confounders such as hyperlipidemia, hypertension, and age. Subsequently, we precisely identify and control for back-door paths in the causal diagram that could introduce bias into our estimates. This approach enables us to estimate the RCT distributions of Diet Coke effects by applying the back-door criterion and its adjustment formulas (Pearl 1995).

Building on the foundational work of (Tian and Pearl 2000), who proposed bounds on probabilities of causation, and further refined by (Mueller, Li, and Pearl 2022)'s use

of covariates to narrow these bounds, this paper calculates the Probability of Necessity and Sufficiency (PNS) across different demographics. Specifically, this involves estimating the proportion of the population for whom Diet Coke consumption leads to obesity, thereby quantifying the causal impact of Diet Coke on obesity rates within varied groups.

Preliminaries & Related Works

In this section, we review some fundamental methods of causal inference. First, we introduce the concept of the structural causal model, or SCM, as outlined in (Pearl 1995; Spirtes et al. 2000; Pearl 2009; Koller and Friedman 2009). Figure 1 exemplifies an appropriate causal diagram, which is essentially a directed acyclic graph (DAG).

Formally, a SCM consists of two sets of variables U and V, and a set of functions f that assigns each variable in V a value based on the values of the other variables in the model. In SCM, exogenous variables (U) are external and have no ancestors, depicted as root nodes in graphs. Endogenous variables (V) depend on exogenous ones and can be predicted using functions in f if all exogenous values are known.

In SCMs, the associated graphical model consists of nodes for each variable in U and V and directed edges that represent functional dependencies. If a variable X depends on Y, there is a directed edge from Y to X. These graphical models are typically directed acyclic graphs (DAGs). Causally, if Y is a parent of X in the graph, Y is a direct cause of X; if Y is an ancestor, it is a potential cause of X.

For instance, consider the following simple SCM:

$$U = \{X, Y\}, V = \{Z\}, F = \{f_Z\}$$

$$f_Z : Z = 2X + 3Y$$

This model represents the salary (Z) that an employer pays an individual with X years of schooling and Y years in the profession. X and Y both appear in f_Z, so X and Y are both direct causes of Z. If X and Y had any ancestors, those ancestors would be potential causes of Z.

The graphical model associated with it is illustrated in Figure 1.

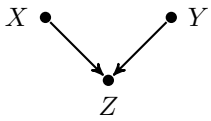


Figure 1: The graphical model of SCM, with X indicating years of schooling, Y indicating years of employment, and Z indicating salary.

One key concept of a causal diagram is called d-separation (Pearl, Glymour, and Jewell 2016).

Definition 1 (d-separation) *In a causal diagram G, a path p is blocked by a set of nodes Z if and only if*

1. *p contains a chain of nodes $A \to B \to C$ or a fork $A \leftarrow B \to C$ such that the middle node B is in Z (i.e., B is conditioned on), or*
2. *p contains a collider $A \to B \leftarrow C$ such that the collision node B is not in Z, and no descendant of B is in Z.*

If Z blocks every path between two nodes X and Y, then X and Y are d-separated conditional on Z, and thus are independent conditional on Z, denoted as $X \perp\!\!\!\perp Y \mid Z$.

With the concept of d-separation in a causal diagram, Pearl proposed the back-door criteria and its associated adjustment formula (Pearl 1995)as follows:

Definition 2 (Back-Door Criterion) *Given an ordered pair of variables (X, Y) in a directed acyclic graph G, a set of variables Z satisfies the back-door criterion relative to (X, Y), if no node in Z is a descendant of X, and Z blocks every path between X and Y that contains an arrow into X.*

If a set of variables Z satisfies the back-door criterion for X and Y, the causal effects of X on Y are given by the adjustment formula:

$$P(y|do(x)) = \sum_z P(y|x, z)P(z). \tag{1}$$

Next, we review the definitions for the three aspects of causation as defined in (Pearl 1999). We use the causal diagrams (Pearl 1995; Spirtes et al. 2000; Pearl 2009; Koller and Friedman 2009) and the language of counterfactuals in its structural model semantics, as given in (Balke and Pearl 2013; Pearl 1999; Halpern 2000).

We use $Y_x = y$ to denote the counterfactual sentence "Variable Y would have the value y, had X been x". For simplicity purposes, in the rest of the paper, we use y_x to denote the event $Y_x = y$, $y_{x'}$ to denote the event $Y_{x'} = y$, y'_x to denote the event $Y_x = y'$, and $y'_{x'}$ to denote the event $Y_{x'} = y'$. For notational simplicity, we limit the discussion to binary X and Y, extension to multi-valued variables are straightforward (Pearl 2009).

Definition 3 (Probability of necessity and sufficiency (PNS)) *(Pearl 1999)*

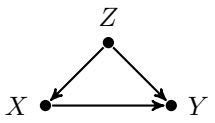
$$PNS = P(y_x, y'_{x'}) \tag{2}$$

PNS stands for the probability that y would respond to x both ways, and therefore measures both the sufficiency and necessity of x to produce y.

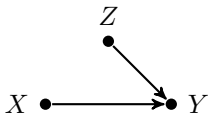
Tian and Pearl (Tian and Pearl 2000) provide a tight bound for PNS without a causal diagram. Li and Pearl (Li and Pearl 2019) provide a theoretical proof of the tight bound for PNS, and other probabilities of causation without a causal diagram.

PNS has the following tight bounds:

$$\max \left\{ \begin{array}{c} 0 \\ P(y_x) - P(y_{x'}) \\ P(y) - P(y_{x'}) \\ P(y_x) - P(y) \end{array} \right\} \leq \text{PNS} \tag{3}$$



(a) Confounder Z



(b) Outcome-affecting covariate Z

Figure 2: Z is not a descendant of X

$$\text{PNS} \leq \min \left\{ \begin{array}{c} P(y_x) \\ P(y'_{x'}) \\ P(x,y) + P(x',y') \\ P(y_x) - P(y_{x'}) + \\ +P(x,y') + P(x',y) \end{array} \right\} \tag{4}$$

Theorems 1 and 2 below provide bounds for PNS when a set Z of variables can be measured which satisfy only one simple condition: Z contains no descendant of X.

Theorem 1 *Given a causal diagram G and distribution compatible with G, let Z be a set of variables that does not contain any descendant of X in G, then PNS is bounded as follows:*

$$\sum_z \max \left\{ \begin{array}{c} 0, \\ P(y_x|z) - P(y_{x'}|z), \\ P(y|z) - P(y_{x'}|z), \\ P(y_x|z) - P(y|z) \end{array} \right\} \times P(z) \leq \textit{PNS} \tag{5}$$

$$\sum_z \min \left\{ \begin{array}{c} P(y_x|z), \\ P(y'_{x'}|z), \\ P(y,x|z) + P(y',x'|z), \\ P(y_x|z) - P(y_{x'}|z) + \\ +P(y,x'|z) + P(y',x|z) \end{array} \right\} \times P(z) \geq \textit{PNS} \tag{6}$$

In figures 2a and 2b, Z is not a descendant of X and further satisfies the back-door criterion. For such cases the PNS bounds can be simplified to read:

Theorem 2 *Given a causal diagram G and distribution compatible with G, let Z be a set of variables satisfying the back-door criterion (Pearl 2011) in G, then the PNS is bounded as follows:*

$$\sum_z \max\{0, P(y|x,z) - P(y|x',z)\} \times P(z) \leq \textit{PNS} \tag{7}$$

$$\sum_z \min\{P(y|x,z), P(y'|x',z)\} \times P(z) \geq \textit{PNS} \tag{8}$$

The significance of Theorem 2 lies in the ability to compute bounds using purely observational data, which is the situation we will face in this paper.

Methodology

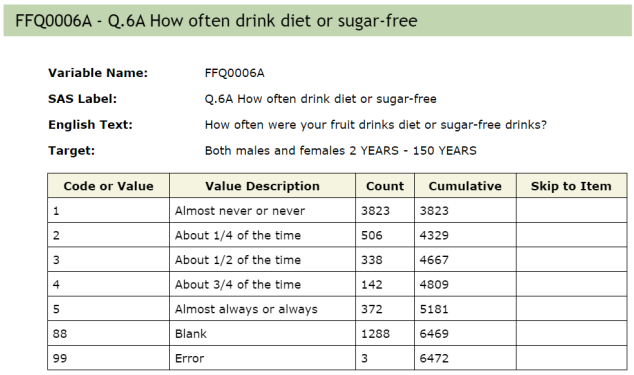
Data Collection

All data in this paper comes from the NHANES program (CDC 2023).The National Health and Nutrition Examination Survey (NHANES) is a key program of the National Center for Health Statistics (NCHS), part of the CDC, which assesses the health and nutritional status of the U.S. population. NHANES annually examines about 5,000 nationally representative individuals through both interviews and physical examinations. The survey collects data on demographic, socioeconomic, dietary, and health-related factors, along with medical measurements and laboratory tests conducted by trained medical staff.

We conducted a thorough review of all data documented in the NHANES project, closely examining the health information contained within various statistical reports. In the diet data segment, the study surveys the frequency of American consumption of a broad range of food items through questionnaires, encompassing everyday food categories such as soft drinks, alcoholic beverages, beef hamburgers, and grain salads, among others. The examination section includes data that can be measured physically, such as blood pressure, body weight, and oral health status. The laboratory tests segment encompasses assays of various biomarkers within the body, including levels of different hormones.

For the topic under discussion, our initial step involves sourcing data relevant to Diet Coke consumption and its association with fatness. Numerous studies (Adab, Pallan, and Whincup 2018; Freedman and Sherry 2009; Goulding et al. 1996), have validated the Body Mass Index (BMI) as an optimal measure for assessing obesity, hence we will utilize BMI data as our indicator of obesity in this paper. However, regarding the intake of Diet Coke, due to changes in dietary survey questionnaires, relevant data was collected only during two survey cycles between 2003 and 2006. Nonetheless, the formulation of Diet Coke has not altered over the past twenty years (The Coca-Cola Company 2023), making the use of this data both reasonable and valid for our analysis.

We hypothesize that the relationship between Diet Coke consumption and fatness is not a simple causal one but is influenced by multiple factors. Thus, it is imperative to establish a criterion for identifying indicators that may impact obesity. (Visser et al. 1997) posits that due to varying body densities, factors such as *age, race, and gender* might influence obesity. (Brown and Biosca 2016)'s research suggests that *education* can alter savings behaviors and consequently affect body fatness. According to (Akbartabartoori, Lean, and Hankey 2005), there is a relationship between *smoking* and obesity, mediated by waist and hip circumferences (WC, HC), and the waist-to-hip ratio (WHR). (Elagizi et al. 2020) argues that *physical activity* can influence an individual's weight through its effects on the cardiovascular system. (Zhou et al. 2018)'s dose-response meta-analysis indicates that within the normal range of body mass indexes, leaner individuals may have a lower risk of developing *hypertension*. (Carr and Brunzell 2004) highlights that the accumulation of abdominal fat, an indicator of obesity, is closely linked to

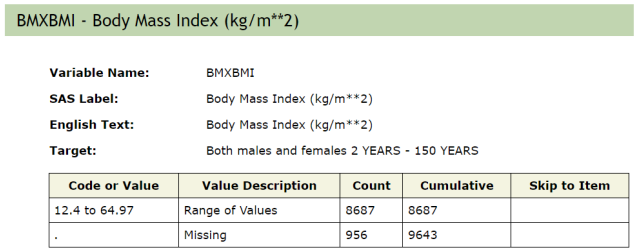


FFQ0006A - Q.6A How often drink diet or sugar-free

Variable Name: FFQ0006A
SAS Label: Q.6A How often drink diet or sugar-free
English Text: How often were your fruit drinks diet or sugar-free drinks?
Target: Both males and females 2 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
1	Almost never or never	3823	3823	
2	About 1/4 of the time	506	4329	
3	About 1/2 of the time	338	4667	
4	About 3/4 of the time	142	4809	
5	Almost always or always	372	5181	
88	Blank	1288	6469	
99	Error	3	6472	

Figure 3: Diet or sugar-free Consumption



BMXBMI - Body Mass Index (kg/m**2)

Variable Name: BMXBMI
SAS Label: Body Mass Index (kg/m**2)
English Text: Body Mass Index (kg/m**2)
Target: Both males and females 2 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
12.4 to 64.97	Range of Values	8687	8687	
.	Missing	956	9643	

Figure 4: BMI measurements

the risks of *hyperlipidemia* and Type 2 *diabetes* associated with metabolic syndrome.

Based on the aforementioned studies, the factors identified as influencing fatness include *age, gender, race/ethnicity, educational level, smoking status, average daily physical activity level, hyperlipidemia, and diabetes mellitus.*

Data Processing

We have extracted the data required for constructing the causal diagram, and before proceeding, it is essential to clarify the data type of each variable. As shown in Figure 3, the survey data on Diet Coke consumption frequency categorizes responses into five frequency levels, assigned specific codes from 1 to 5, representing a range from "Almost never or never" to "Almost always or always," with corresponding counts of respondents. Additionally, the table accounts for special cases such as non-responses ("Blank") and data entry errors ("Error").

Unlike the discrete data associated with Diet Coke consumption, the fatness component, represented by BMI data, consists of a series of continuous values that provide precise BMI measurements for each individual, as shown in Figure 4. The data formats for other influencing factors also fall into these two categories: discrete and continuous. With this data prepared, we can now proceed to construct the causal diagram.

To clearly articulate the relationships among variables in our study and ensure the accuracy of our methodology, we formally define the independent variable (X), the dependent variable (Y), and outline relevant covariates:

Category	**Variable Description**
X	Diet Coke Consumption Frequency
Y	Fatness (BMI)
Covariates	Age
	Gender
	Race / Ethnicity
	Education Level
	Smoking Status
	Average Daily Physical Activity Level
	Hyperlipidemia
	Diabetes Mellitus

Table 1: Summary of Variables

The independent variable (X) is "Diet Coke consumption frequency." This variable quantifies the frequency at which participants consume Diet Coke over a specified period.

The dependent variable (Y) is "fatness," measured by the Body Mass Index (BMI). BMI is calculated based on participants' weight and height to assess their level of fatness.

Additionally, the study will include other covariates that may affect Y, such as age, gender, race/ethnicity,educational level, and health status factors like smoking and physical activity. These factors may confound or modify the relationship between Diet Coke consumption and fatness. By incorporating these covariates, we can more accurately identify and interpret the impact of Diet Coke consumption frequency (X) on fatness (Y).

For the measurement of various health indicators, authoritative standards are essential. According to (Centers for Disease Control and Prevention 2022), a Body Mass Index (BMI) of 30.0 or higher categorizes an individual within the obesity range. Hyperlipidemia, or high cholesterol, is characterized by certain thresholds in cholesterol measurements (Clinic 2024). It is considered high when total cholesterol exceeds 240 mg/dL, LDL (low-density lipoprotein) cholesterol is 160 mg/dL or higher, and triglycerides are above 200 mg/dL. HDL (high-density lipoprotein) cholesterol is ideally above 60 mg/dL, as lower levels may also indicate risk. The A1C test measures the percentage of glycated hemoglobin, providing an average blood glucose level over three months. Levels of 5.7% to 6.4% indicate prediabetes, while 6.5% or higher confirms diabetes.

Causal Diagram

In the field of causal inference, the IC^* algorithm is a pivotal tool for identifying causal relationships among variables from observational data (Pearl 2009). This study employs the IC^* algorithm to construct a causal graph, delineating the causal structures between variables X,Y, and several covariates. The initial step of the IC^* algorithm involves con-

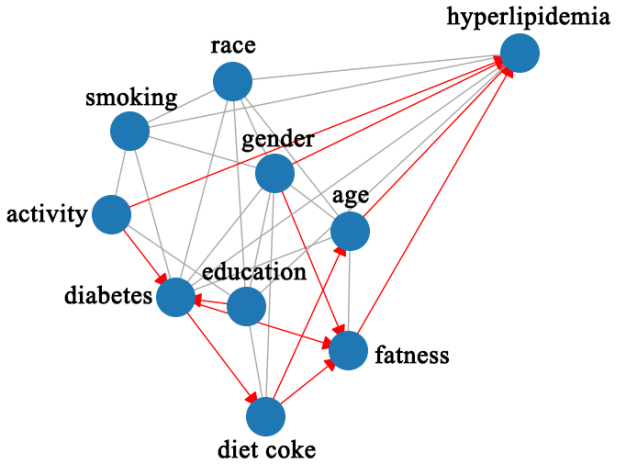


Figure 5: Causal Diagram(IC^* algorithm)

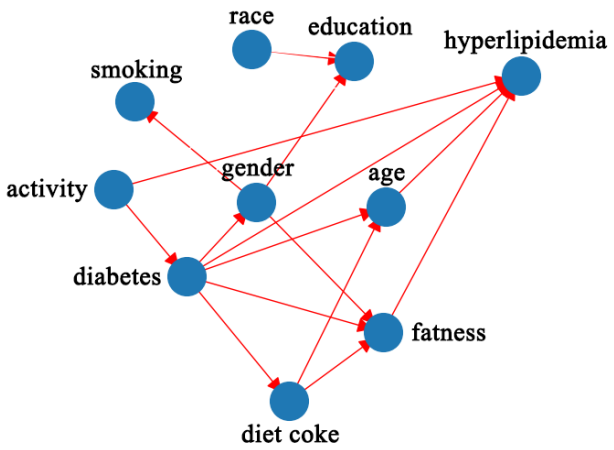


Figure 6: Causal Diagram($NO\ TEARS$ algorithm)

ducting conditional independence tests to analyze relationships among all pairs of variables within the dataset. This process relies on statistical tests to evaluate whether different combinations of variables are independent; if two variables are found to be dependent given other variables, an undirected edge is established between them.

Subsequently, the algorithm enters the structure identification phase, using conditional independence tests—specifically $theRobustRegressionTest$—to detect "$V - structures$" within the data. These structures are unique triplets where one variable is a common cause of the other two variables, which do not directly interact with each other. Identifying these V-structures is crucial for determining the direction of the edges in the graph, as they reveal the causal pathways, either direct or through intermediary variables.

Furthermore, the IC^* algorithm is capable of identifying latent variables within the causal graph and refining the graph structure through iterative optimization processes. In each iteration, the algorithm evaluates whether adding or adjusting the direction of edges better conforms to the evidence of conditional independence from the data. This method allows the algorithm to gradually construct a detailed and accurate map of the causal relationships between the variables. By applying the IC^* algorithm, we successfully portray the complex causal network among variables X, Y, and the covariates, providing a solid foundation for further analysis and model construction.The resulting causal graph is depicted in Figure 5. Red arrows denote clear causal chains, while solid black lines indicate potential but uncertain connections.

To ensure the accuracy of this causal diagram, we employed the DAG with NO TEARS algorithm (Zheng et al. 2018) for verification. This state-of-the-art structural learning method constructs directed acyclic graphs (DAGs) of causal relationships between variables by understanding their conditional dependencies. As shown in Figure 6, a comparison reveals that the core components of the two causal diagrams are nearly identical, thereby validating the reliability of the causal graph generated by the IC^* algorithm.

RCT Calculation

In this section, we focus on calculating counterfactual values (Pearl, Glymour, and Jewell 2016), which represent potential outcomes under the theoretical full control of a specific variable, such as the frequency of Diet Coke consumption. Counterfactual analysis is a central concept in causal inference, enabling us to explore the potential relationships between variables under various interventional scenarios. This means assessing the outcomes when one factor is altered while all others are held constant.

For variable X, representing the frequency of Diet Coke consumption, the original data categorize consumption into five levels ranging from 1 to 5. For computational convenience, we binarized these data, designating a frequency of 1 — indicating those who never drink Diet Coke — as 0, and marking all other levels as 1. For Y (fatness), individuals are classified as 0 for non-obese and 1 for obese, following the criteria previously described.

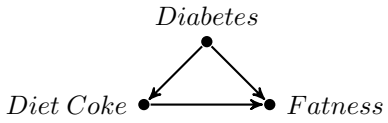


Figure 7: The segment of the causal diagram that satisfies the backdoor criterion.

Upon analyzing the causal diagram derived earlier, we identified a single backdoor pathway as shown in Figure 7, which is characterized by the pathway $DietCoke \rightarrow$

$Diabetes \leftarrow fatness$. Consequently, controlling for diabetes, the only confounder in this pathway, allows us to use the adjustment formula (Pearl 1995) to compute the experimental data. This calculation estimates the causal impact of Diet Coke consumption on fatness, controlled for diabetes status. These are denoted as $P(Y = 1 \mid \text{do}(X = 1))$ and $P(Y = 1 \mid \text{do}(X = 0))$, also represented in our notation as y_x and $y_{x'}$.

PNS Calculation

First of all, We define the following variables and their states:

X: Represents the consumption status of Diet Coke. This is a binary variable where:

- $x = 1$ indicates that Diet Coke is consumed.
- $x' = 0$ or x' indicates that Diet Coke is not consumed.

Y: Represents the obesity status of an individual. This is also a binary variable where:

- $y = 1$ indicates that the individual is obese.
- $y' = 0$ or y' indicates that the individual is not obese.

With the basic variables and states defined, we use counterfactual notation to explore hypothetical changes in these states and their impact on outcomes:

$Y_x = y$: This expression is used to describe a scenario where if X were set to x (consuming Diet Coke), then Y would take the value y (obese).

$Y_{x'} = y$: This expression describes another scenario where, even without consuming Diet Coke ($X = x'$), the individual still reaches an obese state ($Y = y$).

Building on the foundation laid by the clear definitions and the use of counterfactual notation, the computation of the Probability of Necessity and Sufficiency (PNS) in our study is pivotal for understanding the causal impact of Diet Coke consumption on obesity.

PNS evaluates the likelihood that consuming Diet Coke is both necessary and sufficient for the occurrence of obesity within various demographics. **Specifically, it measures the probability that obesity would occur with Diet Coke consumption and would not otherwise.** By calculating PNS, we aim to determine the extent to which Diet Coke consumption can be considered a direct causal factor in obesity.

With the experimental data estimated, we are now positioned to apply the model developed by (Tian and Pearl 2000) to compute the bounds of the Probability of Necessity and Sufficiency (PNS) across the general population. Furthermore, employing the method outlined by (Mueller, Li, and Pearl 2022), we will refine these estimates by narrowing the bounds, as elaborated in Formulas (3)(4)(7)(8) of the $Preliminaries$ section of this paper.

Building on this foundational analysis, we will segment the population into various subgroups based on covariates related to dietary habits and hormone levels to compute the PNS within these subgroups. When forming subgroups, we ensured that each subgroup contained at least 385 individuals. This was done to ensure that we could obtain a margin of error of at most 0.05 for the 95% confidence interval (Li, Mao, and Pearl 2022).This stratification is crucial as different dietary practices can introduce biases in the outcomes. For example, individuals with unhealthy dietary habits—such as frequent consumption of beer, burgers, and doughnuts—may inherently have a higher propensity for obesity, which should not be solely attributed to the consumption of Diet Coke. Although Diet Coke does not contain sugar, its consumption could potentially induce obesity by stimulating certain hormonal secretions or deceiving cerebral mechanisms.

By carefully considering these factors, our aim is to identify specific demographics where Diet Coke consumption significantly leads to obesity (indicating a high lower bound of PNS) or demographics where Diet Coke has virtually no impact on obesity levels (reflecting a low upper bound of PNS). This nuanced approach allows us to deliver a more precise understanding of the causal dynamics between Diet Coke consumption and obesity across different subpopulations.

Subpopulation	**Bounds of PNS**
ActivityHigh	0.10122, 0.39544
Man	0.12298, 0.40726
Age60+	0.14792, 0.43166
Old_Man	0.16865, 0.41567
Old_ActivityLow	0.14637, 0.47675
Old_HyperlipidemiaYes	0.15442, 0.40864
Old_Man_EducationLow	0.16102, 0.41829
Old_Man_ActivityLow	0.22938, 0.53263
Woman	0.08430, 0.39075
Age60-	0.09140, 0.37639
Young_Woman	0.07391, 0.37845
Young_Woman_ActivityHigh	0.05077, 0.34600

Table 2: PNS bounds by Subpopulation for Diet Coke and Fatness

Results

A/B testing, such as Randomized Controlled Trials (RCTs), is regarded as the gold standard for establishing causality in clinical and behavioral research. RCTs offer a controlled experimental setup that allows for a direct comparison, unequivocally demonstrating the direct impact of interventions, such as Diet Coke consumption. Therefore, we initially estimated the counterfactual data for the general population, providing a clear and statistically robust basis for assessing the effects of Diet Coke on obesity prevalence.

$$P(Y = 1 \mid \text{do}(X = 1)) = 0.4157292166$$
$$P(Y = 1 \mid \text{do}(X = 0)) = 0.3198820277$$

Based on the analysis of simulated RCT data, we observed a significant correlation between the consumption of Diet Coke and obesity. Specifically, the probability of

Subpopulation	Bounds of PNS
Old_Man_Hamburger	0.29909, 0.57860
Old_Man_Hotdog	0.24406, 0.48335
Old_Man_Fries	0.20028, 0.47119
Old_Man_Icecream	0.18539, 0.47663
Old_Man_Candy	0.16524, 0.37878
Old_Man_Beer	0.14397, 0.42206
Young_Woman_No_Hamburger	0.00905, 0.22395
Young_Woman_No_Popcorn	0.02081, 0.29678
Young_woman_Salad	0.07663, 0.29593
Young_Woman_No_Syrup	0.01923, 0.23076
Young_Woman_No_Fries	0.00198, 0.21941
Young_Woman_No_Hotdog	0.00000, 0.19487

Table 3: PNS bounds by Subpopulation Based on Dietary Habits

obesity among participants who consumed Diet Coke was 41.57%, compared to 31.99% among those who did not consume Diet Coke. This indicates that the consumption of Diet Coke can increase the likelihood of obesity by approximately 9.58%. While this data suggests that Diet Coke may be a potential risk factor for obesity, the modest increase of less than 10% across the general population may not be entirely convincing in complex real-world scenarios. In this context, the method of calculating the Probability of Necessity and Sufficiency (PNS) boundaries used in this study offers a more detailed and comprehensive analysis, revealing further nuances and complexities associated with Diet Coke's impact on obesity.

Following the methods described previously, we computed and refined the boundaries of the Probability of Necessity and Sufficiency (PNS) for the general population:

$$0.096 \leq PNS \leq 0.405$$

The calculated Probability of Necessity and Sufficiency (PNS) for obesity attributable to Diet Coke consumption ranges from 0.096 to 0.405. The lower bound of approximately 10% suggests that there indeed exists a subset of individuals for whom consuming Diet Coke contributes to obesity. However, the broad range of this boundary indicates variability in the causal impact across the population.

To refine our understanding and identify potentially more characteristic subgroups, we will next calculate the PNS boundaries for various subpopulations. This approach aims to discover specific demographics or behavioral patterns that might exhibit a stronger or more distinct causal relationship between Diet Coke consumption and obesity.

We have selected some subpopulations where the effects are particularly pronounced, as shown in Table 2. After controlling for certain covariates, the boundaries of the Probability of Necessity and Sufficiency (PNS) underwent significant changes. For instance, in the subgroup labeled as $Old\_Man\_ActivityLow$, which refers to men over the age of 60 with low physical activity levels, the lower boundary of the PNS increased by more than double compared to the general population. This indicates that the impact of Diet Coke on weight is more significant in this subgroup.

As shown in Table 3, we further subdivided the population based on dietary habits, yielding significant results. Among individuals with poor dietary habits, such as the subgroup "$Old\_Man\_Hamburger$"—elderly males who frequently consume hamburgers—the lower boundary of the Probability of Necessity and Sufficiency (PNS) rose to nearly 30%. This suggests that a higher proportion of individuals in this group experience obesity potentially linked to consuming Diet Coke, illustrating the significant probability that obesity would occur with Diet Coke consumption and would not otherwise. However, it is likely not the consumption of Diet Coke itself that leads to obesity, but rather the association of Diet Coke with a range of unhealthy dietary behaviors.

Conversely, in subgroups with healthier eating habits, such as "$Young\_Woman\_No\_Hotdog$", the upper boundary being 20% indicates that only a small fraction of this population might become obese due to Diet Coke consumption. In these cases, the practice of consuming Diet Coke while maintaining weight stability appears to be more credible. This differential impact underscores the complex interactions between dietary habits and the effects of specific dietary choices such as Diet Coke on health outcomes.

Discussion

In this paper, we employed SCM and the IC^* algorithm to construct causal diagrams and calculated the probability of interventional effects using adjustment formulas.Subsequently, we proceeded to assess the probability that obesity would occur with Diet Coke consumption and would not occur otherwise. However, our work still has some limitations.

Our ability to perform subgroup analyses was constrained by the total sample size. Each subgroup needed to contain at least 385 individuals to ensure the reliability of the statistical results (Li, Mao, and Pearl 2022). This limitation prevented us from conducting more refined groupings, making it difficult to precisely identify which populations are more or less suitable for consuming Diet Coke. Moreover, although studies (Sylvetsky et al. 2020) suggest that Diet Coke may influence obesity through the induction of hormone secretion, we were unable to obtain specific data on hormone levels, which restricted our comprehensive understanding of its potential health impact mechanisms.

Future research should expand the sample size, including a broader range of regions and populations, to enhance the universality and accuracy of the study findings. Additionally, further exploration of the specific effects of Diet Coke on hormone levels, particularly its potential impact on key hormones such as insulin and leptin, would help deepen our understanding of its relationship with obesity and metabolic health. Future studies could also consider the connections

between Diet Coke consumption and other chronic diseases such as cardiovascular diseases and diabetes.

Through this paper, we aim to provide readers with a comprehensive perspective on the complex relationship between Diet Coke consumption and obesity, as well as the potential implications of these findings for health policies and individual choices.

Conclusion

Through the application of causal inference methods, our study has determined that Diet Coke consumption indeed poses a risk of increased obesity; however, the impact varies significantly across different demographics. Different subgroups experience varying rates of weight gain after consuming Diet Coke, suggesting that the influence of Diet Coke on weight is not a straightforward causal relationship but is likely mediated through multiple factors. These findings emphasize the importance of considering individual differences and dietary habits when assessing the health impacts of foods or beverages. This complexity highlights that the effect of Diet Coke on weight gain is the result of multifaceted interactions, underscoring the necessity to take into account personal health profiles and lifestyles in dietary recommendations.

References

Adab, P.; Pallan, M.; and Whincup, P. H. 2018. Is BMI the best measure of obesity? *BMJ*, 360.

Akbartabartoori, M.; Lean, M.; and Hankey, C. 2005. Relationships between cigarette smoking, body size and body shape. *International journal of obesity*, 29(2): 236–243.

Balke, A.; and Pearl, J. 2013. Counterfactuals and Policy Analysis in Structural Models. arXiv:1302.4929.

Brown, H.; and Biosca, O. 2016. Exploring the relationship between time preference, body fatness, and educational attainment. *Social science & medicine*, 158: 75–85.

Carr, M. C.; and Brunzell, J. D. 2004. Abdominal obesity and dyslipidemia in the metabolic syndrome: importance of type 2 diabetes and familial combined hyperlipidemia in coronary artery disease risk. *The journal of clinical endocrinology & metabolism*, 89(6): 2601–2607.

CDC. 2023. About the National Health and Nutrition Examination Survey. Accessed: 2023-05-09.

Centers for Disease Control and Prevention. 2022. Assessing Your Weight: Adult BMI. https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html. Accessed: 2024-05-11.

Clinic, M. 2024. High blood cholesterol - Diagnosis and treatment. Accessed: 2024-05-14.

Elagizi, A.; Kachur, S.; Carbone, S.; Lavie, C. J.; and Blair, S. N. 2020. A review of obesity, physical activity, and cardiovascular disease. *Current obesity reports*, 9: 571–581.

Freedman, D. S.; and Sherry, B. 2009. The Validity of BMI as an Indicator of Body Fatness and Risk Among Children. *Pediatrics*, 124(Supplement): S23–S34.

Gardener, H.; Rundek, T.; Markert, M.; Wright, C.; Elkind, M.; and Sacco, R. 2012. Diet Soft Drink Consumption is Associated with an Increased Risk of Vascular Events in the Northern Manhattan Study. *Journal of general internal medicine*, 27: 1120–6.

Goulding, A.; Gold, E.; Cannan, R.; Taylor, R.; Williams, S.; and Lewis-Barned, N. 1996. DEXA supports the use of BMI as a measure of fatness in young girls. *International journal of obesity and related metabolic disorders: journal of the International Association for the Study of Obesity*, 20(11): 1014–1021.

Halpern, J. Y. 2000. Axiomatizing causal reasoning. *Journal of Artificial Intelligence Research*, 12: 317–337.

Koller, D.; and Friedman, N. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Li, A.; Mao, R.; and Pearl, J. 2022. Probabilities of Causation: Adequate Size of Experimental and Observational Samples. arXiv:2210.05027.

Li, A.; and Pearl, J. 2019. Unit Selection Based on Counterfactual Logic. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 1793–1799. International Joint Conferences on Artificial Intelligence Organization.

Ma, J.; Bellon, M.; Wishart, J.; Young, R.; Blackshaw, L.; Jones, K.; Horowitz, M.; and Rayner, C. 2009. Effect of the artificial sweetener, sucralose, on gastric emptying and incretin hormone release in healthy subjects. *American journal of physiology. Gastrointestinal and liver physiology*, 296: G735–9.

Mashed. 2023. The Untold Truth Of Diet Coke. https://www.mashed.com/the-untold-truth-of-diet-coke. Accessed: 2024-05-08.

Mueller, S.; Li, A.; and Pearl, J. 2022. Causes of Effects: Learning Individual Responses from Population Data. In Raedt, L. D., ed., *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 2712–2718. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Pearl, J. 1995. Causal diagrams for empirical research. *Biometrika*, 82(4): 669–688.

Pearl, J. 1999. Probabilities of Causation: Three counterfactual interpretations and their identification. *Synthese*, 121(1-2): 93–149.

Pearl, J. 2009. *Causality*. Cambridge University Press, 2 edition.

Pearl, J. 2011. Aspects of graphical models connected with causality.

Pearl, J.; Glymour, M.; and Jewell, N. P. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Peters, J.; Courtney, J.; Cardel, M.; Wyatt, H.; Foster, G.; Pan, Z.; Wojtanowski, A.; Veur, S.; Herring, S.; Brill, C.; and Hill, J. 2015. The effects of water and non-nutritive sweetened beverages on weight loss and weight maintenance: A randomized clinical trial. *Obesity*, 24.

Spirtes, P.; Glymour, C. N.; Scheines, R.; and Heckerman, D. 2000. *Causation, Prediction, and Search*. MIT Press.

Spren. 2023. Brief History of Coke Zero and Diet Coke. https://www.spren.com/brief-history-of-coke-zero-and-diet-coke. Accessed: 2024-05-08.

Sylvetsky, A. C.; Sen, S.; Merkel, P.; Dore, F.; Stern, D. B.; Henry, C. J.; Cai, H.; Walter, P. J.; Crandall, K. A.; Rother, K. I.; et al. 2020. Consumption of diet soda sweetened with sucralose and acesulfame-potassium alters inflammatory transcriptome pathways in females with overweight and obesity. *Molecular nutrition & food research*, 64(11): 1901166.

The Coca-Cola Company. 2023. Is the Diet Coke recipe changing? Accessed: 2023-05-09.

Tian, J.; and Pearl, J. 2000. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28: 287–313.

Visser, M.; Gallagher, D.; Deurenberg, P.; Wang, J.; Pierson Jr, R.; and Heymsfield, S. 1997. Density of fat-free body mass: relationship with race, age, and level of body fatness. *American Journal of Physiology-Endocrinology and Metabolism*, 272(5): E781–E787.

Wu, Y.; Tan, Z.; Zhen, J.; Liu, C.; Zhang, J.; Liao, F.; and Dong, W. 2023. Association between diet soft drink consumption and metabolic dysfunction-associated steatotic liver disease: findings from the NHANES. *BMC Public Health*, 23.

Zheng, X.; Aragam, B.; Ravikumar, P.; and Xing, E. P. 2018. DAGs with NO TEARS: Continuous Optimization for Structure Learning. arXiv:1803.01422.

Zhou, W.; Shi, Y.; Li, Y.-q.; Ping, Z.; Wang, C.; Liu, X.; Lu, J.; Mao, Z.-x.; Zhao, J.; Yin, L.; et al. 2018. Body mass index, abdominal fatness, and hypertension incidence: a dose-response meta-analysis of prospective studies. *Journal of human hypertension*, 32(5): 321–333.