Assessing Causal Discovery Algorithms for Smoking-Outcome Relationships: A NHANES Testbed

Simran Mallik
smallik9@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Yuritzy Ramos
yramos3@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Jeremy Thomas
jthomas483@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Abstract

The Peter Clark (PC) and Greedy Equivalence Search (GES) algorithms have been used in previous work to discover causal relationships between variables in synthetic smoking-related datasets. The literature on the subject has not yet applied these algorithms on real-world data and evaluated their effectiveness in identifying causal relationships in this context. Our contribution to the literature is providing a test bed for testing causal inference algorithms on smoking related outcomes using data from the Center for Disease Control's National Health and Nutrition Examination Survey. This paper explores the PC and GES algorithms on this testbed. The results of this research show that while the algorithms can find some causal relationships, the effectiveness of the model in identifying directed relationships and correctly directed relationships is low. Future work should focus on testing the impact of confounding variables on the algorithms' ability to find causal relationships to smoking.

Keywords

causal inference, epidemiology, NHANES, public health, PC algorithm, GES algorithm, smoking, causal discovery, data science

1 Motivation and Problem Description

Causal inference is a critical tool in epidemiology, enabling researchers to uncover relationships between exposures and health outcomes and thereby inform public health interventions. Understanding the causal links between risk factors, such as smoking, and disease outcomes allows epidemiologists and public health experts to design targeted interventions (such as anti-smoking campaigns, early screening programs, or policies to reduce exposure to harmful substances) to reduce the incidence and burden of preventable diseases.

While the relationship between smoking and various diseases has been extensively studied, the primary motivation of this project is methodological: to compare and evaluate the performance of two prominent causal inference algorithms, PC (Peter-Clark) and GES (Greedy Equivalence Search), when applied to real-world epidemiological data. This comparison is novel because, although these algorithms have been validated on synthetic datasets, they have rarely been systematically applied to real-world data to assess their ability to correctly identify known causal relationships. By focusing on smoking and disease outcomes, which are well-established and can serve as a ground truth, we can rigorously evaluate the strengths and limitations of PC and GES and gain insights into how these methods may be applied to other epidemiological causal inference problems where ground truth is less certain.

The research leverages cross-sectional data from the Centers for Disease Control and Prevention's National Health and Nutrition Examination Survey (NHANES) from 2009–2018, covering approximately 28,500 individuals aged 10–89. NHANES provides a nationally representative dataset with reliable self-reported diagnoses, detailed demographic measures, and standardized anthropometric data. While the dataset has limitations, such as the cross-sectional design and differences between "ever told" diagnoses and current smoking status, the findings from this project are treated as hypothesis-generating.

The research first focused on constructing a causal testbed using the NHANES dataset to evaluate the relationship between individual smoking status and selected disease outcomes. Ground truth causal relationships were defined using accepted findings from the medical literature. Following the creation of the testbed dataset, the PC and GES algorithms were applied to the dataset. These two algorithms were specifically chosen for comparison because they represent fundamentally different approaches to causal inference, allowing us to evaluate how methodological differences impact the identification of causal relationships in real-world data. The PC (Peter-Clark) algorithm is constraint-based: it uses conditional independence tests to iteratively remove edges between variables that are independent and then orients the remaining edges according to rules consistent with these independence relationships. In contrast, the GES (Greedy Equivalence Search) algorithm is score-based: it iteratively adds and removes edges to maximize a scoring function, the Bayesian Information Criterion, thereby identifying the causal graph that optimally balances model fit and complexity. By comparing a constraint-based approach with a score-based approach, we can assess how these methodological differences influence the accuracy, interpretability, and robustness of the inferred causal graphs

Because demographic factors can influence the observed relationships between smoking and disease, subgroup analyses were performed across age, ethnicity, and gender. Differences in algorithm performance within and across demographic subgroups, as well as in comparison to the baseline, are reported in the results. These analyses allow for a more nuanced understanding of how PC and GES perform in heterogeneous populations.

The end deliverables for this project will include a reproducible Jupyter Notebook implementing the causal inference methods, graph visualizations for the NHANES dataset, and an analysis report discussing algorithm methods, results, and performance. By systematically evaluating PC and GES on real-world epidemiological data, this project provides a foundation for applying these causal inference methods to other public health challenges where

identifying causal relationships is critical for designing effective interventions.

2 Ground Truth

Because the NHANES dataset only provides a collection of survey data, it is not directly possible to identify the ground truth relationships between the features that are analyzed. In order to provide a foundation for the causal discovery analysis, prior epidemiological and clinical studies that have established causal relationships between smoking and various chronic diseases will act as ground truth for this study. This "ground truth" will then be compared with the Directed Acyclic Graphs that are produced by the two methods explored to evaluate the effectiveness of each model.

There is notable evidence that smoking has a causal relationship with several health conditions. Respiratory diseases such as Chronic Obstructive Pulmonary Disease (COPD), chronic bronchitis, and emphysema have shown causal patterns [5]. A meta-analysis of 218 studies noted elevated relative risk for current smokers compared to nonsmokers at the following values: COPD: 3.51, chronic bronchitis: 3.41, and emphysema: 4.87, confirming the causal relationships with smoking [4]. Cardiovascular outcomes have been similarly researched in terms of dose response and temporal relationships. Smoking is a determinant of coronary heart disease (CHD) and stroke; also, secondhand smoke exposure is expected to account for more than 30% of global CHD mortality [2, 6]. A meta-analysis of 55 studies found that smoking one cigarette per day produces half the risk of CHD as smoking 20 cigarettes a day would, illustrating the non-linear effects of smoking on vascular health conditions [11].

Oncologic outcomes for smoking have also been established in research. Smoking remains a predominant factor in lung cancer, as it is responsible for 80% of male lung cancer deaths globally [1]. Smoking also increases asthma severity and progression. Adult asthmatics who smoke show worsened corticosteroid response and reduced lung function compared to non-smoking counterparts, and quitting smoking has shown improvements in both lung function and symptoms of asthma patients [3].

Finally, liver and thyroid conditions also arise due to smoking. Cigarette smoke exposure leads to a progression of fibrosis in chronic liver diseases such as nonalcoholic fatty liver disease and primary biliary disease [12]. In addition, a cross-sectional study found that smokers were more likely to have relatively low thyrotropin concentrations. This supported later findings that current smokers were at increased risk for hyperthyroidism. On the other hand, smokers were found to have a negative relationship with hypothyroidism [15] .

3 Related Work

Applying Causal Inference methods to complex, multivariate epidemiological and clinical datasets have been explored but still need to be researched more deeply. Sinha et al. (2025) explore causal analysis for multivariate clinical and environmental exposures data by using random forests to determine feature importance for Directed Acyclic Graphs and using the PC algorithm to infer causal relationships, accounting for confounding variables and providing interpretability regarding the relationship between variables [13]. However, a limitation noted by the authors is that causal relationship(s) may be overlooked if patient data is imbalanced across the subgroups, which is a factor that will be explicitly monitored when using NHANES data. Similarly, Nur et al. apply both the Peter-Clark (PC) and Greedy Equivalent Search (GES) algorithms to model causal factors on health variables that lead to growth stunting [8]. This study highlights how applying PC (a constraint-based approach) to create the structure of the initial graph and combining with GES (a score-based method) to refine directionality of the causal relationships can lead to interpretable causal models. This paper suggests that GES can be useful in identifying directional relationships, highlighting the rationale for comparative analysis.
Zhu et al.introduces a hybrid constrained continuous optimization approach for discovering causal relationships in biological data [14]. The researchers integrate the PC algorithm with NOTEARS to balance accuracy and effect estimation. This hybrid approach displayed robustness to noise, small sample sizes, and hidden confounders, and outperformed non-hybrid, traditional algorithms like PC, GES, and other nonlinear models. This hybrid approach excelled in producing stable causal graphs despite challenging conditions regarding the data, however it still involves conditional independence testing which may not generalize well to nonlinear datasets or large-scale datasets. The researchers highlight that actual causal effects can't be determined using just one algorithm on its own, suggesting combining algorithms, like their hybrid approach, if the results are not robust. Additionally, Montagna et al. compare different causal algorithms on epidemiological data and show that score-based methods like GES and NOTEARs thrive under ideal conditions (linear relationships, low noise, and large datasets), whereas constraint-based methods like PC fall behind [7]. This highlights the balance between theoretical performance and applicability to real-world data, which emphasizes the necessity for robust analyses. This is especially relevant to the NHANES dataset which includes mixed data types, missing values, and complex relationships.
Pearce and Lawlor emphasize that assumptions in causal modeling (causal sufficiency and faithfulness) need to be clearly highlighted and that pluralistic approaches (triangulating methods) should be considered [9]. Though this paper does not delve into hidden confounders and model selection trade-offs, our project addresses such gaps by implementing bootstrapping, testing different parameters, and performing sensitivity analyses to evaluate the interpretability and stability of the causal edges. Additionally, Raghu et al. compares causal structure algorithms like PC and GES on mixed data types, using standard performance metrics like Structural Hamming Distance (SHD) and precision and recall to evaluate accuracy [10]. Though this study provides guidance on improving algorithm performance, it relies on synthetic data and assumes complete data. It also lacks analysis of algorithm interpretability. However, our project will address this by applying these methods to real NHANES data which includes information on demographics, health behaviors, and health outcomes, which will be validated by comparing causal structures against epidemiology knowledge.
Altogether, these studies provide conceptual and methodological guidance for our project. The papers highlight the strengths of constraint-based approaches like PC and score-based methods like GES, how to handle mixed data types, and the necessity for robustness and interpretability checks. Our project builds on this research

by comparing PC and GES for real epidemiology data, implementing bootstrap techniques and sensitivity analysis for addressing missing data and variability in samples, and qualitatively evaluating the causal graphs with previously establish causal relationships between smoking and health outcomes.

4 Testbed Development

The final dataset used for this research has been sourced using the CDC's NHANES (National Health and Nutrition Examination Survey) for the years 2009-2010, 2011-2012, 2013-2014, 2015-2016, and 2017-2018. Each year includes multiple respondent datasets containing information on medical conditions, physical characteristics and body measurements, demographics, and smoking habits.

Because different respondent datasets cover different age groups, left joining all datasets on the respondent sequence number did not result in perfect matches. To ensure consistency and improve the accuracy of causal inference, we opted to inner join the datasets, retaining only individuals present in all datasets. We further filtered the combined dataset to include only respondents with information on their smoking status, as this is our primary variable of interest. This resulted in the final dataset that consists of approximately 28,500 individuals.

Respondents were categorized as never smokers, current smokers, or former smokers. Never smokers are defined as individuals who have smoked fewer than 100 cigarettes in their lifetime, current smokers are those who smoke either some days or every day, and former smokers are individuals who have smoked at least 100 cigarettes in their lifetime but do not currently smoke.

The testbed developed then filters the final dataset to only contain the following features: Smoking Status, Income, Age, Gender, Ethnicity, BMI, Chronic Obstructive Pulmonary Disease, Chronic bronchitis, Emphysema, Heart Attack, Coronary Heart Disease (CHD), Congestive Heart Failure (CHF), Stroke, Cancer or malignancy, Liver conditions, Thyroid problems, Asthma. The following features were selected due to the established causal relationships between smoking and the selected diseases.

5 Exploratory Data Analysis

Our dataset includes 28,551 individuals. Overall, 52% of respondents are female and 48% are male. Across all demographic groups (age, ethnicity, and gender), never smokers are more common than both current and former smokers.

The mean age of participants is approximately 49 years (median: 48 years). The race/ethnicity distribution is as follows: 39% Non-Hispanic White, 22% Non-Hispanic Black, 15% Mexican American, 14% Other/Multi-Racial, and 10% Other Hispanic.

Age-based smoking patterns show clear trends. Among individuals aged 10–49, the proportion of current smokers exceeds that of former smokers. In the 50–59 age group, current and former smoking rates are similar, whereas for ages 59–89, former smokers outnumber current smokers. Gender differences also emerge: females are more likely to be never smokers compared to males and are less likely to be former or current smokers.

Smoking patterns vary by race/ethnicity as well. All groups contain more never smokers than current or former smokers. For Mexican American, Non-Hispanic White, Other Hispanic, and Other/Multi-Racial groups, former smokers are more common than current smokers. In contrast, Non-Hispanic Black individuals show the reverse pattern, with current smokers slightly surpassing former smokers.

BMI distributions are right-skewed for all smoking categories. Median BMI values are 27.3 for current smokers, 28.8 for former smokers, and 28.0 for never smokers.

Finally, disease prevalence varies strongly by smoking status. Conditions such as MCQ160O (chronic bronchitis), MCQ160K, and MCQ010 occur most frequently among current smokers. Other conditions, including MCQ160E, MCQ160C, MCQ160B, MCQ220, MCQ160M, and MCQ160F, occur most frequently among former smokers. Overall, nearly all diseases examined were more prevalent among former and current smokers compared to never smokers.

6 Methodology

For this analysis, the Peter-Clark (PC) and Greedy Equivalence Search (GES) algorithms will be used to identify causal relationships within the NHANES dataset. These two algorithms distinctly identify causality using constraint methods and score-based ordering.

The PC algorithm identifies causal relationships between nodes by performing a series of conditional independence tests for every node pair. The algorithm begins with a complete graph, then edges are removed between nodes that are independent. Kernel conditional independence is used to test for independence between nodes in order to handle nonlinear dependencies. If there is a subset of nodes (S) outside of the full set (V) for which the two nodes tested (Xi , Xj) would be independent, then the edges between those two nodes would be removed. The formal definition can be seen in Equation 1.

$$X_i \perp X_j | S \; \forall \; S \subseteq V / X_i, X_j \tag{1}$$

Once this process is complete, a skeletal graph is produced. This skeletal graph is composed of v-structures where two nodes both point to the same node and undirected edges where the orientation cannot be uniquely identified. After this step, Meek's rules is applied to convert undirected edges to directed edges when possible. This algorithm will output a Completed Partially Directed Acyclic Graph (CPDAG) representing the DAGs with independence relations exhibited in the data. The GES algorithm identifies causal relations by maximizing a likelihood function (Equation 2) with Bayesian Information Criterion (BIC). For this algorithm, G represents a DAG over nodes Xi, ..., Xm. Then, the joint distribution for each node can be represented as shown in Equation 2. The scoring criterion S(G,D) is seen in Equation 3 where D represents each data point in the dataset.

$$P(X_1, ..., X_m) = \Pi_{i=1}^{d} (X_i | P_{aG}(i)) \tag{2}$$

$$(G, D) = \sum_{i=1}^{d} (X_i, P_{aG}(i); D) \tag{3}$$

The GES algorithm begins with an empty graph and iteratively adds edges. There are two phases to the process to identify edges presented between nodes in the dataset. During the forward phase, GES greedily adds edges that maximize the objective function (Equation 3). In its backwards phase, edges are removed that reduce or do not change the objective function to reduce the complexity of the graph. Once this algorithm is completed, a CPDAG is produced, representing the markov equivalence classes with the highest BIC score.

The PC and GES algorithms are tested with 30 bootstraps to generate a more representative graph for the dataset and evaluate the stability of the identified causal edges. Then, to identify demographic shifts that might affect the causal graph produced by PC and GES, the algorithms were also applied to various subgroups based on age, gender, and ethnicity, each tested with 30 bootstraps.

7 Evaluation

The effectiveness of the PC and GES models was evaluated using the edge frequencies between Smoking Status and both health-related and non-health variables, as presented in the Testbed Development section. The edge frequencies for the full testbed represent the baseline causal relationships identified by each algorithm. High-frequency edges indicate relationships that the algorithm consistently discovers across the 30 bootstrapped samples, suggesting stronger or more stable causal connections. In contrast, low-frequency edges indicate relationships that are weakly or inconsistently identified.

After establishing these baseline results, both algorithms were applied to each demographic subgroup following the methodology presented. The subgroups analyzed produced their own edge-frequency results. To assess whether demographic factors influence the causal graphs discovered, we compared differences in edge frequencies across subgroups within each demographic category. Then, these differences were statistically tested using the Kruskal–Wallis test, to compare distributions of edge frequencies. The results of this test reveal how demographic factors can affect the causal structure identified and how the performance of PC and GES differs in their sensitivity to demographic variation and effectiveness when the dataset size decreases. Additionally, edge stability across subgroups is evaluated across all demographic categories for both algorithms. Edges with a high mean frequency and low variance across subgroups are considered stable, suggesting they represent true underlying relationships and are not confounded by demographic relationships. Confidence intervals are also computed for each edge's frequency distribution. These intervals help quantify uncertainty in edge detection across bootstrapped samples. This is used to measure the sensitivity of the algorithms to the partitioned datasets and its ability to identify causal relationships.

8 Results

A baseline was produced for each of the algorithms using the entire testbed data. Both the PC and the GES algorithms were highly consistent in discovering the initial causal structure of the data, identifying the same set of relationships connected to Smoking Status. Their agreement indicates that the underlying causal signals in the data are strong and not an artifact of a single algorithm's design. The algorithms also agreed on the relationships they did not find, ignoring the same edges such as those leading to Thyroid Problem, Congestive Heart Failure, and Age.

The primary difference between both algorithms was the robustness of the link between Smoking Status - BMI While the PC algorithm found this link to be present in the bootstrap samples 100% of the time, the GES algorithm identified this link in about 80% of the samples. This difference was due to the fact that the GES algorithm is less likely to state links based on conditional dependence. This contrast suggests that while the edge is likely real, the PC algorithm's reliance on strict statistical independence tests led to more certain and consistent results compared to GES which focuses on score-optimization which is more sensitive to variations in the resampled data. If including an edge does not improve the score for the graph produced from the data, GES will drop this edge more frequently in the bootstrapped samples.

Next, edge frequency and variation from the baselines were calculated for the demographic factors in the testbed. Each demographic factor was divided into its respective subgroups after which heatmaps and barplots were produced to examine differences in edge frequencies and standard deviation. This will be explained in the sections right below.

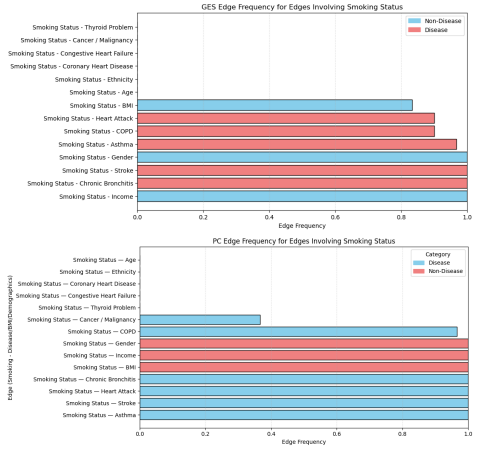

Figure 1: Baseline Edge Frequencies for PC and GES

8.1 Age

The consistency found in the full dataset analysis, where both PC and GES agreed on the same causal relationships, sharply contrasts with the results among the age subgroups. In the heatmaps we see that the frequency of detected edges is much lower for the GES algorithm with several values falling below the 0.4 frequency level. This systematic difference is primarily due to the algorithms' mechanics. The PC algorithm (constraint-based) uses a fixed, lenient statistical threshold (alpha=0.1). This approach makes it more eager to find edges, ensuring that robust signals persist across smaller subgroups, leading to uniformly high frequencies. Conversely, the GES algorithm (score-based) uses the BIC score, which heavily penalizes graph complexity in smaller samples. This penalty forces GES to be highly conservative, dropping edges that are not strictly

necessary to achieve the highest, simplest score. This score-based conservatism drives the systematically lower and more variable frequencies observed in GES, which could be misinterpreted as unique sensitivity to heterogeneity when it may largely reflect the algorithm's inherent penalty for complexity in low-powered samples. The one major exception is the Smoking Status - Gender edge, which exhibits a perfect 1.0 frequency across all age groups for both algorithms. Gender and smoking status seem to exhibit a close relationship, as seen in our exploratory analysis. This stability exists because the strong statistical dependence between smoking rates and gender is a mandatory constraint in the causal model. Both algorithms are forced to include this edge consistently because excluding it would create a severely misspecified model that fails to account for this fundamental, robust dependence in the data, regardless of the subgroup sample size.

Differences in the algorithms' core design also led to sharp contrasts when analyzing subgroup variability for age. The PC algorithm demonstrates high reliability and generalizability. PC uses a fixed, lenient statistical threshold (alpha=0.1), making its SD bars (representing edge variability) remain consistently near zero. This indicates that PC's causal findings are highly stable across all demographic cuts, and they accurately generalize from the overall population to the subgroups.In contrast, the GES algorithm shows greater instability. Its score-based approach (using BIC) causes it to become highly conservative in the smaller subgroups, leading many of its edge frequencies to fall below the 0.4 level. This penalization of complexity leads to significant variability in specific edges, such as Smoking Status - COPD, where GES shows a large difference between internal subgroup variation and the overall baseline. This suggests that the average population conclusion for the COPD link may not hold true for individual subgroups. The one perfectly stable relationship is Smoking Status - Gender, which shows near-zero variability for both algorithms. This link is not a causal conclusion (Gender cannot be caused by Smoking) but a mandatory confounding constraint that must be included to represent the strong, pre-existing statistical relationship in the dataset. (Findings reference Figure 2 and Figure 3).

8.2 Ethnicity

Analyzing the ethnicity heatmaps also highlights that the PC algorithm is more permissive in subgroup analysis, while the GES algorithm is more conservative. The PC algorithm maintains high frequencies (age 0.5) for a range of edges across most ethnic subgroups (e.g., Mexican-American, Other Hispanic, and Other Race). This is likely due to its lenient alpha=0.1 threshold and constraint-based nature, which easily finds statistical support for demographic factors like Age, Gender, and Ethnicity. In contrast, the GES algorithm shows significantly lower frequencies overall because its score-based approach (BIC) penalizes complexity in small subgroups. This conservatism confines its high stability to fewer groups (Non-Hispanic White, Non-Hispanic Black, and Other Race). Both agree on the most robust edge: Smoking Status - Gender is 100% stable across all groups, confirming its role as a mandatory confounding constraint. Also, GES highlights plausible heterogeneity by only conserving the most critical links for each group: the Mexican-American subgroup shows high conservation for Smoking Status -

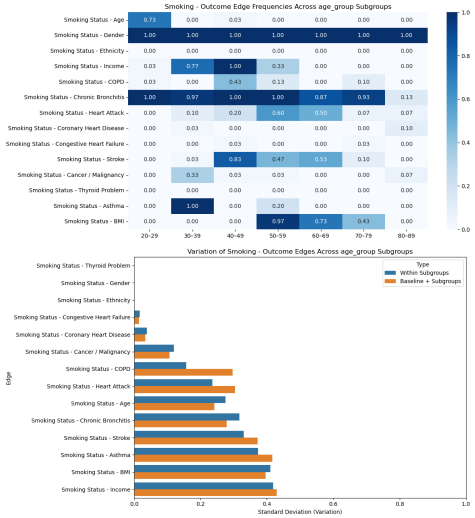

Figure 2: GES: Edge Frequency and Variation from Baseline for Age subgroups

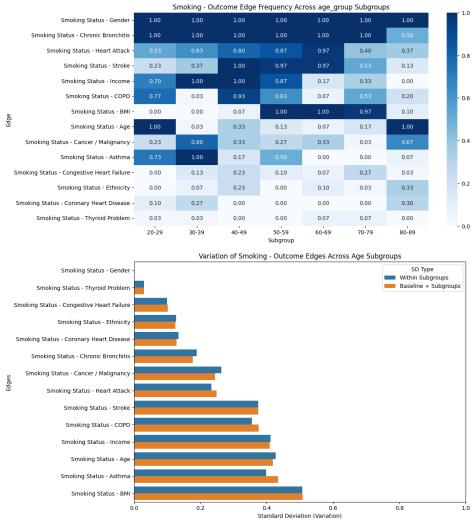

Figure 3: PC: Edge Frequency and Variation from Baseline for Age subgroups

Income (approx 1.0), while the Non-Hispanic Black subgroup shows high conservation for Smoking Status - BMI (approx 1.0). This demonstrates that the dominant pathways through which smoking influences other variables differ between these ethnic groups.

The stability analysis confirms that the PC algorithm may be more robust and generalizable across subgroups, while the GES algorithm highlights the challenge of low statistical power in partitioned data. The PC algorithm maintains high stability: its Within Subgroup standard deviation (blue bars) and Baseline + Subgroup standard deviation (orange bars) are consistently close, demonstrating generalizability across ethnic subgroups due to its permissive alpha=0.1 threshold. In contrast, the GES algorithm shows much

higher instability. Its score-based approach (BIC) penalizes complexity heavily in low-power subgroups, causing the algorithm to frequently drop edges. This instability is most pronounced for the known links, Smoking Status - COPD and Smoking Status - Asthma. The large difference between the blue and orange bars in the GES plot for these edges confirms that the frequency established in the overall population baseline is significantly inconsistent with the frequencies found in the individual ethnic subgroups. This suggests the GES algorithm is too conservative for this low-powered subgroup analysis, as its model penalty overrides the evidence for these links. (Findings reference Figure 4 and Figure 5).

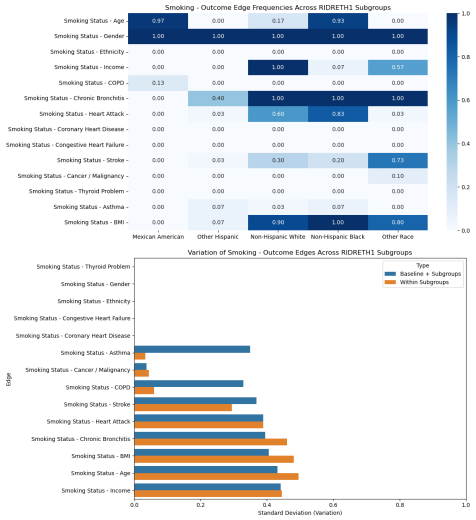

Figure 4: GES: Edge frequency and Variation from Baseline for Ethnicity Subgroups

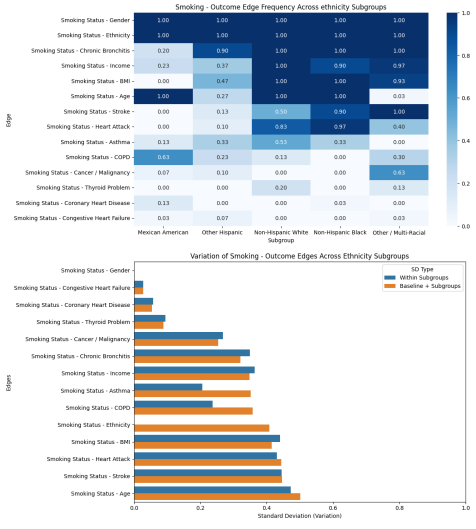

Figure 5: PC: Edge frequency and Variation from Baseline for Ethnicity Subgroups

8.3 Gender

When looking at the Gender demographic group, PC finds more causal relationships that are the same between both genders while GES discovers strong causal relationships that are usually applicable to only one subgroup. For example, we see that PC finds strong causal links between Smoking Status and the following for both male and females: Stroke, Heart Attack, and Chronic Bronchitis. Meanwhile, GES only detects one such relationship: Smoking Status - Chronic Bronchitis. The remaining strong causal relationships are for a single gender such as Smoking Status - Income and BMI for females or Smoking Status - Asthma and Heart Attack for males.

Like the ethnicity and age subgroup analysis, PC maintains high frequencies for a wide range of edges across the gender subgroups, which may be due to its eager nature. GES on the other hand is more conservative. GES has lower frequency values for the edges linking Smoking Status - COPD and Asthma, despite these being known edges. This demonstrates that GES's score-based approach (BIC) penalizes complexity in the smaller gender samples, causing it to drop edges frequently. Consequently, the GES results serve to highlight the few relationships strong enough to overcome the complexity penalty and remain stable across both male and female groups.

In terms of variability, the GES bar plot shows that several causal relationships at the subgroup level have greater variation compared to the baseline + subgroups. We see a similar trend for PC though less pronounced than for GES. This means that there may be some heterogeneity in the Gender demographic group. The large differences between the blue and orange bars in the GES plot confirm that the frequency established in the baseline is significantly inconsistent with the frequencies found in the individual gender subgroups. This implies that the GES algorithm is too conservative for this low-power subgroup analysis, as its model penalty overrides the evidence for these links. (Findings reference Figure 6 and Figure 7)

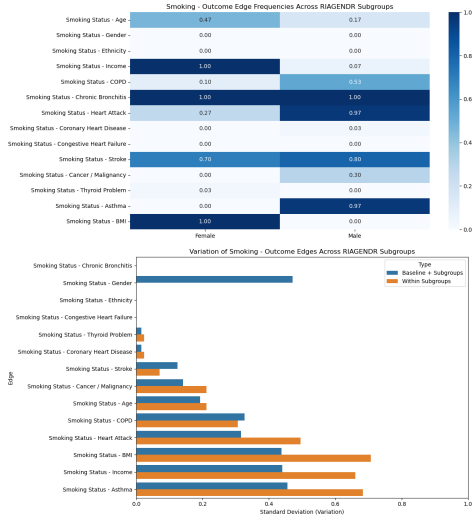

Figure 6: GES: Edge Frequency and Variation from Baseline for Gender Subgroups

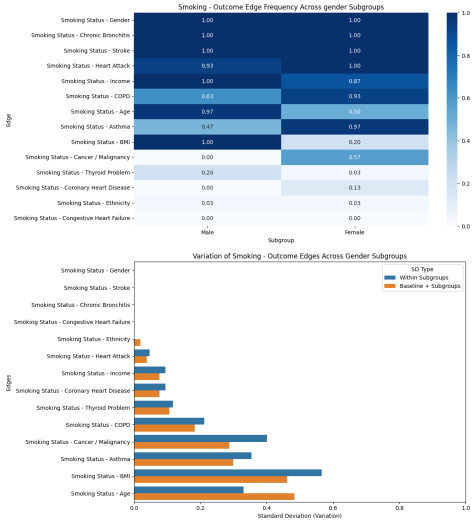

Figure 7: PC: Edge Frequency and Variation from Baseline for Gender Subgroups

Figure 8: Frequent, Stable Edges for GES (top) and PC (bottom)

8.4 Edge Stability Across Subgroups

By examining how edge frequencies vary across all subgroups and demographic categories for both the PC and GES algorithms, we can assess the stability of each edge. A stable edge is one that consistently appears across all groups in the data, suggesting it may represent a true underlying relationship rather than a pattern driven by a specific subgroup.

Under the PC algorithm, the most stable and frequent relationship (indicated by a high mean frequency and low standard deviation) is Smoking Status – Gender, followed by Smoking Status – Chronic Bronchitis. Under the GES algorithm, the relationships between Smoking Status – Gender, Smoking Status – Chronic Bronchitis, and Smoking Status – BMI show similarly high stability.

Overall, the edges Smoking Status – Gender and Smoking Status – Chronic Bronchitis appear as the most consistent and stable across both algorithms and across all demographic groups. (Findings reference Figure 9)

8.5 Confidence Intervals

We observe that the frequency and confidence intervals are higher for PC compared to GES for the following relationships and Smoking Status: BMI, asthma, heart attack, COPD, and cancer/malignancy. The remaining relationships are similar between the two algorithms.

8.6 Kruskal Wallis

To complement our stability analysis from the edge-frequency spreads, we applied a Kruskal–Wallis test to each edge within each demographic (age, ethnicity, gender). For every edge, the test compares the bootstrap edge-frequency distributions across that demographic's subgroups to determine whether the edge is consistent across a demographic group or not. A significant result indicates heterogeneity, meaning the causal relationship is not equally stable across subgroups in a demographic. A non-significant result suggests the edge is consistent across that demographic. Doing this allows us to observe if a causal relationship is stable across all demographics, a few demographics, or a single demographic.

For both PC and GES, the Smoking Status and Gender causal relationship was stable across all three demographic groups (at a

GES Algorithm
Baseline Edge Frequency with 95% Confidence Interval

Edge	Edge Frequency	CI Lower Bound	CI Upper Bound
Smoking Status – Income	1.00	0.89	1.00
Smoking Status – Chronic Bronchitis	1.00	0.89	1.00
Smoking Status – Stroke	1.00	0.89	1.00
Smoking Status – Gender	1.00	0.89	1.00
Smoking Status – Asthma	0.97	0.83	0.99
Smoking Status – COPD	0.90	0.74	0.97
Smoking Status – Heart Attack	0.90	0.74	0.97
Smoking Status – BMI	0.83	0.66	0.93
Smoking Status – Age	0.00	0.00	0.11
Smoking Status – Ethnicity	0.00	0.00	0.11
Smoking Status – Coronary Heart Disease	0.00	0.00	0.11
Smoking Status – Congestive Heart Failure	0.00	0.00	0.11
Smoking Status – Cancer / Malignancy	0.00	0.00	0.11
Smoking Status – Thyroid Problem	0.00	0.00	0.11

PC Algorithm
Baseline Edge Frequency with 95% Confidence Interval

Edge	Edge Frequency	CI Lower Bound	CI Upper Bound
Smoking Status – Gender	1.00	0.886487	1.00
Smoking Status – Income	1.00	0.886487	1.00
Smoking Status – Asthma	1.00	0.886487	1.00
Smoking Status – Heart Attack	1.00	0.886487	1.00
Smoking Status – Chronic Bronchitis	1.00	0.886487	1.00
Smoking Status – Stroke	1.00	0.886487	1.00
Smoking Status – COPD	0.966667	0.833296	0.99
Smoking Status – Cancer / Malignancy	0.366667	0.218739	0.54
Smoking Status – Age	1.00	0.00	0.11
Smoking Status – Ethnicity	1.00	0.00	0.11
Smoking Status – Coronary Heart Disease	1.00	0.00	0.11
Smoking Status – Congestive Heart Failure	1.00	0.00	0.11
Smoking Status – Thyroid Problem	1.00	0.00	0.11

Figure 9: Baseline Edge Frequency with 95% Confidence Interval for GES and PC

5% significance level). While this was the only stable relationship across all groups for PC, Smoking Status and Ethnicity was also determined to be a stable edge for GES. Some of the remaining causal relationships were determined to be stable for one or two of the demographic groups. For example, the following relationships with smoking status were stable in the ethnicity demographic group for GES (at the 5% significance level): cancer/malignancy, asthma, and coronary heart disease.

Overall, the PC algorithm appears to have more unstable edges compared to GES when applying this test. However, GES is not able to identify relationships frequently given its conservative nature, making it less varied than PC. Thus, PC appears to have more unstable edges despite its ability to identify causal relationships at a higher rate.

PC Algorithm

Stable: All Demographics

- Smoking Status – Gender**

Stable: ≤ 2 Demographics

- Smoking Status – MCQ160B**
- Smoking Status – MCQ160F**
- Smoking Status – MCQ160K**
- Smoking Status – Ethnicity**

Unstable: All Demographics

- Smoking Status – BMI**
- Smoking Status – INDHHIN2**
- Smoking Status – MCQ160E**
- Smoking Status – MCQ160M**
- Smoking Status – MCQ160O**
- Smoking Status – MCQ160C**

GES Algorithm

Stable: All Demographics

- Smoking Status – Gender**
- Smoking Status – Ethnicity**
- Smoking Status – INDHHIN2*

Stable: ≤ 2 Demographics

- Smoking Status – MCQ220**
- Smoking Status – MCQ010**
- Smoking Status – MCQ160C**
- Smoking Status – RIDAGEYR**
- Smoking Status – MCQ160F**
- Smoking Status – MCQ160K**

Unstable: All Demographics

- Smoking Status – BMI**
- Smoking Status – MCQ160E**
- Smoking Status – INDHHIN2**
- Smoking Status – MCQ160O**

** p-value < 0.05, ** p-value < 0.01*

9 Limitations and Future Work

The results of our analysis show that while PC and GES are able to discover causal relationships between Smoking Status and other variables in the NHANES data set, their different approaches (constraint-based vs score-based) impact their ability to find generalizable insights. We also observed that PC was unable to determine the directionality of these relationships, while those found by GES did not adhere to our ground truths. There are a few reasons why we observed these findings. One possibility is the existence of confounding variables that affect the algorithms' ability to find causal relationships. For example, age may have a strong association with multiple variables in the dataset which can lead the algorithms to view the addition of these edges as a necessary step to satisfy their associated independence test or score. This can cause the algorithms to overlook weaker causal relationships between behavioral and disease variables. A related possibility is that the algorithms are picking up on latent variables not present in the data, which would naturally serve as a stepping stone from one variable to another. In our case, age might be associated with a latent behavioral variable that would cause a disease. This would lead the algorithms to view the addition of an edge between age and this disease as improving the overall DAG. Meanwhile, the absence of such variables could also weaken the associations between smoking and diseases, causing their frequency in the bootstrap samples and directionality to be obscured. Lastly, the cross-sectional structure of our NHANES testbed data might compound some relationships, such as those between demographics and smoking status, while weakening others. Although we compiled data from 2009 to 2018, time was not a factor in our data set that could be used by the algorithms. This means that the order of events was not taken into account by the algorithms when searching for causal relationships. This makes it harder to observe feedback loops and confounding/latent variables that could be driving relationships between two variables.

Future iterations of this work should consider running experiments with and without demographic factors to determine if these confound potential relationships between smoking and disease variables. Another recommendation is to use longitudinal data, as time is a factor that could affect the relationship between smoking and the progression of a disease (i.e., how long a person has smoked).

10 Conclusions

Both PC and GES successfully identified key causal edges in the baseline NHANES data, particularly links between Smoking Status and demographic or health variables. In subgroup analyses, PC consistently produced higher edge frequencies than GES, likely due to its higher alpha threshold and constraint-based, eager approach. GES, by contrast, was more conservative and often dropped edges in smaller subgroups, but this conservatism occasionally highlighted potential heterogeneity, as edges with large frequency differences across subgroups may indicate real subgroup-specific effects.

While GES can estimate edge directionality, most inferred directions did not match known relationships, suggesting limitations when applying these algorithms to cross-sectional data with complex confounding. Both algorithms repeatedly identified edges between Smoking Status and demographic outcomes, highlighting the strong influence of demographic factors and the need to account for them in causal analyses of health behaviors.

Overall, PC appears more robust for subgroup analyses, providing stable and generalizable edge discovery, while GES may be useful for identifying potential heterogeneity but struggles with low-powered subgroups. Future work should explore longitudinal data and the inclusion/exclusion of demographic variables to better disentangle causal relationships between smoking and health outcomes.

References

[1] Feiling Ai, Jian Zhao, Wenyi Yang, and Xia Wan. 2023. Dose–response relationship between active smoking and lung cancer mortality/prevalence in the Chinese population: a meta-analysis. *BMC Public Health* 23, 747 (2023). doi:10.1186/s12889-023-15529-7

[2] Dagfinn Aune, Sabrina Schlesinger, Teresa Norat, and Elio Riboli. 2019. Tobacco smoking and the risk of heart failure: A systematic review and meta-analysis of prospective studies. *European Journal of Preventive Cardiology* 26, 3 (2019), 279–288. doi:10.1177/2047487318806658

[3] Vanesa Bellou, Athena Gogali, and Konstantinos Kostikas. 2022. Asthma and Tobacco Smoking. *Journal of Personalized Medicine* 12, 8 (2022), 1238. doi:10.3390/jpm12081238

[4] Chiwook Chung, Kyu Na Lee, Kyungdo Han, Dong Wook Shin, and Sei Won Lee. 2023. Effect of smoking on the development of chronic obstructive pulmonary disease in young individuals: a nationwide cohort study. *Frontiers in Medicine* 10 (2023), 1190885. doi:10.3389/fmed.2023.1190885 Open Access under CC BY 4.0.

[5] Barbara A. Forey, Alison J. Thornton, and Peter N. Lee. 2011. Systematic review with meta-analysis of the epidemiological evidence relating smoking to COPD, chronic bronchitis and emphysema. *BMC Pulmonary Medicine* 11, 36 (2011), 1–61. doi:10.1186/1471-2466-11-36

[6] Giuseppina Gallucci, Alfredo Tartarone, Rosa Lerose, Anna Vittoria Lalinga, and Alba Maria Capobianco. 2020. Cardiovascular risk of smoking and benefits of smoking cessation. *Journal of Thoracic Disease* 12, 7 (2020), 3866–3876. doi:10.21037/jtd.2020.02.47

[7] F. Montagna et al. 2023. Assumption Violations in Causal Discovery and the Robustness of Score Matching. *arXiv preprint arXiv:2310.13387* (2023). https://export.arxiv.org/pdf/2310.13387v2

[8] Y. S. R. Nur, A. Sa'adah, D. Aldo, and B. Masulah. 2025. Causal Modeling of Factors in Stunting Using the Peter-Clark and Greedy Equivalence Search Algorithms. *JITK* 10, 3 (2025), 523–533. https://ejournal.nusamandiri.ac.id/index.php/jitk/article/view/6184

[9] N. Pearce. 2017. Causal Inference—So Much More Than Statistics. *International Journal of Epidemiology* 46, 5 (2017), 1603–1605. doi:10.1093/ije/dyx116

[10] A. Raghu et al. 2019. Evaluation of Causal Structure Learning Methods on Mixed Data Types. *PLoS ONE* 14, 6 (2019), e0217038. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6510516/

[11] Mahfuzur Rahman, Mohammad Alatiqi, Mohammad Al-Jarallah, Maryam Yousef Hussain, Abdul Monayem, Prashant Panduranga, and Rajesh Rajan. 2025. Cardiovascular Effects of Smoking and Smoking Cessation: A 2024 Update. *Global Heart* 20, 1 (2025), 15. doi:10.5334/gh.1399 Open Access under CC BY 4.0.

[12] Stephanie M. Rutledge and Amon Asgharpour. 2020. Smoking and Liver Disease. *Gastroenterology & Hepatology* 16, 12 (2020), 617–625. https://www.gastroenterologyandhepatology.net/archives/december-2020/smoking-and-liver-disease/

[13] M. Sinha, P. Haaland, A. Krishnamurthy, et al. 2025. Causal Analysis for Multivariate Integrated Clinical and Environmental Exposures Data. *BMC Medical Informatics and Decision Making* 25 (2025), 27. doi:10.1186/s12911-025-02849-4

[14] Y. Zhu, P. V. Benos, and M. Chikina. 2024. A Hybrid Constrained Continuous Optimization Approach for Optimal Causal Discovery from Biological Data. *Bioinformatics* 40, Supplement 2 (2024), ii87–ii97. doi:10.1093/bioinformatics/btae411

[15] Bjørn O. Åsvold, Trine Bjøro, Tom I. L. Nilsen, and Lars J. Vatten. 2007. Tobacco Smoking and Thyroid Function: A Population-Based Study. *Archives of Internal Medicine* 167, 13 (2007), 1428–1432. doi:10.1001/archinte.167.13.1428