The Table 2 Fallacy and Overfitting: A Persistent Problem in Contemporary Research?

Víctor Juan Vera-Ponce[1,2,*], Jhosmer Ballena-Caicedo[1,2], Lupita Ana Maria Valladolid-Sandoval[1,2], Fiorella E. Zuzunaga-Montoya[1,2] and Carmen Inés Gutierrez De Carrillo[1,2]

[1]*Instituto de Investigación de Enfermedades Tropicales, Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM), Amazonas, Perú*

[2]*Facultad de Medicina (FAMED), Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM), Amazonas, Perú*

Abstract: The "Table 2 fallacy" represents a common methodological error in medical research, characterized by indiscriminate statistical adjustment for multiple variables without considering their causal nature. This article examines the theoretical foundations of the problem, distinguishing between studies with descriptive, predictive, and explanatory objectives, and emphasizing how the research purpose should determine the adjustment strategy. We highlight the fundamental role of Directed Acyclic Graphs (DAGs) in correctly identifying confounding, mediating, and colliding variables, thus avoiding overadjustment and resulting biases. To illustrate these considerations, we present two practical examples: the relationship between obesity and colorectal cancer, and between coffee consumption and breast cancer. In the first case, we demonstrate how adjustment for intestinal dysbiosis (a mediator) can attenuate the association between obesity and colorectal cancer, reducing the adjusted relative risk from 1.78 (95% CI: 1.20–2.65) to 1.49 (95% CI: 0.97–2.29) and eliminating statistical significance (p=0.072). In the second example, we show how including insomnia (a collider) in the model can create artificial associations between coffee consumption and breast cancer, dramatically increasing the adjusted relative risk to 1.94 (95% CI: 1.34-2.81) with high statistical significance (p<0.001), when a correctly specified model shows no such association. We conclude that, in explanatory studies, it is essential to develop causal reasoning prior to statistical analysis, using DAGs to guide the selection of adjustment variables. This rigorous methodological approach prevents both the dilution of real causal effects and the generation of spurious associations, increasing the internal validity of epidemiological findings and their utility for clinical decision-making.

Keywords: Bias, Confounding Factors, Epidemiologic, Mediation Analysis, Causality.

INTRODUCTION

The "Table 2 fallacy" describes a common error in epidemiological and clinical studies where results of multivariate models—typically presented in Table 2—include inappropriate adjustments. The problem arises when researchers control not only for confounding variables but also for mediators (factors in the causal pathway between exposure and outcome) or colliders (variables influenced by multiple causes), distorting the true magnitude of associations and leading to erroneous causal conclusions [1].

While the scientific literature emphasizes adjusting for confounders to isolate exposure effects, other variable types are often incorrectly handled. Inappropriate adjustment for mediators or colliders causes overadjustment bias and generates spurious associations, resulting in biased estimates that compromise interpretation [2].

*Address correspondence to this author at the Instituto de Investigación de Enfermedades Tropicales, Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM), Amazonas, Perú;
E-mail: vicvepo@gmail.com

This phenomenon affects any discipline that employs multivariate analyses, including nutritional epidemiology, oncology, public health, and clinical research [3-5]. The root problem is the absence of a solid causal framework and confusion between variable types, which can lead to underestimating or overestimating exposure effects, thereby compromising internal validity.

This review explores the conceptual foundations of the Table 2 fallacy, provides empirical evidence of the issue, and offers practical recommendations for prevention. Understanding variable types and applying causal logic before statistical adjustment yields more accurate estimates, thereby reinforcing research validity and providing direct implications for clinical and public health decision-making.

Theoretical and Conceptual Foundations

A key point often overlooked is aligning the study objective with the statistical adjustment strategy employed. In the medical literature, it is common to find articles that mix descriptive, predictive, and explanatory objectives, leading to confusion and unhelpful conclusions [6]. Each of these purposes requires a

distinct approach to variable selection and the application of statistical models [7].

It is crucial to distinguish between two uses of the term "overfitting" that may cause confusion. In predictive modeling and machine learning, "statistical overfitting" refers to a model that fits the training data too closely, capturing noise rather than true patterns, thus performing poorly on new data [8]. In contrast, "overadjustment bias" in causal inference—the focus of this manuscript—occurs when controlling for variables that mediate or collide causal pathways, thereby distorting the true causal effect estimate. While both involve "too much adjustment," they represent fundamentally different problems: statistical overfitting is about prediction accuracy, whereas overadjustment bias is about causal validity. Throughout this manuscript, we use "overfitting" and "overadjustment" to refer to the causal inference problem, not the statistical prediction problem.

When the primary interest is to describe the frequency of a disease or compare its distribution among different groups (regions, socioeconomic levels, etc.), the emphasis is not on isolating the "pure" effect of a particular factor or establishing causality. In this scenario, it is neither necessary nor advisable to include numerous adjustment variables in the analysis, as the objective is limited to showing reality as it is. For example, suppose one wants to document the prevalence of a disease in different regions. In that case, it is sufficient to present the percentages or rates for each area, without the need to correct for multiple factors [9].

At the opposite end, when attempting to build a predictive model—for example, to estimate the probability that a patient will develop a certain pathology—the selection of variables is justified mainly by their ability to improve the accuracy of the prediction. Here, the causal role of each variable is less relevant; in fact, factors can be included that simply increase the model's ability to "guess" the outcome, even though their biological significance may be uncertain. Even so, it is essential to explicitly state that the objective is merely predictive so that these associations are not confused with cause-effect relationships [10].

Finally, in studies of an explanatory or etiological nature, the priority is to understand why an outcome occurs. Under this approach, the selection of variables and the adjustment strategy must respond to causal reasoning, with tools such as Directed Acyclic Graphs (DAGs) guiding the choice of control factors. The goal is to estimate the genuine effect of the exposure of interest, minimizing the influence of other variables that could distort the relationship being studied. This perspective is of great value in the clinical field, where elucidating the role of a risk factor can motivate preventive interventions or guide therapeutic decision-making [11, 12].

In summary, clarity regarding the objective type—descriptive, predictive, or explanatory—is fundamental for precisely selecting the variables to be included in the statistical analysis. A merely associative study without a well-defined purpose runs the risk of making unnecessary adjustments or overlooking important factors for the question posed. Having clear objectives from the outset optimizes study design, facilitates the interpretation of results, and produces more robust knowledge applicable to medical practice [13].

This article emphasizes explanatory or etiological studies, as they most frequently present the so-called "Table 2 fallacy" by attempting to control for multiple variables without adequate causal justification. Precisely, in this type of study—where the aim is to isolate the effect of a risk factor or estimate the magnitude of a causal relationship—it is essential to distinguish between confounding and overfitting and to establish a solid conceptual model. At this point, DAGs become relevant, as they facilitate the selection of adjustment variables based on causal logic, preventing both the omission of important confounders and inappropriate adjustment for mediators or colliders.

The Role of Directed Acyclic Graphs (DAGs)

Designing a DAG is extremely useful in undertaking an explanatory study that clarifies the causal relationship between an exposure and an outcome. These diagrams have the quality of presenting in a visual and organized manner, in which variable influences are always present in a unidirectional way and without forming loops. In this way, the study's causal hypothesis is captured, showing the possible pathways through which exposure might affect the outcome and the factors that may interfere with this relationship [14].

Within this representation, variables can occupy different positions. One of these is the confounding variable, which is associated with both the exposure of interest and the outcome, without being part of the

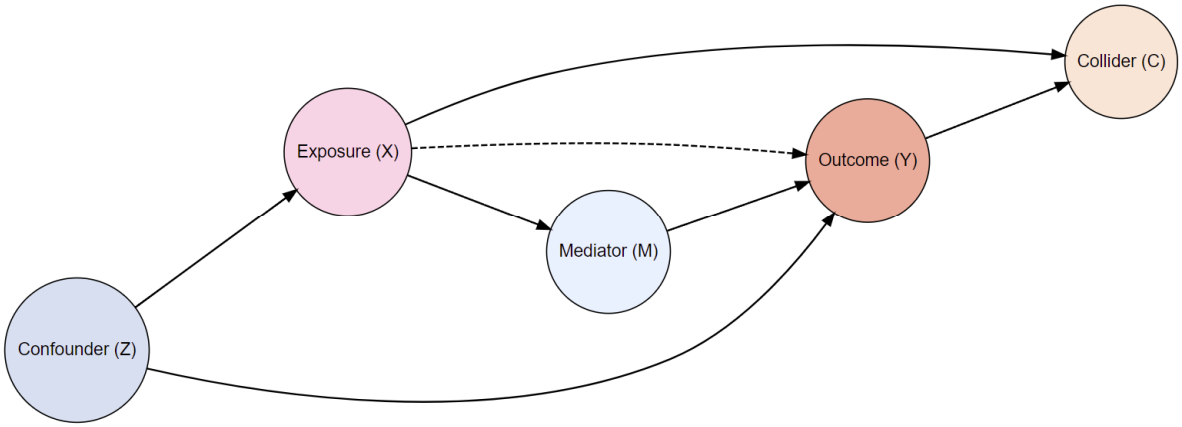

Figure 1: Directed acyclic graph.

causal pathway between them. Its presence can mask or exaggerate the existing relationship. A classic example consists of studying the association between coffee consumption (exposure) and lung cancer (outcome): if smoking appears in the DAG as a factor that influences coffee consumption and, at the same time, increases the risk of lung cancer, it is identified as a confounder. Adjusting for smoking in the analysis allows for better isolation of the effect of coffee itself on lung cancer [15, 16].

Another fundamental role that can be illustrated in a DAG is that of the mediating variable, which forms part of the causal pathway connecting the exposure to the outcome. In graphical terms, the mediator receives an arrow from the exposure and, in turn, points toward the outcome. Adjusting for this variable can cause overfitting when the main interest is knowing the total effect of the exposure. For example, if studying the association between obesity (exposure) and type 2 diabetes mellitus (outcome), insulin resistance, which is caused by obesity and which, in turn, increases the risk of diabetes, would be positioned as a mediator. Including it in the model means underestimating the true magnitude of the relationship between obesity and diabetes if the objective is to understand the global etiology of the disease [15, 16].

Finally, the DAG also facilitates the identification of a "collider" or colliding variable, which arises when two (or more) different causal factors converge on it [15, 16]. Visually, arrows directed toward the collider from more than one variable can be distinguished. Improperly adjusting for a collider can, paradoxically, generate a spurious association that did not exist before [17]. To give another specific example, this happens if studying the relationship between socioeconomic level and the presence of obesity, but a variable that is a consequence of both (for example, diet quality) is included in the adjustment, opening an artificial pathway of association and distorting the results instead of clarifying them.

In summary, developing a solid DAG coherent with prior theory or biological evidence guides the researcher on when it is appropriate or not to adjust for certain variables. In this way, biases such as overfitting or the appearance of spurious correlations are avoided, and the "Table 2 fallacy" is prevented. This graphical and logical-causal approach leads to more reliable interpretations by clearly distinguishing whether a variable should be treated as a confounder, mediator, or collider, based on its position in the causal network.

To operationalize these concepts in practice, Figure **2** presents a decision tree that systematically guides researchers through the process of identifying whether a variable should be included in the adjustment set. This tool assumes that a preliminary DAG has been constructed based on subject-matter knowledge, biological plausibility, and existing literature. The decision tree should be used in conjunction with, not as a replacement for, careful causal reasoning. Each pathway through the tree leads to a clear recommendation: adjust only for true confounders (shown in green), while avoiding adjustment for mediators, colliders, and unrelated variables (shown in red/coral), thereby preventing the Table 2 fallacy.

Caption: Decision tree for determining adjustment strategy in explanatory studies. Variables should be classified based on their causal relationships established through DAG construction, not their statistical associations. Green endpoints indicate variables requiring adjustment; red/coral endpoints indicate variables that should not be adjusted to avoid bias.

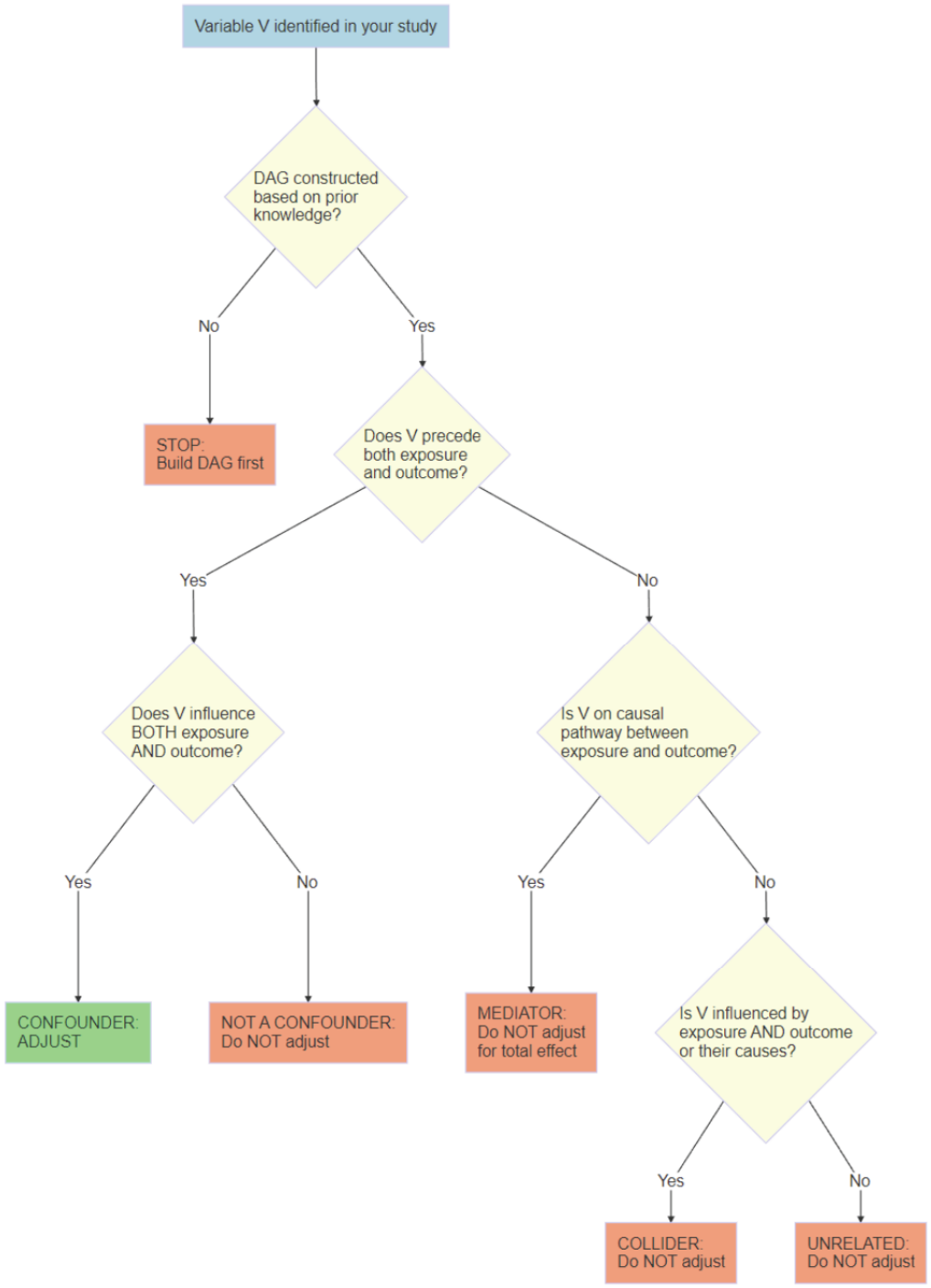

Figure 2: Decision tree for variable adjustment in causal inference studies.

First Example: Obesity And Colorectal Cancer

Below, we present an applied example to illustrate using a DAG in epidemiological research, focusing on the relationship between obesity and colorectal cancer (Figure **3**). This association is particularly relevant due to the sustained increase in obesity prevalence and the high incidence of colorectal neoplasms worldwide. Although the model we will describe is a simplified version of reality, it offers a didactic guide to understanding which variables to control for and which we should not adjust based on their role as potential confounders, mediators, or even colliders.

In the diagram, we position obesity as the exposure that could trigger a carcinogenic process in the colon, but we also recognize the importance of age as a key confounder. With the passage of years, both the

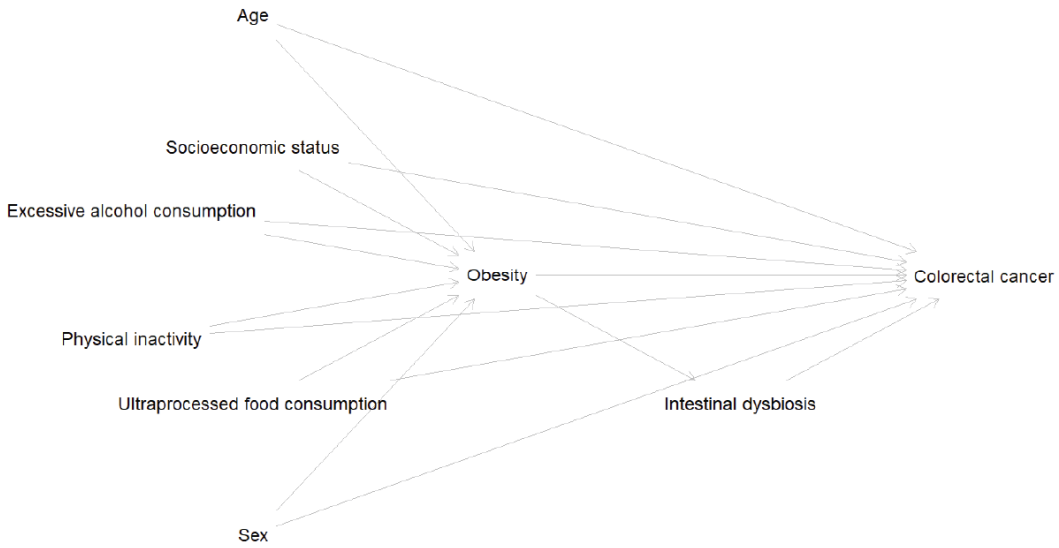

Figure 3: Directed acyclic graph of obesity as a risk for colorectal cancer.

probability of suffering from obesity and developing colorectal cancer increase, so not adjusting for age could distort the true association between obesity and cancer. Another factor we consider is socioeconomic status: people with lower socioeconomic status tend to have a higher risk of obesity, partly due to limitations in access to healthy foods and opportunities for physical activity; at the same time, they could face greater difficulties in accessing health services that favor screening and early detection of cancer. Ultra-processed food consumption is also relevant, as it directly relates to a higher risk of obesity and, potentially, to the presence of carcinogenic compounds that increase the probability of developing a colorectal tumor. Sex, for its part, is incorporated as another factor that influences both susceptibility to obesity (due to differences in fat distribution and metabolism) and the probability of presenting colorectal cancer, perhaps mediated by various hormonal aspects.

To represent a possible mediating pathway within the relationship between obesity and colorectal cancer, we add intestinal dysbiosis. Obesity could alter the composition of the microbiota, triggering a state of chronic inflammation and the production of carcinogenic metabolites that, in turn, increase the risk of colorectal neoplasia. Considering dysbiosis as a mediator implies that if we adjust for this factor when the interest is to estimate the total effect of obesity, we would be underestimating its real contribution to colorectal cancer, as we suppress part of the causal pathway.

Next, two analyses perfectly illustrate the difference between adjusting according to the DAG (avoiding over adjusting for mediating or colliding variables) and an "associated factors" approach in which almost any related variable is introduced, even without discriminating its role in the causal chain. Table **1** could correspond to the DAG-based model, while Table **2** would reflect the expanded model of associated factors.

Table 1: Adjusted Association between Obesity and Colorectal Cancer Incidence Based on Dag-Informed Confounding Control

Characteristic	aRR*	95% CI	p-value
Obesity			
No	Ref.	—	—
Yes	1.78	1.20 - 2.65	<0.001

*Obesity has been adjusted for age, sex, socioeconomic status, and ultra-processed food consumption as identified through directed acyclic graph analysis of causal pathways.
aRR: Adjusted risk ratio. 95% CI: 95% Confidence Interval.

In Table **1**, the adjustment is limited to those variables considered true confounders in the DAG (for example, age, socioeconomic status, ultra-processed food consumption, or physical activity if they have been identified as confounders, but intestinal dysbiosis is not included as it is a possible mediator). Under this scheme, the association between obesity and the outcome (colorectal cancer) remains statistically significant: obesity shows an aRR of 1.78 (95% CI: 1.20–2.65), indicating a substantial increase in relative risk after considering the relevant confounders. This finding supports the hypothesis that obesity plays a pertinent role in the etiology of colorectal cancer, at least in the context of the adjusted variables.

In contrast, Table **2** illustrates the "associated factors" approach, in which multiple variables are introduced without a clear causal criterion, including possible mediators (such as intestinal dysbiosis) or variables that might not act as confounders. When "adjusting for everything," the estimated effect of obesity decreases to an aRR of 1.49 (95% CI: 0.97–2.29) and loses statistical significance (p=0.072). From a quick reading, one could erroneously conclude that obesity "is not associated" with the outcome. However, this drop in the strength of the association is largely due to the inclusion of a possible mediating variable (dysbiosis), "blocking" part of the real effect that obesity exerts on colorectal cancer. This phenomenon exemplifies the Table 2 fallacy or "overfitting." When controlling for variables in the causal pathway, the signal of the main exposure is diluted, and a relationship that does exist can be masked.

Table 2: Comprehensive Multivariable Analysis of Potential Risk Factors for Colorectal Cancer: Fully Adjusted Poisson Regression Model

Characteristic	aRR	95% CI	p-value
Obesity			
No	Ref.	—	—
Yes	1.49	0.97 - 2.29	0.072
Age			
< 45	Ref.	—	—
≥ 45	4.45	2.89 – 7.16	< 0.001
Physical activity			
No	Ref.	—	—
Yes	1.23	0.80 - 1.90	0.354
Ultra-processed food consumption			
No	Ref.	—	—
Yes	1.28	0.94 - 1.75	0.111
Intestinal dysbiosis			
No	Ref.	—	—
Yes	1.52	0.96 - 2.40	0.074
Socioeconomic status			
Low	Ref.	—	—
Moderate/High	0.81	0.50 - 1.29	0.366
Alcohol consumption			
Low	Ref.	—	—
High	1.48	0.90 - 2.41	0.12

*Each factor has been simultaneously adjusted for all other variables in the model, including obesity, physical activity, ultra-processed food consumption, intestinal dysbiosis, socioeconomic status, and alcohol consumption.
aRR: Adjusted risk ratio. 95% CI: 95% Confidence Interval.

Additionally, it is important to note that the other variables presented in these "associated factors" models should not even be reported or interpreted. The common practice of presenting and interpreting all model covariates as if they were "independently associated factors" is conceptually flawed. Each variable in the model fulfills different roles according to the underlying causal structure (confounder, mediator, or collider); therefore, interpreting their coefficients as independent effects is meaningless. For example, the coefficient of low physical activity or alcohol consumption does not represent their true effect on colorectal cancer, as this effect is distorted by the inclusion of other variables in the causal pathway. Researchers should avoid the temptation to interpret these secondary coefficients and focus solely on estimating the primary exposure, which is the objective of an explanatory study.

In summary, Table **1** represents a model based on the DAG, in which only confounders have been identified and adjusted for, leaving out mediators. Table **2**, for its part, presents a "multifactorial" analysis that includes both confounders and mediators and other variables of possible interest. Comparing both estimates highlights the relevance of establishing a causal framework before adjustment, avoiding the Table 2 fallacy and reinforcing the validity of conclusions about the association between obesity and colorectal cancer.

Second Example: Coffee Consumption and Breast Cancer

In this new example, we position coffee consumption as an exposure factor in the relationship with breast cancer, recognizing that multiple variables can influence both coffee consumption and the development or early detection of this neoplasm. Through DAG, we identify which variables act as possible confounders and which could function as mediators or even colliders (Figure **4**). This distinction is fundamental for making justified decisions about when statistical adjustment is appropriate and when not to avoid underestimating or "inventing" associations.

First, we consider lack of physical activity (or low activity), consumption of ultra-processed foods, smoking, and socioeconomic status as possible confounders between coffee consumption and breast cancer. Physical activity and diet (represented here by ultra-processed foods) are related not only to the risk of chronic pathologies but also to behavioral patterns that

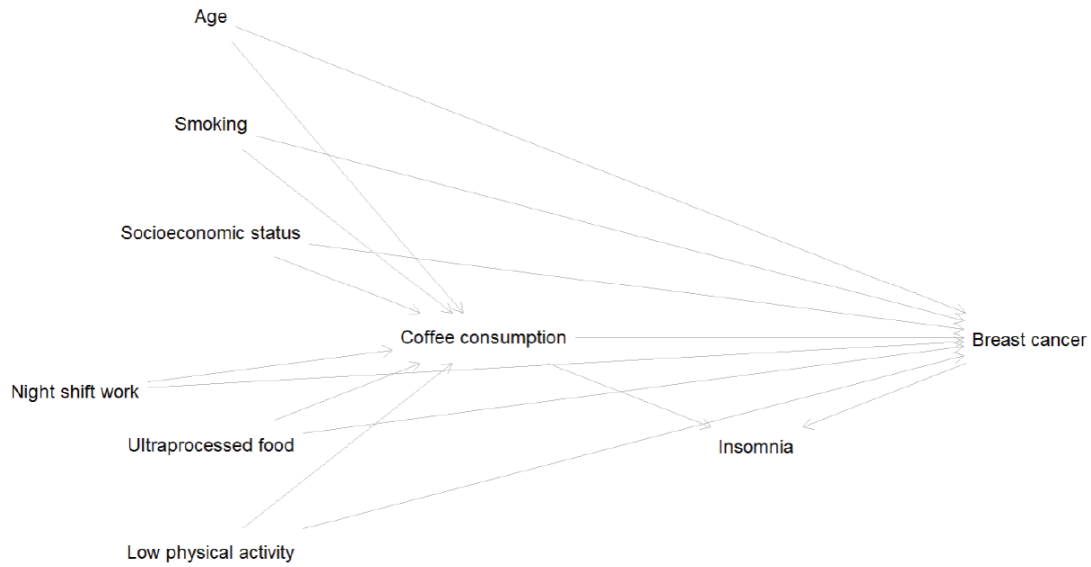

Figure 4: Directed acyclic graph of coffee consumption as a risk for breast cancer.

may include coffee intake. Smoking, for its part, usually has correlations with different habits, including coffee, and is associated with a higher risk of various neoplasms, including some types of breast cancer. Socioeconomic status influences access to food and general lifestyle while simultaneously conditioning the probability of regularly consuming coffee and accessing early cancer detection. All these variables were selected in the DAG as genuine candidates for confounding since they could distort the true relationship between coffee consumption and the neoplasm.

Similarly, night work is introduced as another factor that may share pathways with coffee consumption (e.g., higher intake in night workers) and is also associated with a differentiated risk of certain cancers due to alterations in circadian rhythms and rest habits. The resulting DAG positions these variables as factors to consider in the adjustment to isolate the specific effect of coffee on colorectal cancer.

On the other hand, insomnia emerges as a possible collider: it can be a consequence of (or be exacerbated by) high coffee intake and, at the same time, be associated with the appearance or perception of symptoms that lead to an earlier diagnosis of cancer (or with other stress and hormonal factors that affect carcinogenesis). Adjusting for insomnia in the model, when this variable receives causal arrows from coffee consumption and the mechanisms that lead to cancer, can introduce spurious associations, making it difficult to interpret the real influence of coffee.

Under this approach, Table **3** shows the analysis in which adjustment is made only for the confounders identified through the DAG (smoking, socioeconomic status, consumption of ultra-processed foods, low physical activity, and night work), leaving out insomnia due to its possible character as a collider. The aRR for coffee consumption in this scenario is 1.26 (95% CI: 0.88–1.80), with a p-value of 0.206, which indicates that there is no robust statistical evidence that coffee increases or reduces risk, but the adjustment is methodologically "clean" concerning the causal objective.

As in the previous example, it is worth emphasizing that these covariates serve exclusively as adjustment factors to obtain an unbiased estimate of the exposure-outcome relationship. Their individual coefficients should not be interpreted as independent effects on breast cancer, since their inclusion in the model is solely justified by their role as confounders in the causal framework, not as factors of interest whose direct effects we aim to quantify.

Table 3: Adjusted Association between Obesity and Breast Cancer Incidence Based on Dag-Informed Confounding Control

Characteristic	aRR*	95% CI	p-value
Coffee consumption			
No	Ref.	—	—
Yes	1.26	0.88 - 1.80	0.206

*Coffee consumption has been adjusted for smoking, socioeconomic status, ultra-processed food consumption, low physical activity, and night shift work as identified through directed acyclic graph analysis of causal pathways.
aRR: Adjusted risk ratio. 95% CI: 95% Confidence Interval.

Conversely, Table **4** presents a model of "associated factors," in which all variables are included simultaneously, including insomnia (a possible collider). The estimate for coffee consumption (aRR=1.94, 95% CI: 1.34–2.81) shows a significant increase in the relative risk ($p < 0.001$), demonstrating how the inclusion of a collider can create or amplify associations that may not reflect the true causal effect. This is particularly noteworthy compared to the more conservative and methodologically sound estimate in Table **1**. Additionally, variables such as low physical activity or socioeconomic status acquire a statistically significant role consistent with their character as genuine confounders. Still, the simultaneous presence of insomnia could distort the association pathways, creating artificial relationships or magnifying existing ones.

Table 4: Comprehensive Multivariable Analysis of Potential Risk Factors for Breast Cancer: Fully Adjusted Poisson Regression Model

Characteristic	aRR	95% CI	p-value
Coffee consumption			
No	Ref.	—	—
Yes	1.94	1.34 - 2.81	< 0.001
Age			
< 45	Ref.	—	—
≥ 45	1.08	0.82 - 1.42	0.596
Insomnia			
No	Ref.	—	—
Yes	1.08	0.82 - 1.42	0.595
Smoking			
Low	Ref.	—	—
High	0.79	0.47 - 1.33	0.369
Socioeconomic status			
No	Ref.	—	—
Yes	1.89	1.47 - 2.44	<0.001
Ultra-processed food consumption			
No	Ref.	—	—
Yes	1.02	0.56 - 1.88	0.941
Low physical activity			
No	Ref.	—	—
Yes	1.58	1.18 - 2.12	0.002
Night shift work			
No	Ref.	—	—
Yes	1.64	0.95 - 2.80	0.073

*Each factor has been simultaneously adjusted for all other variables in the model, including coffee consumption, insomnia (collider), smoking, socioeconomic status, ultra-processed food consumption, low physical activity, and night shift work.
aRR: Adjusted risk ratio. 95% CI: 95% Confidence Interval.

This contrast between a DAG-informed model (Table **3**) and an indiscriminate "multifactorial" model (Table **4**) illustrates the importance of defining, from the beginning, how and why each variable enters the adjustment. While the first approach is based on causal reasoning that isolates only confounders, the second adds a variable that may be a collider, opening the possibility of artificial associations or additional biases. In this case, we observe how conditioning on a collider has dramatically altered our conclusions, suggesting a strong harmful effect of coffee consumption that wasn't evident in the properly specified model. This example clearly demonstrates how the "Table 2 fallacy" can lead to erroneous conclusions that might inappropriately influence clinical practice or public health recommendations.

DISCUSSION

The Table 2 fallacy remains a frequent challenge in observational studies aimed at establishing causal relationships. As demonstrated throughout this review, the root problem lies in the inadequate differentiation between confounding, mediating, and colliding variables, leading many researchers to adjust "for everything" and thereby distort or misrepresent the associations under investigation.

Our examples illustrate two fundamentally different analytical approaches. The DAG-informed approach carefully selects adjustment variables based on their causal role, while the "associated factors" model indiscriminately includes multiple variables without considering their position in the causal chain. Although "associated factors" analyses may provide value in exploratory research or hypothesis generation, they carry substantial methodological risks when used for causal inference.

Proper causal inference requires researchers to establish the relationship between exposure and outcome before analysis. DAGs, grounded in biological plausibility, existing literature, and epidemiological theory, clarify whether variables act as confounders (affecting both exposure and outcome), mediators (lying on the causal pathway), or colliders (influenced by multiple causes). This distinction is crucial: adjusting for confounders is necessary, but adjusting for mediators attenuates true effects, while adjusting for colliders creates spurious associations [17]. Without

this framework, multivariate analyses that adjust "for everything" risk nullifying real effects or generating artificial correlations [3-5].

The "associated factors" approach suffers from several fundamental flaws. First, including mediators or colliders produces biased estimates that fail to capture the true effect of the primary exposure. Second, selecting variables based solely on statistical significance promotes model instability and increases the risk of chance findings. Third, this atheoretical approach often ignores biological plausibility, potentially overlooking important causal mechanisms while emphasizing spurious relationships [15, 16].

Perhaps most critically, interpreting all coefficients from indiscriminate models as "independent effects" fundamentally misunderstands what such models estimate. A single variable may simultaneously act as a confounder for one relationship while serving as a mediator or collider for another. Without a clear causal framework, these coefficients lack coherent interpretation and should not guide clinical or public health decisions. The seemingly comprehensive Table **2** becomes a statistical amalgamation rather than a meaningful analysis of causal relationships.

For explanatory studies, the methodology is clear: construct a comprehensive causal framework before analysis to identify true confounders while excluding mediators and colliders. This approach yields the most accurate estimation of causal effects. Even in exploratory research, having a preliminary causal map helps avoid basic methodological errors. In predictive modeling, understanding each variable's causal role helps determine whether its inclusion improves or compromises prediction accuracy [3-5].

Therefore, the choice between DAG-informed and "associated factors" approaches is not merely methodological preference but determines the validity of causal conclusions. Proper causal reasoning maximizes internal validity and produces clinically meaningful findings, while indiscriminate adjustment undermines both scientific rigor and practical utility. Understanding these distinctions is essential for researchers, reviewers, and consumers of scientific literature to properly evaluate and apply research findings.

LIMITATIONS AND STRENGTHS

This methodological review uses simulated examples rather than empirical data, reflecting a practical constraint in causal research: most observational datasets lack the complete variable sets required for proper DAG implementation. Databases are typically designed for descriptive or administrative purposes, not for testing specific causal hypotheses, often omitting crucial mediators or unmeasured confounders. This limitation paradoxically reinforces our argument—researchers frequently adjust for "available" rather than "appropriate" variables, perpetuating the Table 2 fallacy. Additionally, while our examples necessarily simplify complex causal relationships for pedagogical clarity, this simplification reflects proper methodology: the research question and causal framework must be defined a priori, not determined by available data.

Our approach has several strengths. First, by using controlled examples, we can demonstrate the exact magnitude of bias introduced by inappropriate adjustment, something impossible with real data where true causal effects are unknown. Second, our framework emphasizes that the "associated factors" approach is not merely suboptimal but methodologically incorrect for causal inference—each variable's role must be theoretically justified before model inclusion. The decision tree (Figure **4**) operationalizes these concepts, requiring explicit causal assumptions rather than hiding them behind automated statistical procedures.

Finally, while uncertainty about causal relationships exists in practice, this should be addressed through sensitivity analyses with alternative DAG specifications, not by abandoning causal reasoning altogether. The common practice of presenting all adjusted coefficients as independent effects when models include mediators or colliders violates fundamental principles of causal inference. Our contribution lies in clarifying these principles and providing practical tools for their implementation, addressing a persistent methodological problem that undermines the validity of many published findings.

CONCLUSIONS

In conclusion, the practical utility of causal reflection lies not only in methodological soundness but translates into more informed clinical or public health decisions. By avoiding confusing mediators with confounders and not introducing colliders into adjustment models, it is ensured that the reported associations reflect, with the greatest possible fidelity, the underlying etiological relationships, which results in real advances in knowledge and health action.

PRACTICAL RESOURCES AND TOOLS

To facilitate the implementation of the methodological approaches discussed in this manuscript, we direct readers to established resources for constructing and analyzing DAGs.

- DAGitty (www.dagitty.net): Web-based interface requiring no programming knowledge, with automatic generation of adjustment sets and R code export capabilities [19].
- ggdag R package: For creating publication-quality DAGs within R's ggplot2 framework [20]
- dagitty R package: For analyzing causal diagrams and identifying adjustment sets programmatically [19].

These tools can reproduce the examples presented in this manuscript and assist researchers in developing their own DAGs. Rather than providing code that may become outdated, we encourage readers to consult the comprehensive documentation and tutorials available for each tool.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Since this manuscript is a secondary data base study it was not required.

Informed consent: Since this is a secondary data analysis, informed consent was not required.

Clinical trial number: not applicable.

ACKNOWLEDGEMENTS

A special thanks to the members of Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM), Amazonas, Peru, for their support and contributions throughout the completion of this research.

FINANCIAL DISCLOSURE

This study was financed by Vicerectorado de Investigación de la Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY

Data are available upon request to the corresponding author.

AUTHORS' CONTRIBUTION

Víctor Juan Vera-Ponce: Conceptualization, Investigation, Methodology, Resources, Writing - Original Draft, Writing - Review & Editing

Fiorella E. Zuzunaga-Montoya: Methodology, Software, Data Curation, Formal analysis, Writing - Review & Editing

Jhosmer Ballena-Caicedo: Investigation, Project administration, Writing - Original Draft, Writing - Review & Editing

Lupita Ana Maria Valladolid-Sandoval: Investigation, Project administration, Writing - Original Draft, Writing - Review & Editing

Carmen Inés Gutierrez De Carrillo: Methodology, Resources, Supervision, Funding acquisition, Writing - Review & Editing

REFERENCES

[1] Dharma C, Fu R, Chaiton M. Table 2 Fallacy in Descriptive Epidemiology: Bringing Machine Learning to the Table. Int J Environ Res Public Health 2023; 20(13): 6194. https://doi.org/10.3390/ijerph20136194

[2] Schisterman EF, Cole SR, Platt RW. Overadjustment Bias and Unnecessary Adjustment in Epidemiologic Studies. Epidemiol Camb Mass 2009; 20(4): 488-95. https://doi.org/10.1097/EDE.0b013e3181a819a1

[3] Kaufman JS. Causal Inference Challenges in the Relationship Between Social Determinants and Cardiovascular Outcomes. Can J Cardiol 2024; 40(6): 976-88. https://doi.org/10.1016/j.cjca.2024.02.005

[4] Akinkugbe AA, Simon AM, Brody ER. A scoping review of Table 2 fallacy in the oral health literature. Community Dent Oral Epidemiol 2021; 49(2): 103-9. https://doi.org/10.1111/cdoe.12617

[5] Bandoli G, Palmsten K, Chambers CD, Jelliffe-Pawlowski LL, Baer RJ, Thompson CA. Revisiting the Table 2 fallacy: A motivating example examining preeclampsia and preterm birth. Paediatr Perinat Epidemiol 2018; 32(4): 390-7. https://doi.org/10.1111/ppe.12474

[6] Hernán MA, Hsu J, Healy B. A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. CHANCE 2019; 32(1): 42-9. https://doi.org/10.1080/09332480.2019.1579578

[7] Lederer DJ, Bell SC, Branson RD, Chalmers JD, Marshall R, Maslove DM, *et al*. Control of Confounding and Reporting of Results in Causal Inference Studies. Guidance for Authors from Editors of Respiratory, Sleep, and Critical Care Journals. Ann Am Thorac Soc 2019; 16(1): 22-8. https://doi.org/10.1513/AnnalsATS.201808-564PS

[8] Alba AC, Agoritsas T, Walsh M, Hanna S, Iorio A, Devereaux PJ, *et al*. Discrimination and Calibration of Clinical Prediction Models: Users' Guides to the Medical Literature. JAMA 2017; 318(14): 1377-84. https://doi.org/10.1001/jama.2017.12126

[9] Conroy S, Murray EJ. Let the question determine the methods: descriptive epidemiology done right. Br J Cancer 2020; 123(9): 1351-2. https://doi.org/10.1038/s41416-020-1019-z

[10] Fox MP, Murray EJ, Lesko CR, Sealy-Jefferson S. On the Need to Revitalize Descriptive Epidemiology. Am J Epidemiol 2022; 191(7): 1174-9. https://doi.org/10.1093/aje/kwac056

[11] Hernan MA, Robins JM. Causal Inference: What If. CRC Press; 2024. 312 p.

[12] Huitfeldt A. Is caviar a risk factor for being a millionaire? BMJ 2016; 355: i6536. https://doi.org/10.1136/bmj.i6536

[13] Vera-Ponce VJ, Zuzunaga-Montoya FE, Vásquez-Romer LEM, Sanchez-Tamay NM, Loayza-Castro JA, Gutierrez De Carrillo CI. Scientific Tasks in Biomedical and Oncological Research: Describing, Predicting, and Explaining. J Cancer Res Updat 2024; 13: 52-65. https://doi.org/10.30683/1929-2279.2024.13.08

[14] Werlinger F, Cáceres DD, Werlinger F, Cáceres DD. Aplicación de grafos acíclicos dirigidos en la evaluación de un set mínimo de ajuste de confusores: un complemento al modelamiento estadístico en estudios epidemiológicos observacionales. Rev Médica Chile 2018; 146(7): 907-13. https://doi.org/10.4067/s0034-98872018000700907

[15] Tennant PWG, Murray EJ, Arnold KF, Berrie L, Fox MP, Gadd SC, *et al*. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. Int J Epidemiol 2021; 50(2): 620-32. https://doi.org/10.1093/ije/dyaa213

[16] Ferguson KD, McCann M, Katikireddi SV, Thomson H, Green MJ, Smith DJ, *et al*. Evidence synthesis for constructing directed acyclic graphs (ESC-DAGs): a novel and systematic method for building directed acyclic graphs. Int J Epidemiol 2020; 49(1): 322-9. https://doi.org/10.1093/ije/dyz150

[17] Holmberg MJ, Andersen LW. Collider Bias. JAMA 2022; 327(13): 1282-3. https://doi.org/10.1001/jama.2022.1820

[18] Suttorp MM, Siegerink B, Jager KJ, Zoccali C, Dekker FW. Graphical presentation of confounding in directed acyclic graphs. Nephrol Dial Transplant Off Publ Eur Dial Transpl Assoc - Eur Ren Assoc 2015; 30(9): 1418-23. https://doi.org/10.1093/ndt/gfu325

[19] Textor J, van der Zander B, Gilthorpe MS, Liskiewicz M, Ellison GT. Robust causal inference using directed acyclic graphs: the R package "dagitty". Int J Epidemiol 2016; 45(6): 1887-94. https://doi.org/10.1093/ije/dyw341

[20] Barrett M. ggdag: An R Package for visualizing and analyzing causal directed acyclic graphs. R package version 0.2.7 2022. Available from: https://CRAN.R-project.org/package=ggdag

Received on 20-08-2025 Accepted on 21-09-2025 Published on 23-10-2025

https://doi.org/10.6000/1929-6029.2025.14.61