PharmaSUG 2026 - Paper RW-253

Embracing Novel Approaches to Automated Causal Inference Framework

Laura Watson Smith and Sherrine Eid, SAS Institute

ABSTRACT

Our industry perhaps has the deepest understanding of the statement, correlation does not imply causation. Causal claims are directly actionable for policy implementation in a way that traditional association-based claims might not be. This leaves us with a rather glaring and consequential question: how can we infer causation? As the evolution of technology and methods grows, we find ourselves at an exciting time for innovation and potential to exponentially impact causality in our industry.

This paper outlines an applied framework for performing causal analysis on observational data in SAS Viya, leveraging automation with the trusted logic in SAS and novel open-source packages. Researchers can now robustly analyze causal inference in real world data, as well as visually analyze the insights generated by the best-in-class analytics on an end-to-end platform. Moreover, a low-code-no-code and programming interface allows easier collaboration for research study teams to access advanced analytics like machine learning models to generate evidence for a regulatory dialogue. This solution allows epidemiologists and researchers to assess real time what-if scenarios and comparisons to generate intelligence more quickly.

INTRODUCTION

PUBLISHED EXAMPLE

Consider the question, “Does smoking cessation cause weight change?” Framed as an associational question, this becomes, “Are smoking cessation and weight change associated?” In contrast, a causal inference approach asks, “Would the same weight change occur in the absence of smoking cessation as was observed when individuals quit smoking?” Smoking is a major public health concern, contributing to many chronic diseases and long-term changes in body weight.[1-4] Being overweight or obese is linked to increased chronic health risks and higher all-cause mortality.[5-7] Understanding how smoking cessation affects weight change is therefore important for public health research and for informing clinical guidelines. This study uses observational, cross-sectional data from the US National Health and Nutrition Examination Survey (NHANES) to estimate the causal effect of smoking cessation on weight change while adjusting for confounding factors, following the framework outlined in the textbook What If by Hernán and Robins.

ORIENTATION FOR CAUSAL INFERENCE

Causality describes a relationship wherein one event, the cause, directly influences another, the outcome, such that the outcome would not have occurred without the cause.[8] Causal inference is the process of determining whether such a relationship exists between variables in the general population and quantifying its strength.[8]

WHAT IS THE DIFFERENCE BETWEEN CAUSAL INFERENCE AND STATISTICAL ANLYSIS

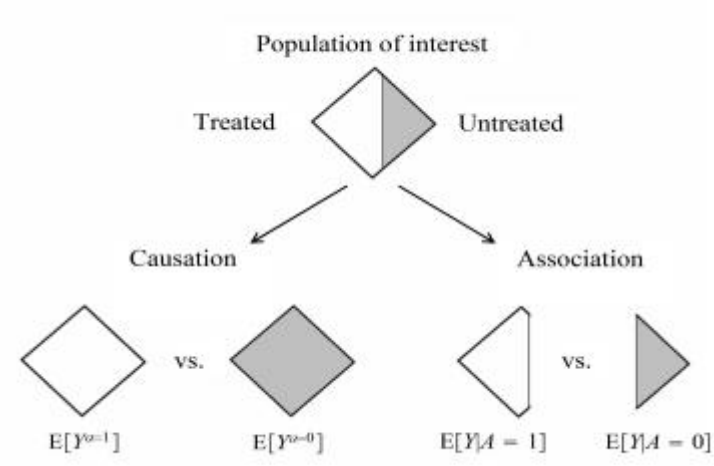
QUESTIONS

Statistical analysis and causal inference address different questions. Statistical analysis focuses on observed associations, examining the joint distribution of variables to answer associational questions.[9,10] In contrast, causal inference tackles counterfactual “what-if” scenarios, asking how outcomes would change under alternative conditions or interventions.[9] Because the same data distribution can align with multiple causal structures, answering causal questions requires knowledge of the underlying data-generating process, formalized through constructing a causal model and asserting causal assumptions.9 The key distinction is that while statistical analysis describes observed associations, causal inference explains what would have occurred under different, hypothetical circumstances.

POPULATIONS

Figure 1 taken from What If by Hernán and Robins, helps to illustrate the difference in considered populations between causal inference and statistical analysis in a general cross-sectional example.[11] Consider the question: "Is the risk of outcome the same for treated and untreated individuals?", where Y denotes outcome and A defines treatment assignment.

Statistical analysis and causal inference differ also in the populations they conceptually compare. Statistical analysis divides the observed sample into treated and untreated groups, then compares average outcomes between these real individuals—Person X versus Person Z—using the data we have.[11,12] In contrast, causal inference imagines the same population under two hypothetical scenarios: everyone treated and everyone untreated.[11] Conceptually, it compares Person X under treatment to that same Person X had they not been treated, which requires defining potential outcomes for each individual for each exposure condition.[11] Because only one outcome is observed for each person, one potential outcome is counterfactual and unobservable, making causal inference fundamentally a problem of missing data. While statistical analysis could, in principle, observe more data through additional sampling, causal inference uses observed data and assumptions to learn about outcomes that can never be directly observed.[12]



X (Observed) X (Hypothetical) X (Observed) Z (Observed)

Figure 1. Illustration of the Populations Considered in Causal vs. Associational Statistics

Figure 1 was captured from What If by Hernán and Robins, see reference 11.

ASSUMPTIONS

The final key distinction lies in the assumptions each approach makes. Associational assumptions are testable in principle, given a sufficiently large sample size.[9,13] These include assumptions about distributions or model structures, such as normality, linearity, and independence. Causal inference, on the other hand, requires additional assumptions beyond those of standard statistical analysis. These causal assumptions allow an observational study to be conceptualized as a conditionally randomized experiment.[11]

WHY DO WE DO CAUSAL INFERENCE INSTEAD OF STATISTICAL ANALYSIS

We use causal inference when we want to understand the effect of an intervention, not just whether variables are related. Statistical analysis only identifies associations, while causal inference asks what would happen if conditions changed—making it more actionable for research, policy implementation, and medicine than statistical analysis.[14,15] For example, "Smoking cessation is associated with weight change" describes correlation, but a causal claim like "Smoking cessation causes a 3-pound weight increase compared to not quitting" explains the impact of intervention and supports decision-making.

WHY IS CAUSAL INFERENCE DIFFICULT?

UNOBSERVABLE DATA

Causal inference is often described as a problem of missing data. As previously noted, it conceptually compares the same population under two hypothetical exposure conditions. This framework yields two potential outcomes for each individual: one observed, and the other—the counterfactual—unobservable. This is the fundamental problem of causal inference: we can never observe both potential outcomes for the same individual at the same time.[16]

ASSOCIATION ≠ CAUSATION

Mentioned earlier, traditional statistical analysis focuses on associational questions, whereas causal inference addresses counterfactual questions. This distinction leads directly to a familiar problem: association is not the same as causation. The presence of an association alone is insufficient to establish a causal relationship. Association and causation are defined with respect to different underlying populations, and an observed association may reflect phenomena other than a causal relationship, such as bias or confounding.[16-21]

TO IMPLY CAUSATION

Establishing causality requires conditions beyond an observed association between two variables.[17,20,21] Satisfying these conditions amounts to making appropriate assumptions, including identifying assumptions that imply the causal effect is measurable from observed data. This process of identification will be elaborated on in the next section.

COMMON CAUSAL TASKS

DIRECTED ACYCLIC GRAPHS (DAGS)

Directed acyclic graphs (DAGs) visualize assumed relationships among variables included in the system as nodes linked through one-way directed arrows which form no loops or cycles.[11,22-31] They clarify causal pathways in complex studies, informing study design by making causal assumptions explicit.[11,22-31] By explicitly showing pathways, DAGs help identify confounders, mediators, and colliders, guiding the selection of covariates to control confounding and improve causal estimates.[11,22-31] DAGs also provide a means of visually assessing potential biases in causal inference.[24-26,30] Using a DAG helps to bridge gaps in bias assessment by identifying biases from all sources, including data selection, study design, and analysis method.[32]

IDENTIFICATION

Identification determines whether the causal estimand—such as the average treatment effect (ATE) in this case—can be expressed using observed data.[11,33] When the identifying assumptions underlying causal analysis hold, it implies that the observed effect is the only possible effect that could have arisen given the data; if identification fails, the analysis remains associational.[11,34] This process is complicated by the fundamental problem of causal inference: we cannot observe both potential outcomes for the same individual at the same time.[34] As a result, causal inference operates at the population level, comparing averages rather than individual outcomes.[11] In causal identification, the goal is to prove that it is possible to express the causal estimand as a function of the observed data, a task complicated by the very nature of this fundamental problem. To justify that the causal estimand is identifiable from our data, certain causal identification assumptions must be asserted.[34]

- Positivity: There is a nonzero probability that individuals in the observed data receive every level of exposure for each combination of values for included covariates.[35-38]
- Consistency: The exposure is defined well enough that on the individual level, there are no multiple versions of the exposure leading to different potential outcomes.[39-41]
- No Interference: An individual's treatment status has no impact on the potential outcomes for any other individual.[37,41]
- Exchangeability: For all levels of exposure, the potential outcomes are independent of the exposure assignment conditional on the observed covariates.[37,42]

Taken together, these assumptions allow us to assert that the exposed and unexposed groups are identical in all respects, whether observable or unobservable, except for their treatment assignments.[37,41-43] This allows us to assert that, under these assumptions, our observed data identifies the causal estimand, which may be interpreted as reflecting a true causal relationship.[11,33,34]

EFFECT ESTIMATION

Effect estimation quantifies the impact of an exposure on an outcome across a population. After consulting a casual diagram, and once the causal parameter is identified, statistical methods are used to estimate its magnitude, providing insight into how changes in treatment would influence outcomes at the population level.

CAUSAL INFERENCE ACROSS PROGRAMMING LANGUAGES

Today, researchers and practitioners employ a range of options for conducting causal inference analyses, including using different general programming language platforms, which each offer different strengths or specific functional capabilities within the causal inference framework.

We found that SAS excels through its advanced computational capabilities combined with user-friendly interfaces that make it accessible to a wide range of users. It inherently has automated analytics frameworks, machine learning pipelines, and other programming languages built into the platform, enabling the development of workflows tailored to specific research questions. Additionally, SAS has strong data handling capabilities and built-in procedures for causal inference.

Python is an open-source language with a large, active user community and extensive documentation. It provides specialized libraries for causal inference, along with flexible visualization and customization options, and enables machine learning packages and broader automated analytics frameworks. However, compared to SAS, Python is a language with emerging packages being explored by some regulators. The following section provides an overview of the Python-based causal inference packages evaluated in comparison to the corresponding SAS procedures.

PYTHON PACKAGE – CAUSAL LEARN

Causal-learn is a Python library for causal discovery in observational data, providing a range of classical and modern algorithms to learn causal structure directly from data. It supports constraint-based methods (e.g., PC), score-based methods (e.g., GES), and functional causal model approaches such as LiNGAM, with LiNGAM uniquely producing a fully oriented DAG suitable for downstream identification and effect estimation. Originally developed in 2021 by researchers affiliated with Carnegie Mellon University, Causal-learn is actively maintained and continues to evolve with regular releases.[44]

PYTHON PACKAGE – DOWHY

DoWhy is an end-to-end Python library for causal inference that integrates causal identification and effect estimation within a single, unified framework. It supports standard identification strategies such as backdoor, frontdoor, general adjustment, and instrumental variables, and offers a broad range of estimation methods, including regression, propensity score techniques, and doubly robust estimators. Developed at Microsoft Research in 2019 and actively maintained, DoWhy is widely adopted for applied causal analysis due to its flexibility, strong theoretical grounding, and extensive documentation.[45,46]

EXAMPLE

This example demonstrates effect estimation using SAS Viya to assess whether quitting smoking causes weight change, while appropriately adjusting for confounders. It is implemented as an interactive dashboard, leveraging native SAS procedure code as well as Python packages through embedded SAS Jobs. At each step in the dashboard, embedded SAS Jobs allow users to specify model inputs - such as exposure and outcome variables - providing an easy, low code way to control the underlying code being executed in a reproducible manner.

CONSTRUCT A CAUSAL DIAGRAM

Causal analyses begin by constructing a DAG to visualize the relationships among variables, particularly how they relate to the exposure and outcome. This helps identify confounders that must be controlled to obtain unbiased effect estimates. A DAG can be created manually or generated algorithmically using either SAS's PROC CAUSALDISCOVERY or Python's CAUSALLEARN (**Figure 2**; **Figure 4**). These methods each produce an Estimated Adjacency Matrix that describes the learned structure of the DAG through parent-child relationships among variables (**Figure 3**; **Figure 5**).

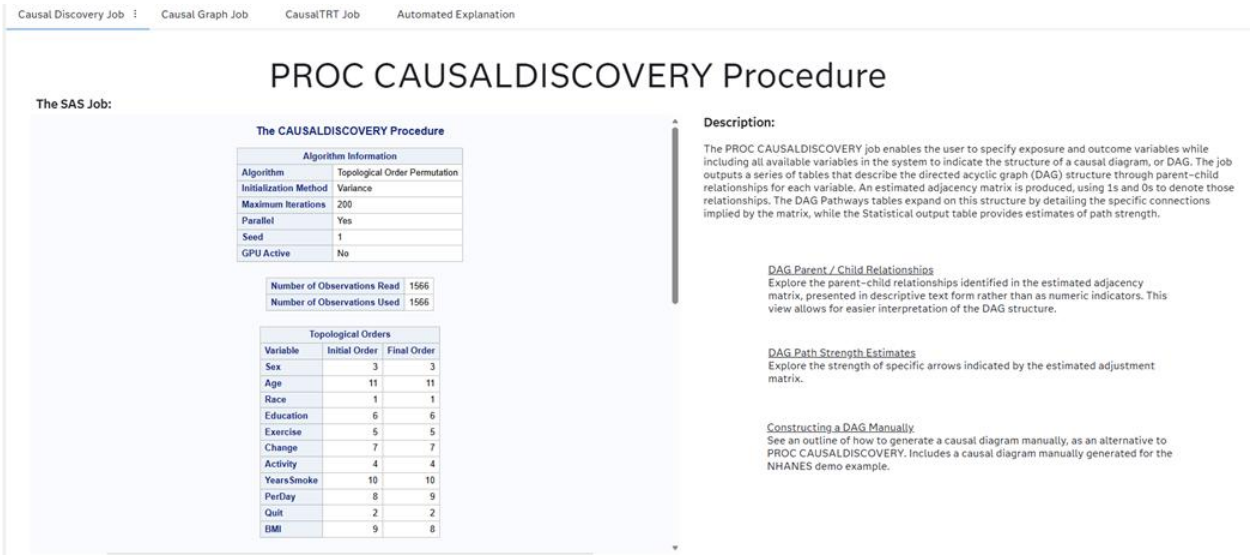


Figure 2. Effect Estimation Dashboard: Proc CausalDiscovery Page

Estimated Adjacency Matrices

Alpha	Variable	Sex	Age	Race	Education	Exercise	Change	Activity	YearsSmoke	PerDay	Quit	BMI
1E-6	Sex	0	1	0	1	1	0	1	0	1	0	1
	Age	0	0	0	0	0	0	0	0	0	0	0
	Race	1	0	0	0	1	0	0	0	1	1	1
	Education	0	1	0	0	0	1	0	1	1	0	1
	Exercise	0	0	0	1	0	0	0	1	0	0	1
	Change	0	0	0	0	0	0	0	0	0	0	0
	Activity	0	0	0	1	1	0	0	0	0	0	1
	YearsSmoke	0	1	0	0	0	0	0	0	0	0	0
	PerDay	0	0	0	0	0	0	0	1	0	0	0
	Quit	1	1	0	1	1	1	1	0	1	0	1
	BMI	0	1	0	0	0	0	0	1	1	0	0

Figure 3. PROC CAUSALDISCOVERY Output: Estimated Adjacency Matrix

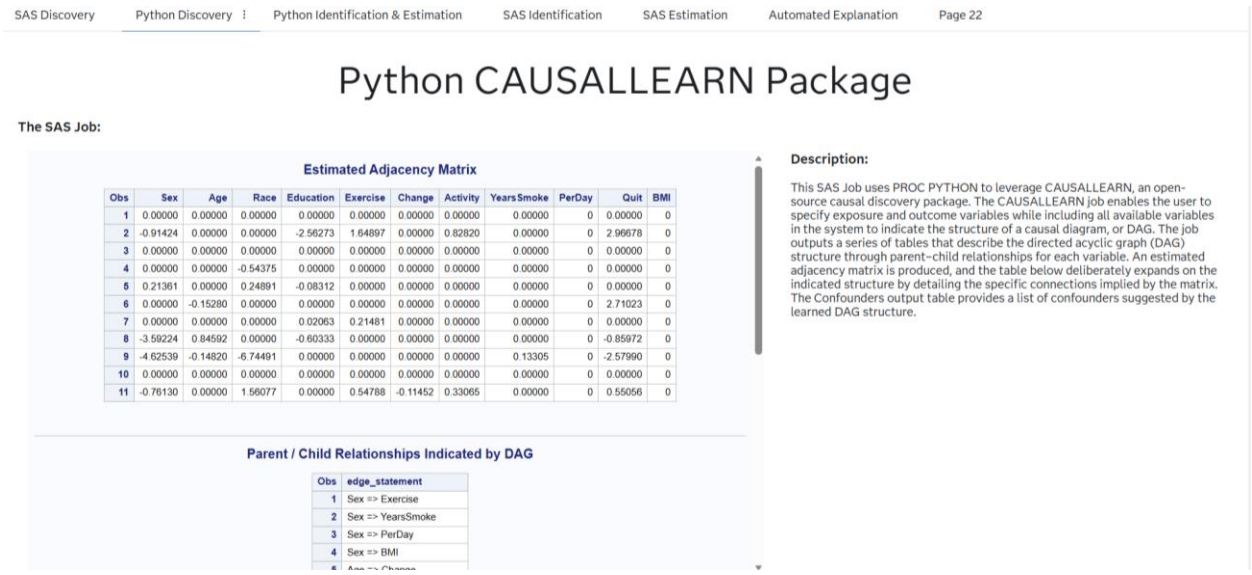


Figure 4. Effect Estimation Dashboard: CausalLearn Page

Estimated Adjacency Matrix

Obs	Sex	Age	Race	Education	Exercise	Change	Activity	YearsSmoke	PerDay	Quit	BMI
1	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0	0.00000	0
2	-0.91424	0.00000	0.00000	-2.56273	1.64897	0.00000	0.82820	0.00000	0	2.96678	0
3	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0	0.00000	0
4	0.00000	0.00000	-0.54375	0.00000	0.00000	0.00000	0.00000	0.00000	0	0.00000	0
5	0.21361	0.00000	0.24891	-0.08312	0.00000	0.00000	0.00000	0.00000	0	0.00000	0
6	0.00000	-0.15280	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0	2.71023	0
7	0.00000	0.00000	0.00000	0.02063	0.21481	0.00000	0.00000	0.00000	0	0.00000	0
8	-3.59224	0.84592	0.00000	-0.60333	0.00000	0.00000	0.00000	0.00000	0	-0.85972	0
9	-4.62539	-0.14820	-6.74491	0.00000	0.00000	0.00000	0.00000	0.13305	0	-2.57990	0
10	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0	0.00000	0
11	-0.76130	0.00000	1.56077	0.00000	0.54788	-0.11452	0.33065	0.00000	0	0.55056	0

Figure 5. CausalLearn Output: Estimated Adjacency Matrix

These causal diagram construction methods in the Effect Estimation dashboard help identify the structure of a causal diagram but do not generate a visual DAG, so this analysis is supplemented with a manual construction method. A detailed description of this manual approach is provided in the linked GitHub page.

The completed causal diagram (**Figure 6**) summarizes the relationships among variables in the data and can reveal structures not evident from statistical analysis alone. In the smoking cessation example, the DAG confirmed Age, Sex, and Years Smoked as confounders and additionally identified Exercise as a confounder because it influences both quitting status and BMI, which affects weight change. A subsequent statistical review supported this finding, illustrating how DAGs help uncover complex dependencies that may be missed by traditional analyses alone.

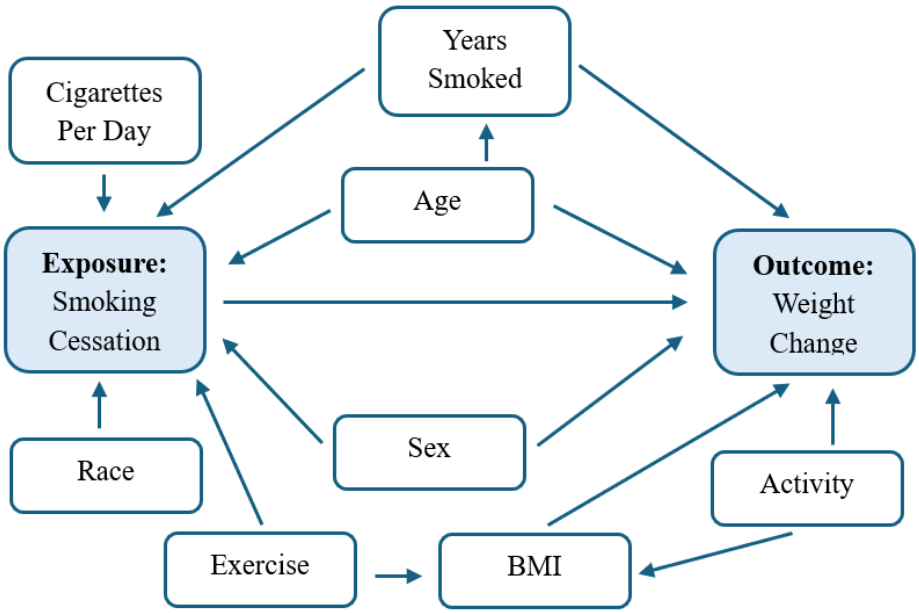


Figure 6. Manually Constructed Causal Diagram of NHANES Example

CAUSAL IDENTIFICATION

The next step, Identification, verifies whether the average treatment effect can be estimated from the observed data, moving the analysis beyond traditional associational methods into causal inference. SAS performs causal identification and estimation as separate steps using dedicated procedures, whereas the Python DoWhy package carries out identification and estimation jointly within one integrated process.

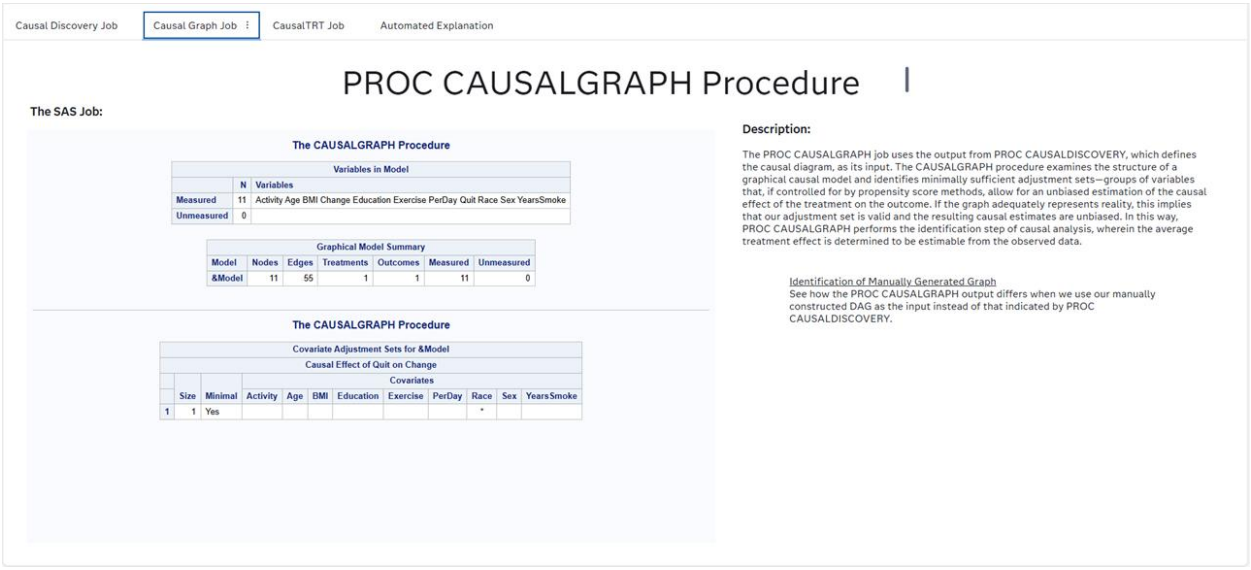


Figure 7. Effect Estimation Dashboard: Causal Graph Page

The SAS approach to causal identification is PROC CAUSALGRAPH (**Figure 7**). The procedure takes the defined causal diagram from either the manual construction process, PROC CAUSALDISCOVERY, or CAUSALLEARN, then evaluates the graph—given the user specified exposure and outcome—to identify

covariate adjustment sets for unbiased effect estimation.[47] The key output is the covariate adjustment set table (**Figure 8**); finding at least one adjustment set indicates that the average treatment effect is identifiable when those variables are controlled for as confounders.

Covariate Adjustment Sets for TESTID Demo

Causal Effect of SmokingCessation on WeightChange

	Size	Minimal	Covariates							
			Activity	Age	BMI	Exercise	PerDay	Race	Sex	YearsSmoke
1	4	Yes		*	*				*	*
2	4	Yes		*		*			*	*
3	5	No	*	*	*				*	*
4	5	No	*	*		*			*	*
5	5	No		*	*	*			*	*
6	5	No		*	*		*		*	*
7	5	No		*	*			*	*	*
8	5	No		*		*	*		*	*
9	5	No		*		*		*	*	*

Figure 8. PROC CAUSALGRAPH Output: Covariate Adjustment Sets

From PROC CAUSALGRAPH two minimally sufficient adjustment sets are reported for the manually generated diagram. The second set aligns exactly with the manually selected confounders—Age, Exercise, Sex, and Years Smoked—providing additional support for that confounder selection. This establishes identification, confirming that the average treatment effect is estimable from our data.

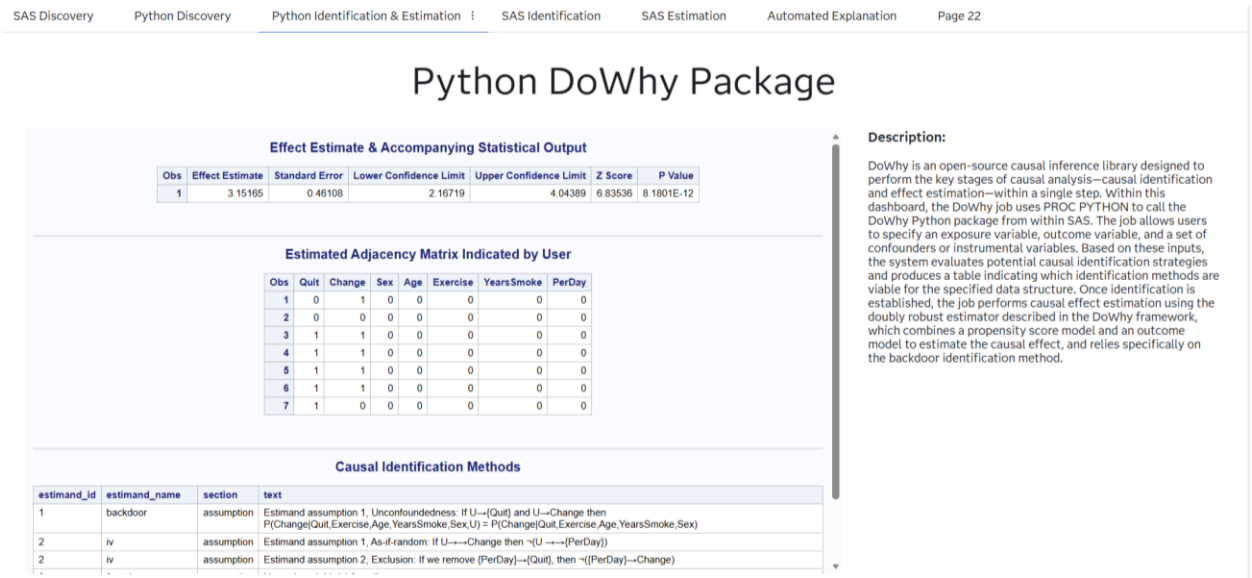


Figure 9. Effect Estimation Dashboard: DoWhy Page

Causal Identification Methods

estimand_id	estimand_name	section	text
1	backdoor	assumption	Estimand assumption 1, Unconfoundedness: If U→{Quit} and U→Change then P(Change|Quit,Exercise,Age,YearsSmoke,Sex,U) = P(Change|Quit,Exercise,Age,YearsSmoke,Sex)
2	iv	assumption	Estimand assumption 1, As-if-random: If U→→Change then ¬(U →→{PerDay})
2	iv	assumption	Estimand assumption 2, Exclusion: If we remove {PerDay}→{Quit}, then ¬({PerDay}→Change)
3	frontdoor	assumption	No such variable(s) found!
4	general_adjustment	assumption	Estimand assumption 1, Unconfoundedness: If U→{Quit} and U→Change then P(Change|Quit,Age,Exercise,YearsSmoke,Sex,U) = P(Change|Quit,Age,Exercise,YearsSmoke,Sex)

Figure 10. DoWhy Output: Identification Methods

The Python approach to causal identification is integrated directly into the DoWhy framework with causal estimation. DoWhy incorporates user specified confounders and instrumental variables, defined as variables that point only to the exposure in the causal diagram, which are identified for users through SAS's PROC CAUSALDISCOVERY or Python's CAUSALLEARN (**Figure 9**). The framework then evaluates several identification strategies and summarizes in the Identification Method Table which approaches can or cannot identify the causal estimand (**Figure 10**). When the required assumptions for a given method appear in the table, that method is considered sufficient to establish identification, provided those assumptions are met—typically by adjusting for the listed confounders, which is done automatically by DoWhy in effect estimation.

For our example DoWhy identified several methods—backdoor, instrumental variables (IV), and general adjustment—as sufficient to establish identification for our analysis, while the frontdoor criterion was determined to be insufficient. In principle, any of the sufficient identification strategies could be paired with an appropriate estimator, such as a doubly robust or general regression estimator, for effect estimation within the DoWhy framework. In our implementation, we rely on the doubly robust estimator in combination with the backdoor identification method. As such, confirmation that the backdoor criterion is sufficient to identify the causal estimand ensures that our analysis can proceed.

EFFECT ESTIMATION

In the final step, the treatment effect is estimated as the difference in predicted outcomes across the population under each treatment level. This can be done with SAS using the PROC CAUSALTRT procedure, which integrates a propensity score model and an outcome regression model into a pair of estimating equations to compute the average treatment effect (**Figure 11**). PROC CAUSALTRT runs through all supported causal inference methods, producing predicted outcomes by treatment level, average treatment effect estimates, propensity scores, and related statistical results. The main output is the Analysis of Causal Effects table (default method: AIPW), which reports all these statistical outputs (**Figure 12**). A summary table in the Effect Estimation dashboard also compiles average treatment effect estimates and confidence metrics across all methods (**Figure 13**).

In the smoking cessation example, Figure 8 and Figure 9 show an AIPW estimated average treatment effect of 3.37 kg, indicating that quitting smoking leads to roughly a 3.4 kg long term weight gain after adjusting for confounders.

Effect estimation is performed automatically when using the Python DoWhy package (**Figure 9**). Our implementation relies on a doubly robust estimator, which internally and automatically specifies both a propensity score model and an outcome model for the input data. It uses the backdoor identification method to adjust for the user provided confounders. DoWhy returns a table that reports the estimated causal effect along with associated confidence measures (**Figure 14**).

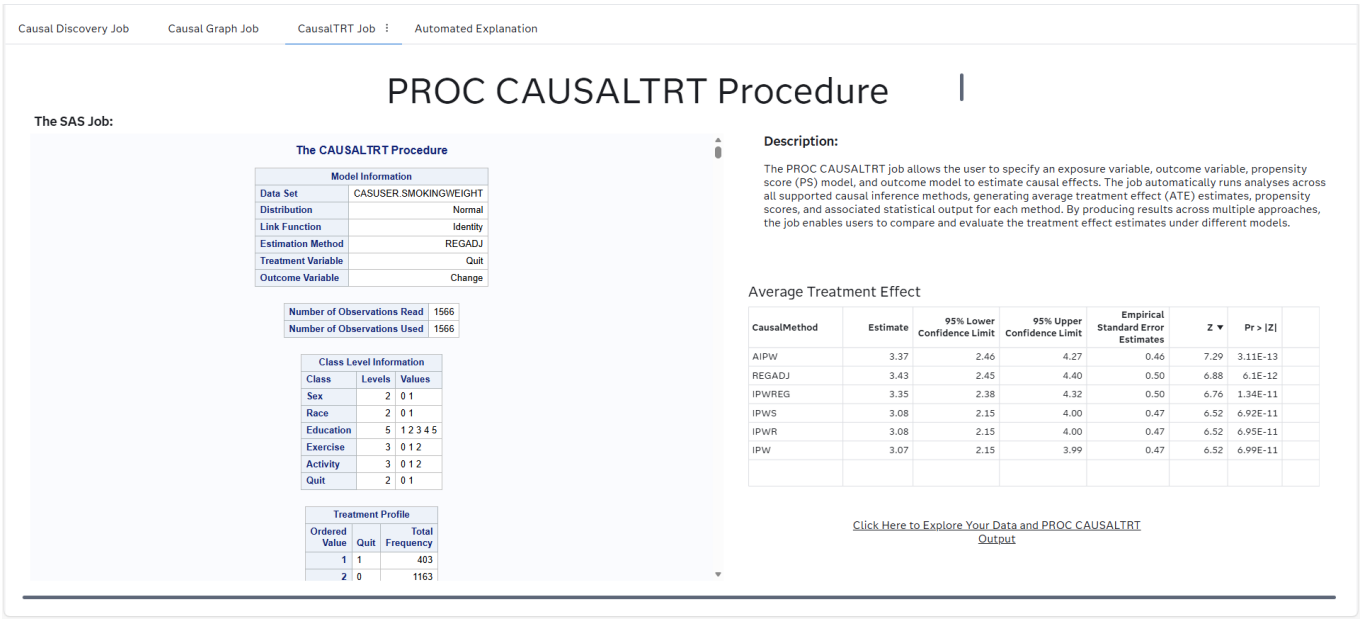


Figure 11. Effect Estimation Dashboard: CausalTrt Page

The CAUSALTRT Procedure

Analysis of Causal Effect

Parameter	Treatment Level	Estimate	Robust Std Err	Wald 95% Confidence Limits		Z	Pr > |Z|
POM	1	5.1467	0.4177	4.3282	5.9653	12.32	<.0001
POM	0	1.7782	0.2152	1.3564	2.2001	8.26	<.0001
ATE		3.3685	0.4621	2.4628	4.2742	7.29	<.0001

Figure 12. PROC CAUSALTRT Output: Analysis of Causal Effects AIPW

CausalMethod	Estimate	95% Lower Confidence Limit	95% Upper Confidence Limit	Empirical Standard Error Estimates	Z ▼	Pr > |Z|
AIPW	3.37	2.46	4.27	0.46	7.29	3.11E-13
REGADJ	3.43	2.45	4.40	0.50	6.88	6.1E-12
IPWREG	3.35	2.38	4.32	0.50	6.76	1.34E-11
IPWS	3.08	2.15	4.00	0.47	6.52	6.92E-11
IPWR	3.08	2.15	4.00	0.47	6.52	6.95E-11
IPW	3.07	2.15	3.99	0.47	6.52	6.99E-11

Figure 13. PROC CAUSALTRT Output: Analysis of Causal Effects Summary

Effect Estimate & Accompanying Statistical Output

Obs	Effect Estimate	Standard Error	Lower Confidence Limit	Upper Confidence Limit	Z Score	P Value
1	3.15165	0.46108	2.16719	4.04389	6.83536	8.1801E-12

Figure 14. DoWhy Output: Effect Estimate Measures

For the NHANES example, DoWhy reported a causal effect estimate of 3.15, indicating that smoking cessation leads to an approximate 3.15 kg increase in weight compared with not quitting after adjusting for confounders. This estimate aligns with the range of results included in SAS's PROC CAUSALTRT Analysis of Causal Effects Summary table. Notably, both the SAS and Python analyses provide consistent evidence that smoking cessation results in a meaningful increase in weight. With these findings, the causal effect estimation analysis is complete, though additional visualizations could still be explored.

SUMMARY OF FINDINGS

ARE THE METHODS DIFFERENT?

SAS and Python support the core causal inference tasks—discovery, identification, and estimation—but differ in their methodology. For causal discovery, Python's Causal learn offers a set of structure learning algorithms, including PC, GES, and LiNGAM, with LiNGAM uniquely producing a fully oriented DAG. In contrast, PROC CAUSALDISCOVERY in SAS focuses on two algorithms (TOP and MCV) designed for reliability and interpretability in regulated environments, with TOP as the default.

Differences also arise in causal identification. DoWhy relies primarily on covariate adjustment strategies such as backdoor and general adjustment, whereas PROC CAUSALGRAPH supports a wider range of identification approaches, including constructive backdoor, general backdoor, and instrumental variable methods, using constructive backdoor as the default.

Finally, while both SAS and DoWhy use doubly robust estimators for effect estimation, DoWhy does not support effect modifiers. SAS explicitly accommodates instruments, confounders, and effect modifiers within the causal structure, providing greater flexibility in model specification.

ARE THE PROCESSING STEPS DIFFERENT?

While SAS and Python follow the same high level causal workflow—discovery → identification → estimation, they differ in downstream processing and visualization support. In SAS, outputs from causal procedures are natively structured for post estimation visualization, including propensity scores, weights, and potential outcomes, enabling immediate exploratory analysis with minimal additional effort. In contrast, Python based workflows require additional post processing to generate comparable visualizations, such as manually constructing datasets for propensity scores, weights, or potential outcomes. Because DoWhy does not estimate individual level potential outcomes by default, achieving similar post estimation visualizations typically requires extra coding and data preparation compared with SAS.

WHEN SHOULD YOU USE EACH LANGUAGE?

At the level of core procedural code—independent of the low code SAS Job interface—SAS is generally more user friendly, particularly for users who are new to causal inference, programming, or both. Its documentation is clear, comprehensive, and centralized, making it easier to adopt and interpret correctly than the more fragmented Python ecosystem, which spans GitHub repositories, PyWhy resources, and third party publications. This reflects SAS's long standing investment in developing a mature, well documented analytical environment that supports users across a wide range of experience levels.

CONCLUSION

This work introduced the foundational principles of causal inference and demonstrated how these methods can be systematically implemented and integrated in SAS Viya using both SAS procedures and Python-based open source libraries. While Python offers flexibility, extensive community resources, and specialized packages for causal inference, SAS continues to set a high standard for reliability, interpretability, and regulatory readiness—particularly in regulatory environments where methodological transparency and reproducibility are essential. By providing robust built in procedures, integrated visualization capabilities, and clear, consolidated documentation, SAS enables developers and researchers to meet consistent analytical standards and to make informed decisions about the most appropriate tools and workflows for their specific research needs. Together, SAS and Python provide complementary options within SAS Viya, empowering users to leverage automation, advanced analytics, and flexible programming paradigms to generate high quality causal evidence in an efficient, end to end environment.

REFERENCES

1. Driva S, Korkontzelou A, Tonstad S, Tentolouris N, Katsaounou P. The Effect of Smoking Cessation on Body Weight and Other Metabolic Parameters with Focus on People with Type 2 Diabetes Mellitus. Int J Environ Res Public Health. 2022 Oct 14;19(20):13222. doi: 10.3390/ijerph192013222. PMID: 36293800; PMCID: PMC9603007.
2. Mishra A, Chaturvedi P, Datta S, Sinukumar S, Joshi P, Garg A. Harmful effects of nicotine. Indian J Med Paediatr Oncol. 2015 Jan-Mar;36(1):24-31. doi: 10.4103/0971-5851.151771. PMID: 25810571; PMCID: PMC4363846. https://pmc.ncbi.nlm.nih.gov/articles/PMC4363846/
3. Audrain-McGovern J, Benowitz NL. Cigarette smoking, nicotine, and body weight. Clin Pharmacol Ther. 2011 Jul;90(1):164-8. doi: 10.1038/clpt.2011.105. Epub 2011 Jun 1. PMID: 21633341; PMCID: PMC3195407. https://pmc.ncbi.nlm.nih.gov/articles/PMC3195407/#S3
4. Hu, Y., Zong, G., Liu, G., Wang, M., Rosner, B., Pan, A., Willett, W. C., Manson, J. E., Hu, F. B., & Sun, Q. (2018). Smoking cessation, weight change, type 2 diabetes, and mortality. New England Journal of Medicine, 379(7), 623–632. https://doi.org/10.1056/nejmoa1803626
5. National Institute of Diabetes and Digestive and Kidney Diseases. (September 2021). Overweight & Obesity Statistics. US National Institute of Health. https://www.niddk.nih.gov/health-information/health-statistics/overweight-obesity
6. National Institute of Diabetes and Digestive and Kidney Diseases. (May 2023). Health Risks of Overweight & Obesity. US National Institute of Health. https://www.niddk.nih.gov/health-information/weight-management/adult-overweight-obesity/health-risks
7. Kim HJ, Kim BS, Lee JH, Shin JH. Impact of underweight on 3-year all-cause mortality in patients with acute severe hypertension: a retrospective cohort study. Sci Rep. 2022 Mar 21;12(1):4798. doi: 10.1038/s41598-022-08892-9. PMID: 35314748; PMCID: PMC8938442. https://pmc.ncbi.nlm.nih.gov/articles/PMC8938442/
8. Garcia-Huidobro D, Michael Oakes J. Squeezing observational data for better causal inference: Methods and examples for prevention research. Int J Psychol. 2017 Apr;52(2):96-105. doi: 10.1002/ijop.12275. Epub 2016 Apr 20. PMID: 27094382; PMCID: PMC5549466.
9. Pearl, J. (2009). Caual inference in statistics: An overview. Statistical Surveys. Vol. 3 (2009): 96-146. ISSN:1935-7516. DOI: 10.1214/09-SS057.
10. Reizinger, P. (Nov 6, 2021). Pearls of Causality #4: Causal Queries. Casual Causality. Retrieved from https://rpatrik96.github.io/posts/2021/11/poc4-causal-queries/
11. Hernán, MA., Robins, JM. (2020). Causal Inference: What If. Boca Raton: Chapman & Hall/CRC.
12. Masten, M. (Oct 30, 2015). Statistical vs Causal Inference: Causal Inference Bootcamp. Mod-U: Powerful Concepts in Social Science. Youtube. Retrieved from https://www.youtube.com/watch?v=jfkysJxSifg
13. Reizinger, P. (Nov 6, 2021). Pearls of Causality #5: Statistical vs Causal Inference. Casual Causality. Retrieved from https://rpatrik96.github.io/posts/2021/11/poc5-stats-vs-causality/
14. Krohn, J., Taylor, S. (Oct 16, 2022). Causal Modeling: Why and when is it helpful? Super Data Science: ML & AI Podcast with Jon Krohn. Youtube. Retrieved from https://www.youtube.com/watch?v=F159s-Ar4vc
15. Hershaff, J. (Sept 4, 2025). Why Causal Inference Matters: From Marketing to Policy. Udacity. Youtube. Retrieved from https://www.youtube.com/watch?v=h0rHctkP7fQ
16. Lee, H., Aaronson, J., Nunan, D. (2019) Association or causation? How do we know? Catalogue of Bias, Nuffield Department of Primary Care. Retrieved from https://catalogofbias.org/2019/03/05/association-or-causation-how-do-we-ever-know/
17. Aandahl, EM. (). The difference between association, correlation, and causation. Lediди. Retrieved from https://lediди.com/academy/the-difference-between-association-correlation-and-causation
18. Hernán MA. A definition of causal effect for epidemiological research. J Epidemiol Community Health. 2004 Apr;58(4):265-71. doi: 10.1136/jech.2002.006361. PMID: 15026432; PMCID: PMC1732737.
19. McCormick, K. (March 18 2022). Machine Learning and AI Foundations: Prediction, Causation, and Statistical Inference. LinkedIn Career Hub. Retrieved from https://www.linkedin.com/learning/machine-learning-and-ai-foundations-prediction-causation-and-statistical-inference/prediction-causation-and-statistical-inference?dApp=218089040&leis=LAA&u=2101305

20. Irizarry, R. () Introduction to Data Science: Linear Models: Association is Not Causaltion. Github. Retrieved from https://rafalab.dfci.harvard.edu/dsbook-part-2/linear-models/association-not-causation.html
21. Barrat, H., Kirwan, M., Shantikuman, S. (2018). Association and Causation. Faculty of Public Health. Retrieved from https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1a-epidemiology/association-causation
22. Faries, D., Obenchain, R., Haro, J., & Leon, A. (2010). Analysis of Observational Health Care Data Using SAS. Sas Institute.
23. Faries, D. E., Zhang, X., Kadziola, Z., Siebert, U., Kuehne, F., Obenchain, R. L., & Haro, J. M. (2020). Real World Health Care Data Analysis: Causal methods and implementation using SAS. SAS Institute.
24. Primbs, M., Bijlstra, G., Holland, R., & Thoemmes, F. (2024, August 1). Causal Inference for Dummies: A Tutorial on Directed Acyclic Graphs and Balancing Weights.
25. Darren Dahly. (September 25, 2019). Intro to causal inference and directed acyclic graphs.
26. CodeEmporium. (January 3, 2022). Causal Inference – EXPLAINED!
27. 21. Epidemiology Stuff. (March 1, 2022). Directed Acyclic Graphs (DAGs).
28. Pearl, J. (2010). An Introduction to Causal Inference. The International Journal of Biostatistics.
29. Wang SV, Schneeweiss S. (January 2022). Assessing and Interpreting Real-World Evidence Studies: Introductory Points for New Reviewers. Clin Pharmacol Ther.
30. Schneeweiss S, Rassen JA, Brown JS, Rothman KJ, Happe L, Arlett P, Dal Pan G, Goettsch W, Murk W, Wang SV. (2019 Mar 19). Graphical Depiction of Longitudinal Study Designs in Health Care Databases. Ann Intern Med.
31. Emmott, N., Nafie, M., Kim, D., & Hendricks-Sturrup, R. (June 18, 2024). Real-World Evidence to Support Causal Inference: Methodological Considerations for Non-Interventional Studies. Duke-Margolis Institute for Health Policy.
32. SAS Help Center: Causal Inference Overview
33. Kiciman, E., Sharma, A. (March 29, 2021). Getting Started With Causal Inference; Chapter 3: Identification. Causalinference.gitlab.io. Retrieved from https://causalinference.gitlab.io/causal-reasoning-book-chapter3/
34. Unal, M. (February 22, 2023). Identification: The Key to Credible Causal Inference. Towards Data Science. Retrieved from https://towardsdatascience.com/identification-the-key-to-credible-causal-inference-c3023143349e/
35. Cole, Stephen R.a; Frangakis, Constantine E.b. The Consistency Statement in Causal Inference: A Definition or an Assumption?. Epidemiology 20(1):p 3-5, January 2009. | DOI: 10.1097/EDE.0b013e31818ef366
36. Murray, E. (January 26, 2019). Positivity: What it is and why it matters for data science. Medium. Retrieved from https://medium.com/@EpiEllie/positivity-what-it-is-and-why-it-matters-for-data-science-d5e9c0bc1fcb
37. Didelez, V., Evans, R. (November 26, 2025). Causal Inference; Chapter 9: Causal Assumptions. Stats.ox.ac.uk. Retrieved from https://www.stats.ox.ac.uk/~evans/APTS/causassmp.html
38. Neal, B. (September 7, 2020). Positivity/Overlap and Extrapolation. Youtube. Retrieved from https://www.youtube.com/watch?v=4xc8VkrF98w&list=PLoazKTcS0Rzb6bb9L508cyJ1z-U9iWkA0&index=15
39. Neal, B. (September 7, 2020). No Interference and Consistency. Youtube. Retrieved from https://www.youtube.com/watch?v=3t4HPI8Gmto
40. Rehkopf DH, Glymour MM, Osypuk TL. The Consistency Assumption for Causal Inference in Social Epidemiology: When a Rose is Not a Rose. Curr Epidemiol Rep. 2016 Mar;3(1):63-71. doi: 10.1007/s40471-016-0069-5. Epub 2016 Feb 16. PMID: 27326386; PMCID: PMC4912021.
41. Perry, R. (January 19, 2020). SUTVA vs. Exchangeability. Github. Retrieved from https://rflperry.github.io/posts/sutva_vs_exchangeability/
42. Neal, B. (September 7, 2020). Conditional Exchangeability and the Adjustment Formula. Youtube. Retrieved from https://www.youtube.com/watch?v=FFjL5Hkeap4&list=PLoazKTcS0Rzb6bb9L508cyJ1z-U9iWkA0&index=14

43. Unal, M. (March 7, 2023). Why are Randomized Experiments the Gold Standard in Causal Inference? Towards Data Science. Retrieved from https://towardsdatascience.com/why-are-randomized-experiments-the-gold-standard-in-causal-inference-f3fa240a1d02/
44. Zheng, Y., et. al. (2024). Causal-learn:Causal discovery in python. Journal of Machine Learning Research. 25(60):1-8.
45. Amit Sharma, Emre Kiciman. DoWhy: An End-to-End Library for Causal Inference. 2020. https://arxiv.org/abs/2011.04216
46. Patrick Blöbaum, Peter Götz, Kailash Budhathoki, Atalanti A. Mastakouri, Dominik Janzing. DoWhy-GCM: An extension of DoWhy for causal inference in graphical causal models. 2024. MLOSS 25(147):1−7. https://jmlr.org/papers/v25/22-1258.html
47. SAS Institute Inc. 2018. SAS/STAT® 15.1 User's Guide: The CAUSALGRAPH Procedure. Cary, NC: SAS Institute Inc. https://support.sas.com/documentation/onlinedoc/stat/151/causalgraph.pdf
48. SAS Institute Inc. (Nov 24, 2025). Overview: PROC CAUSALDISCOVERY Procedure. https://go.documentation.sas.com/doc/en/pgmsascdc/v_072/casecon/casecon_causaldiscovery_overview.htm

ACKNOWLEDGMENTS

The authors would like to thank Clay Thompson and Joe Zhao, subject matter experts who were consulted in that capacity and provided guidance to ensure the technical correctness and clarity of this paper. The authors also thank Jim Box for editorial support and mentorship throughout the development of the paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Laura Watson

Company: SAS Institute, Inc

Address: 100 SAS Campus Drive, Cary, NC27513

Work Phone: 919.531.4894

Email: Laura.Watson@sas.com

Website: www.sas.com