Causal Explanations for Disparate Trends: Where and Why?

Anonymous Authors

ABSTRACT

During data analysis, we are often perplexed by certain *disparities* observed between two groups of interest within a dataset. To better understand an observed disparity, we need *explanations* that can pinpoint the data regions where the disparity is most pronounced, along with its causes, i.e., factors that alleviate or exacerbate the disparity. This task can be complex and tedious, particularly when the dataset is large and high-dimensional, demanding an automatic system for discovering *explanations* (data regions and causes) of an observed disparity in a dataset. When offering explanations for disparities, it is critical that they are not only interpretable but also actionable—enabling users to make informed, data-driven decisions. This requires explanations to go beyond surface-level correlations and instead capture *causal* relationships. We introduce ExDis, a framework for discovering causal Explanations for Disparities between two groups of interest. ExDis identifies data regions (subpopulations) where disparities are most pronounced (or reversed), and associates specific factors that causally contribute to the disparity within each identified data region. We formally define the ExDis framework and the associated optimization problem, analyze its complexity, and develop an efficient algorithm to solve the problem. Through extensive experiments over three real-world datasets, we demonstrate that ExDis generates meaningful causal explanations, outperforms prior methods, and scales effectively to handle large, high-dimensional datasets.

ACM Reference Format:

Anonymous Authors. 2026. Causal Explanations for Disparate Trends: Where and Why?. In *Proceedings of International Conference on Management of Data (SIGMOD '26).* ACM, New York, NY, USA, 19 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Data is the main building block of modern, data-driven decision-making. People rely on the trends observed in the data to gain insights, and in turn, use those insights to draw conclusions or even make important decisions. For large, high-dimensional datasets, certain observed trends often require further drilling down or *explanations*. For example, after observing a *disparate* trend that females are more likely to experience nervous breakdowns and anxiety attacks than non-females, one might wonder: "For which subpopulation is this disparity more pronounced?", "Which factors contribute to this disparity?", "Are there particular countries/races where this trend is reversed?", and so on. Manually searching for these answers is like finding a needle in a haystack, which demands automated ways to pinpoint the subpopulation or data region where a trend is amplified or substantially reversed from the global trend, and identify factors that are connected to these disparate trends.

Table 1: A sample (toy) dataset over a partial schema of the Stack Overflow Annual Developer Survey dataset [1].

Gender	Ethnicity	Education	Role	YrsProfCoding	TC
Non-binary	White	BS	Business analyst	6-8	83K
Male	South Asian	PhD	Data analyst	4-6	124K
Female	South Asian	MS	Back-end developer	2-4	75K
Male	East Asian	BS	Back-end developer	6-8	59K

Understanding *causal reasons* behind disparities in outcomes between two groups is essential for making informed, data-driven decisions to address inequities. For instance, if a policymaker identifies the factors that causally lead to lower average salaries in a specific subpopulation compared to the rest of the population, they can design targeted interventions to mitigate the gap.

In this paper, we propose ExDis for automatically Explaining an observed Disparate trend. We proceed to provide three examples, highlighting unique use cases, to motivate the need to discover *explanations* (data regions and associated causes) of an observed disparate trend between two groups of interest within a dataset.

Example 1.1 (Investigating a disparate trend). *A social analyst Miro, is examining tech workers' total-compensation data, such as the Developer Survey by Stack Overflow dataset [1], a sample of which is shown in Table 1. The dataset contains information about individuals' demographics (gender, ethnicity, age, etc.), role, experience in coding professionally, their own and their parents' education, total compensation (TC), etc. Contrary to the common knowledge, Miro observes that the average TC of data or business analysts ($106K) is about 10% higher than the average TC of back-end developers ($96K)—a surprising trend!*

Miro wants to identify large subpopulations that contribute significantly to this observed trend, and uncover causes behind it. After some digging, he discovers that one of the larger subpopulations is White individuals aged between 25–34, which constitutes 35% of the survey population. Within this subpopulation, on average, analysts ($115K) earn 9.5% more than developers ($105K). Miro further discovers that having worked as a professional coder is a major contributing factor to this disparity, particularly for this subpopulation. Specifically, among White individuals aged between 25–34, having 6–8 years of professional coding experience causes a TC increase of $44K for `analysts`*, and only $10K for* `back-end developers`*, further exacerbating the TC-gap.*

After this discovery, Miro wonders if there are other significant data regions with similar properties. What are the major causes of the disparity in those regions? Unfortunately, the dataset contains a number of multi-valued attributes, which makes manual exploration of all possible data regions an infeasible option. Furthermore, exploring all possible factors that can cause a significant disparity in the target attribute (TC) requires a more involved search. This makes it impossible for Miro to find explanations of the disparity manually.

SIGMOD '26, May 31–June 5, 2026, Bengaluru, India

ACM ISBN 978-1-4503-XXXX-X/2018/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

Example 1.2 (Debugging bias). *While analyzing a health insurance coverage dataset [64], Soha observes a disparate trend that individuals with an occupation that involves manual labor have a 13% lower chance (78%) of being covered by health insurance than the overall population across all occupations (91%). This indicates a "blue-collar bias" [21] and Soha wishes to discover data regions where this bias is more pronounced and uncover why. Turns out that among non-natives, which make up 16% of the population, the disparity is even more severe. Within this subpopulation, manual-labor workers have a 65% chance of being insured, which is 19% lower than the 84% coverage rate for all occupations. Furthermore, among not-natives, earning between 25K–55K boosts the chance of having health insurance for* `manual-labor workers` *by 2%, where it hurts the chance for* `people with any occupation` *by 1%.*

Example 1.3 (Discovering reverse trends). *Generally, males have a lower likelihood (37%) of feeling nervous frequently than non-males (45%). Madison is investigating a Medical Expenditure Panel Survey dataset [2] and she is interested in finding subpopulations where a reverse trend exists, i.e., males have a higher likelihood of feeling nervous than non-males (this phenomenon is known as Simpson's Paradox [68]). One such subpopulation is divorced people with age between 51–63 who have a recommendation to exercise from their doctor, where males have a higher likelihood (47%) of feeling nervous than non-males (43%). Interestingly, within this subpopulation, not currently smoking exacerbates the situation for* `males` *(increases the likelihood of feeling nervous by 21%) but improves the situation for* `non-males` *(decreases the likelihood of feeling nervous by 14%). However, to discover such a reverse trend, they need to examine all possible subpopulations and try out all possible treatments within each subpopulation, which is tedious.*

The above examples motivate investigating disparities in an aggregated *outcome variable* (e.g., TC) between *two (possibly overlapping) groups* of interest (e.g., analysts vs back-end developers), aiming to identify (1) *where* the disparities are most pronounced (or reversed), such as a specific data region or subpopulation and (2) *why*, i.e., what factors further alleviate/exacerbate the disparity.

Desiderata. There are three key desiderata for the aforementioned problem. **First,** a single causal explanation rarely accounts for the disparity observed across different subpopulations. Different subpopulations may exhibit disparity for different underlying reasons. Thus, the first goal is to identify *high-utility* subpopulations - groups where a strong and meaningful causal explanation for the disparity exists. **Second,** while small subpopulations may exhibit strong causal explanations, insights derived from such groups are often not generalizable. To ensure broader applicability and avoid misleading insights, chosen subpopulations should exhibit sufficient *support*, i.e., they should cover a reasonable portion of the overall dataset. **Third,** selecting the top-k subpopulations purely based on utility and support may lead to redundant or overlapping results. For example, "principal engineers" and "people with age between 35–45" may comprise the same individuals, as most principal engineers are 35–45 years old. Therefore, beyond finding high-utility and high-support subpopulations, we must minimize the overlap among the reported k subpopulations. This alludes to the notion of *diverse* selection of subpopulations [41, 70, 78].

Problem. Given a database D, an outcome variable O, causal background knowledge in the form of a causal DAG $\mathcal{G}_D$ by Pearl's graphical causal model [44], two groups of interest g_1 and g_2, and a parameter k, our goal is to generate a set of k *disparity explanations* that, collectively, best explain the disparities between g_1 and g_2 w.r.t the aggregated outcome variable O, according to $\mathcal{G}_D$. In this work, we consider `AVERAGE` as the aggregation function since it satisfies the requirements for causal analysis (more details are in Section 4.1). Each explanation consists of two components: (1) a *subpopulation* where the disparity is pronounced (or reversed), and (2) a *treatment pattern* that causally affects g_1 and g_2 disparately, within that subpopulation. The quality of a set of disparity explanations is primarily determined by the causal strength of the treatment patterns they reveal, measured by the *Average Treatment Effect (ATE)* [44]. This forms our main optimization objective: to identify a set of subpopulations for which the associated treatments strongly explain the observed disparity. In addition to maximizing causal explainability, two important constraints guide the selection process: (1) Each explanation must have sufficient *support*—i.e., the subpopulation it describes should cover a sizable fraction of the data, ensuring representativeness and generalizability. (2) The selected subpopulations must exhibit low *overlap*, encouraging *diversity* in the explanation set. This is quantified using the Jaccard similarity between subpopulations, which helps avoid redundant or overly similar explanations.

Challenges and limitations of prior work. The key challenge lies in identifying subpopulations that are associated with strong causal explanations for observed disparities, while simultaneously satisfying constraints on support and diversity. Prior work has typically focused on selecting the top-k subpopulations based on observed disparities or coverage [3, 43], often neglecting causal explainability or redundancy. Moreover, unlike approaches that rely solely on observed outcome disparities [61], we emphasize the importance of discovering subpopulations where the disparity is causally explained by specific treatments. In summary, we extend prior work in two key directions: (1) We address a more difficult variant of the problem by optimizing over a set of k subpopulations under support and diversity constraints, and (2) We focus on identifying high-quality *causal* explanations, not only data regions with high disparities. Recent works [36, 75] have used causal inference to explain aggregate query results and disparities between two groups within it. However, they do not support overlapping groups, which is essential for bias debugging as demonstrated in Example 1.2. Moreover, we aim to identify causal explanations within different subpopulations rather than providing a single explanation for the entire data or query outcome. Finally, operating on the entire data, rather than the aggregate view, poses another challenge as the search space becomes significantly larger (we explain this in Section 5.2).

Contributions. We make the following contributions:

(1) We **formalize the problem** of generating a set of causal explanations to account for disparities in outcomes between two (possibly overlapping) groups of interest. Specifically, we define an optimization problem that seeks to *maximize the causal utility* of the selected explanations—i.e., their ability to causally explain the observed disparity, subject to three constraints: (1) a bound on the number of explanations (size k), (2) minimum support for each explanation to ensure representativeness, and (3)

limited overlap between subpopulations to reduce redundancy. We show that this problem is NP-hard (Section 4).

(2) We **develop the ExDis framework**, which operates in three steps. First, ExDis finds candidate subpopulations. Then it identifies local explanations for each candidate subpopulation by adapting a previous work on finding treatments with substantial causal effect [75]. Finally, it uses an effective greedy strategy to find a k-sized explanation set (Section 5).

(3) We present a thorough **empirical analysis** over 3 real-world datasets and present **3 case studies** that include 5 baselines, and 4 variations of our approach as additional comparison points. We show that ExDis generates higher quality explanations than the existing approaches and can find alternative explanations that existing approaches miss. We also find ExDis scalable and efficient in practice, with its runtime being linear w.r.t the number of data tuples (Section 6).

We review related work in Section 2, provide background on causal inference in Section 3, and discuss our limitations and directions for future work in Section 7.

2 RELATED WORK

Identifying interesting subgroups in high-dimensional data. Prior work exists to identify the most intriguing data subsets for exploration [5, 9, 19, 24, 33, 40, 43, 53–56, 74]. Other studies have focused on uncovering compelling data visualizations [66, 79], or pinpointing data subsets where models underperform [12, 43, 61]. A key part of our objective is identifying subpopulations where there is a significant disparity in the average outcomes between two groups of interest.

DivExplorer [43] is designed to analyze the behavior of classification models, with the primary goal of identifying data regions where a performance metric (e.g., false positive or false negative rates) deviates significantly from its value over the entire dataset. FairDebugger [61] aims to identify data subsets responsible for fairness violations in the outcomes of a random forest classifier. It pinpoints the most impactful data samples that significantly influence the model's predictions. However, unlike ExDis, both of these systems focus solely on detecting data regions with unexpected behavior, without uncovering the underlying *causal* factors driving the observed trends. We empirically compare our approach against these baselines in Section 6.

Query results explanation. Extensive research has been devoted to explaining the results of aggregate SQL queries. Multiple works have leveraged *data provenance* to generate explanations for query results [7, 10, 13, 30, 32, 37, 38, 63]. Other methods have explored non-causal interventions [8, 14, 15, 47, 48, 62, 72], entropy-based techniques [16], and identifying counterbalancing patterns [39]. This line of work is different than ours, as our goal is to explain the disparity among two, possibly overleaping, groups of interest via a small set of causal explanations.

Recent works [36, 52, 75, 76] have used causal inference to explain aggregate query results. Prior work [52, 76] proposed methods to find confounding variables that explain the correlation relationship between the grouping attribute and the outcome in group-by-average queries. In both the same explanation is provided for all groups in the query results. CauSumX [75] focuses on providing causal explanations for group-by-average queries, aiming to explain the overall aggregate view by identifying factors influencing the outcome within each group in the query results. In contrast, our objective is to identify subpopulations within the data where the disparity between two groups is most pronounced and to offer specific causal explanations for the observed disparity within these subpopulations. Moreover, unlike CauSumX, which derives explanations from the aggregate view, we analyze the entire dataset, which is typically much larger. As a result, we introduce multiple optimizations to improve runtime efficiency. While our goals differ, we adapt their treatment mining algorithm for our approach (Section 5.2).

XInsight [36] identifies both causal and non-causal patterns to explain disparities between two groups in aggregate queries. In contrast, our approach supports overlapping groups and identifies specific causal explanations within different subpopulations rather than seeking a single treatment for the entire dataset. We argue that, in many cases, no universal explanation suffices, and disparities are better understood through localized causal insights. This observation is supported by our experiments (Section 6.2).

Rule mining. Association rule mining is a widely studied problem [22, 28], which aims to identify frequent relationships in datasets. Rule-based interpretable models utilize these techniques to derive predictive rules, aiming to balance accuracy and interpretability [27, 35, 51, 57]. Recent works have explored rule generation from causal relationships [45, 46, 60] and heterogeneous treatment effect estimation [69, 73]. However, they do not address our problem of identifying where disparities are significant and their causes.

3 BACKGROUND ON CAUSAL INFERENCE

We use Pearl's model for *observational causal analysis* [44] and present below a few concepts according to it. The broad goal of *causal inference* is to estimate the effect of a *treatment variable* T on an outcome variable O (e.g., the effect of `YrsProfCoding` on `TC`).

The Average Treatment Effect (ATE) is a commonly used measure in causal analysis, quantifying the difference in expected outcomes between treated and untreated groups [44, 50]. ATE conceptually assumes a scenario where treatment is assigned randomly. However, in observational data, treatment is not assigned randomly, and *confounding variables* that influence both treatment and outcome must be accounted for. To estimate ATE for a binary treatment T on an outcome O, the following definition is used:

$$ATE(T, O) = \mathbb{E}_Z\left[\mathbb{E}[O \mid T = 1, \mathbf{Z}] - \mathbb{E}[O \mid T = 0, \mathbf{Z}]\right]$$

Here, $\mathbf{Z}$ represents the set of confounding variables that affect both the treatment T and the outcome O, ensuring that the causal effect is isolated from confounding influences.

Example 3.1. *Suppose that we want to estimate the causal effect of* `YrsProfCoding` *on* `TC` *from the Stack Overflow (SO) dataset. Since the values of* `YrsProfCoding` *were not assigned at random, and having more or fewer years of professional coding experience and obtaining a high total compensation may depend on other attributes like* `Ethnicity`, `Education`, *and* `Role`, *we must control for these confounding variables when estimating the causal effect of* `YrsProfCoding` *on* `TC`.

Pearl's model provides ways to account for these confounders $\mathbf{Z}$ to get an unbiased causal estimate under additional assumptions:

(1) The unconfoundedness assumption states that if we condition on the confounders $\mathbf{Z}$, then the treatment T is independent of the outcome O given $\mathbf{Z}$. Intuitively, it means that after conditioning on $\mathbf{Z}$, the treatment T is as good as randomly assigned. (2) The Overlap assumption ensures that for every combination of confounders, there is a nonzero probability of receiving the treatment, allowing for valid comparisons across groups.

Pearl's model gives a systematic way (e.g., the backdoor criterion [44]) to find a sufficient set of confounding variables $\mathbf{Z}$ when a *causal DAG* is available. Causal DAGs are graphical models that provide a simple way to graphically represent causal relationships among variables. A causal DAG is a specific type of Bayesian network, where nodes represent random variables (i.e., data attributes) and edges signify potential direct causal influence.

Example 3.2. *Figure 1 depicts a partial causal DAG for the SO dataset over a subset of attributes in Table 1 as variables. Given this causal DAG, we can observe that* `YrsProfCoding` *depends on the values of an individual's* `Role`, `Ethnicity`, *and* `Education`.

A causal DAG can be constructed by a domain expert as in the above example, or using existing *causal discovery* algorithms [20]. In this work, we assume the causal DAG is given as part of the input, which is a common assumption in prior work [17, 31, 75].

In our framework for providing causal explanations for disparities between two groups, we focus on estimating the causal effect of a treatment T on an outcome O within a specific subpopulation, characterized by a *pattern* ψ. (We define the set of patterns considered in Section 4.) Consequently, our goal is to compute the Conditional Average Treatment Effect (CATE) [23, 49] rather than the ATE, as CATE captures the treatment effect within a targeted subpopulation. To estimate the CATE for a binary treatment T on an outcome O for a subpopulation ψ, we use the following definition:

$$CATE(T, O|\psi) = \mathbb{E}_Z\left[\mathbb{E}[O \mid T = 1, \mathbf{Z}, \psi] - \mathbb{E}[O \mid T = 0, \mathbf{Z}, \psi]\right]$$

where $\mathbf{Z}$ represents a sufficient set of confounding variables.

4 DISCOVERING DISPARITY EXPLANATIONS

We consider a single-relation database over a schema $\mathbb{A}$. The schema is a vector of attributes $\mathbb{A}=(A_1, \ldots, A_m)$, where each A_i is associated with a domain $\mathsf{dom}(A_i)$. We have a categorical or continuous outcome attribute $O \in \mathbb{A}$. We also assume all other attributes in $\mathbb{A}$ are categorical (when an attribute is not categorical, we can discretize them to make them categorical). A database instance D populates the schema with a set of tuples $t=(a_1, \ldots, a_m)$ where $a_i \in \mathsf{dom}(A_i)$. We use bold letters to represent a subset of attributes $\mathbf{A} \subseteq \mathbb{A}$.

To specify a *subpopulation* (subset of tuples) from the database D, we use *patterns* [16, 34, 48, 72] that comprise conjunctive *predicates* on attribute values. Patterns are commonly used in the literature of query results explanations [32, 72, 75].

Definition 4.1 (Pattern). *Given a database instance D over schema $\mathbb{A}$, a simple predicate φ is an expression of the form A_i op a_i, where $A_i \in \mathbb{A}$, $a_i \in \mathsf{dom}(A_i)$, and* op $\in \{=, \neq\}$. *A* pattern *ψ is a conjunction of simple predicates, i.e., $\psi = \varphi_1 \wedge \ldots \wedge \varphi_k$. We use $\psi(D) \subseteq D$ to denote the subpopulation within D defined by ψ.*

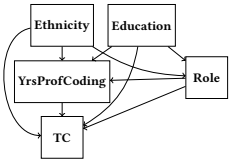


Figure 1: Partial causal DAG for the Stack Overflow dataset.

Example 4.1. *Examples of two simple predicates in the SO dataset (Table 1) are* `Ethnicity = Asian` *and* `Role = Data analyst`. *An example of a pattern is* `Ethnicity = Asian` $\wedge$ `Role = Data analyst`.

The groups of interest, g_1 and g_2, are defined by the patterns ψ_{g_1} and ψ_{g_2} respectively. In this work, we only consider equality or inequality predicates, in line with previous work on explanations that deem such predicates intuitive and understandable [3, 16, 47].

4.1 Disparity Explanations

Given a database D defined over a schema $\mathbb{A}$ and an outcome attribute $O \in \mathbb{A}$, we aim to discover explanations for an observed disparity in average(O) between g_1 and g_2. Our building blocks are *disparity explanations* that identify *where* the average outcomes for g_1 and g_2 differ significantly and *why*.

Example 4.2. *An analyst over the SO dataset (Table 1) is interested in finding explanations of a surprising disparate observation: the average* `TC` *of data or business analysts (*\$106k*) is* \$10k *higher than the average* `TC` *of back-end developers (*\$96k*).*

We focus on comparing the average outcomes of two groups. Our framework is designed to uncover *causal explanations* for differences in these averages, leveraging the concept of CATE as discussed in Section 3. Since CATE inherently relies on *expectations* (weighted averages), it is particularly suited for analyzing aggregate averages of outcomes. Aggregate functions such as `SUM` or `COUNT`, on the other hand, depend on the *number of tuples* in the data, which does not directly align with causal effect estimates. While non-causal approaches to explanations [29, 32, 39, 48, 71, 72] can support a variety of aggregate functions, methods that are based on causal estimates typically focus on averages [31, 36, 52, 75, 76].

Mutable and immutable attributes. We assume the attributes set $\mathbb{A} \setminus \{O\}$ is partitioned into two disjoint sets: *mutable* attributes that can be used to define what affects the outcome (e.g., years of coding professionally, education) and *immutable* attributes, which are inherent and cannot be changed (e.g., ethnicity, gender). We use immutable attributes to define the subpopulations. Formally, let $\mathbf{I} \subset \mathbb{A}$, denote the set of immutable attributes and $\mathbf{M} \subset \mathbb{A}$ denote the set of mutable attributes, where $\mathbf{M} \cap \mathbf{I} = \emptyset$ and the outcome $O \notin \mathbf{M} \cup \mathbf{I}$. This categorization makes sure that explanations consist solely of mutable attributes that can imply corrective measures to reduce the disparity among the groups. We assume that a domain expert provides this categorization of attributes. This is similar to prior work on counterfactual explanations, where certain attributes are excluded in the explanation as they are non-actionable [18, 25, 26].

Definition 4.2 (disparity explanation). *Given a database instance D over a schema $\mathbb{A}$, an outcome variable $O \in \mathbb{A}$ and two groups of interest g_1 and g_2, a disparity explanation ϕ is defined as a pair of patterns (ψ_g, ψ_e) where:*

(1) ψ_g is defined by attributes in $\mathbf{I} \subset \mathbb{A}$, highlighting a subpopulation with significant disparity between g_1 and g_2 in terms of AVG*(O).*
(2) ψ_e is defined by attributes in $\mathbf{M} \subset \mathbb{A}$, indicating a treatment that can explain the disparity between g_1 and g_2 within $\psi_g(D)$.

To assess the impact of the treatment ψ_e on the outcome O within the subpopulation $\psi_g(D)$, we compare the causal effect of ψ_e on O within the two subpopulations: $(\psi_g \wedge \psi_{g_1})(D)$ and $(\psi_g \wedge \psi_{g_2})(D)$.

Example 4.3. *Continuing with our running example. where g_1 is* `analysts` *and g_2 is* `back-end developers`*, an example disparity explanation is: Among white individuals aged between 25–34, having 6-8 years of professional coding experience boosts the* TC *for* `analysts` *more than* `back-end developers`*. Here, the subpopulation pattern ψ_g is defined by* `Ethnicity = White` $\wedge$ `Age = 25 − 34` *and the treatment pattern ψ_e is* `YrsProfCoding = 6–8`*. Within this subpopulation, the average* TC *for* `analysts` *and* `back-end developers` *are $115K and $105K, respectively (gap is $10K).*

4.2 Problem Formulation

We now formally define the problem of finding disparity explanations. We assume that we are given a database instance D with schema $\mathbb{A}$, a causal model $\mathcal{G}_D$ on $\mathbb{A}$, outcome $O \in \mathbb{A}$, and two groups g_1 and g_2 defined by the patterns ψ_{g_1} and ψ_{g_2}. Let $\{\phi_1, \phi_2, \ldots, \phi_l\}$ be a set of possible disparity explanations of size l. Our goal is to find a bounded-sized set of disparity explanations Φ to identify subsets of the data that (1) provide insights into the disparity between g_1 and g_2, and (2) avoid redundancy across different subsets to cover different data regions.

Usefulness of an explanation. A disparity explanation $\phi = (\psi_g, \psi_e)$ is considered useful if (i) it reveals a significantly different effect of the treatment ψ_e on the outcome O for the subpopulation $\psi_g(D)$ between g_1 and g_2, and (ii) it constitutes a significant portion of the data. To this end, we define the *disparity score*, which measures the magnitude of the disparity, and the *support* of a disparity explanation that allows us to eliminate disparity explanations that constitute only minor portions of the data.

The disparity score Δ of a disparity explanation $\phi = (\psi_g, \psi_e)$ measures the absolute difference between the two CATE values: one computed over the subpopulation $(\psi_g \wedge \psi_{g_1})(D)$ and the other over $(\psi_g \wedge \psi_{g_2})(D)$. The difference is normalized by the maximal outcome value. Formally,

$$\Delta(\phi) = \frac{\left|CATE_{\mathcal{G}_D}(\psi_e, O|\psi_g \wedge \psi_{g_1}) - CATE_{\mathcal{G}_D}(\psi_e, O|\psi_g \wedge \psi_{g_2})\right|}{\max\{|o| \mid \in O\}}$$

In order to prioritize disparity explanations that cover a large portion of the given database, we use the notion of *support*. The support of a disparity explanation $\phi = (\psi_g, \psi_e)$ is defined by the fraction of tuples $\in D$ that take part in the explanation, namely, tuples that satisfy the patterns in the disparity explanation. Formally,

$$support(\phi) = \frac{|\psi_{g \wedge g_1}(D) \cup \psi_{g \wedge g_2}(D)|}{|D|}$$

Intuitively, the higher the support of a disparity explanation, the more interesting it is, as it applies to a larger portion of the population. We prefer disparity explanations with high support.

Example 4.4. *Continuing from Example 4.3, the support for the disparity explanation, over the subpopulation White individuals aged between 25 to 34—working as either analysts or back-end developers—is $\frac{16{,}508}{47{,}702} = 34.6\%$. The disparity score for the corresponding explanation—with the identified cause of having 6 to 8 years of professional coding experience—is $\frac{|44{,}058 - 10{,}552|}{2{,}000{,}000} = 0.016$. Note that this is an actual disparity explanation that we found during our empirical analysis over the SO dataset (Table 4, row 5).*

Diversity among the disparity explanations. We are interested in a diverse set of disparity explanations to reveal and explain the difference in outcome for the two groups of interest. Given two groups of interest g_1 and g_2, we use $D_{g_1 \cup g_2}$ to denote the subset of D containing tuples that belong to at least one of the groups. More formally, $D_{g_1 \cup g_2} = \psi_{g_1}(D) \cup \psi_{g_2}(D)$. Given two disparity explanations $\phi = (\psi_g, \psi_e)$ and $\phi' = (\psi_{g'}, \psi_{e'})$, defined over subpopulations g and g', respectively, and the same outcome variable O, we use the Jaccard similarity between $\psi_g(D_{g_1 \cup g_2})$ and $\psi_{g'}(D_{g_1 \cup g_2})$ to measure the similarity between ϕ and ϕ'. Formally:

$$\textsc{sim}(\phi, \phi') = \frac{|\psi_g(D_{g_1 \cup g_2}) \cap \psi_{g'}(D_{g_1 \cup g_2})|}{|\psi_g(D_{g_1 \cup g_2}) \cup \psi_{g'}(D_{g_1 \cup g_2})|}$$

We are now ready to formally define the problem of selecting disparity explanations. At a high level, our goal is to select a bounded-sized diverse set of disparity explanations with support above a given threshold, such that their combined disparity score is maximized, with bounded pairwise similarity to reduce redundancy.

Problem 1 (Disparity Explanation Selection). *Given a database instance D with schema $\mathbb{A}$, a causal model $\mathcal{G}_D$ on $\mathbb{A}$, outcome $O \in \mathbb{A}$, two groups of interest g_1 and g_2, a set of possible disparity explanations $\{\phi_1, \phi_2, \ldots, \phi_l\}$, a fixed budget $k \in \mathbb{N}^+$, a support threshold σ and a similarity threshold τ, select a disparity explanation set $\Phi \subseteq \{\phi_1, \phi_2, \ldots, \phi_l\}$, such that:*

*(1) **(size constraint)** $|\Phi| \leq k$*
*(2) **(support constraint)** $\forall \phi_i \in \Phi,\ support(\phi) \geq \sigma$*
*(3) **(diversity constraints)** $\forall \phi_i, \phi_j \in \Phi,\ \textsc{sim}(\phi_i, \phi_j) \leq \tau$*
*(4) **(objective)** $\Delta(\Phi) = \sum_{\phi \in \Phi} \Delta(\phi)$ is maximized*

We show that Problem 1 is intractable: given a bound B over the minimal value of the objective function, the corresponding decision problem is NP-Hard.

Proposition 4.1. *Given a set of candidate disparity explanations Φ_c, a budget k, a support threshold σ, a similarity threshold τ, and a bound B, determining whether $\exists \Phi \subseteq \Phi_c$ s.t $|\Phi| \leq k$, $\forall \phi_i \in \Phi,\ support(\phi) \geq \sigma$, $\forall \phi_i, \phi_j \in \Phi,\ \textsc{sim}(\phi_i, \phi_j) \leq \tau$ and $\sum_{\phi \in \Phi} \Delta(\phi) \geq B$ is NP-hard.*

The proof is based on a reduction from the Independent Set (IS) problem and is given in the Appendix.

5 THE EXDIS ALGORITHM

Since the number of possible disparity explanations may be exponential in the number of attributes and their domain values, considering all possible disparity explanations is inefficient. ExDis avoids this by generating a set of potentially promising ones. Then it applies a greedy search to select the top-k with the highest disparity scores, adhering to the support and diversity constraints.

The ExDis framework comprises three steps: (1) the *subpopulation miner*, which generates patterns to identify subpopulations with sufficient support; (2) the *explanation miner*, which uncovers causal explanations for each candidate subpopulation; and (3) the *greedy search*, which efficiently selects the top-k disparity explanations using a greedy strategy. We build upon and adapt existing methods (e.g.,[4, 75]) where applicable, and introduce novel techniques when necessary.

5.1 Subpopulation Miner

Our first goal is to identify candidate data regions (subpopulations) where the disparity between groups g_1 and g_2 is significant and supported by sufficient data. To do this, we adapt the first step of the classic Apriori algorithm [4] to find all subpopulations with support above a given threshold, which requires only a single pass over the data. This step is restricted to immutable attributes **I**, ensuring that the generated candidate subpopulations are based solely on the immutable attributes.

As noted in Section 1, the ExDis framework is versatile and can be applied to various use cases, such as investigating surprising observations, debugging fairness issues, or identifying reverse trends. In each scenario, one may search for specific subpopulations where the average outcome for g_1 is either higher or lower than that for g_2. To accommodate this, we introduce a filtering step which returns only subpopulations that meet the relevant condition.

5.2 Explanation Miner

The next step is to explain the disparity within each candidate subpopulation found in the previous step. Given a subpopulation pattern ψ_g, an outcome variable O, and two groups g_1 and g_2, our goal is to identify the treatment pattern ψ_e with the highest disparity score. To this end, we adapt the treatment mining step of CauSumX [75] to our setting. CauSumX provides causal explanations for the results of aggregate queries. An explanation pattern consists of a set of tuples from the aggregate view (i.e., the query results), and a treatment pattern is used to quantify the causal effect (in terms of the CATE value) of the treatment on the outcome within the relevant subview. CauSumX employs a heuristic lattice traversal approach to identify promising treatment patterns with high CATE values. However, unlike CauSumX, which estimates the CATE value, we estimate the disparity score of the treatment pattern under consideration. Furthermore, we adjust the search to the context, e.g., when exploring parts of the data where the average outcome for g_1 exceeds that of g_2, the explanation should elucidate this phenomenon—specifically, we aim to identify treatments that favor g_1 over g_2, and filter out the others.

We note that CauSumX operates only on aggregate views (i.e., the results of aggregate queries), where considering a relatively small number of grouping patterns is sufficient (in [75], the average was 24). In contrast, our setting requires searching for subpopulations with significant disparity across the entire dataset, as we aim to identify a treatment for each subpopulation. This requires consideration of a much larger number of potential grouping patterns (in our experiments, the average was 184). As a result, this module is the primary bottleneck of our system. To mitigate this, we introduce the following optimizations to improve runtime:

Limiting patterns. To ensure conciseness (and thus interpretability) of explanations, while reducing runtime, we restrict the search for treatment patterns to at most two predicates (similar to [3]).

Parallelization. Since the treatment patterns for the subpopulations can be identified independently, we leverage parallelism. Additionally, within a subpopulation, we compute the disparity score of each treatment pattern in parallel to reduce runtime.

Caching. Computing the disparity score for a given treatment pattern requires adding it as a node to the underlying causal DAG [75]. Often, this may lead to a DAG that has been previously encountered. To avoid redundant computations, we cache results related to previously encountered causal DAGs.

Sampling. As was done in [75], instead of focusing on obtaining precise CATE values, we estimate CATE from a random sample of the data. We use a fixed sample size of 500,000 tuples, guided by our empirical findings, which indicates that this sampling size achieves highly accurate CATE estimations while maintaining a relatively low runtime. However, the sampling ratio is a customizable system parameter and the user is free to tune it for more accurate results.

5.3 Greedy Search

Given the set of candidate disparity explanations $\{\phi_i\}_{i=1}^{l}$ obtained in the previous two steps, our goal is to identify the top-k disparity explanations with the highest disparity scores, adhering to the constraints (as defined in Problem 1).

To achieve this, we first cluster the candidate explanations using a hierarchical clustering algorithm (using Scipy [67] implementation) based on the symmetric difference among the subpopulations as the similarity measure. Then, from each cluster, we select a random representative explanation and assign its disparity score to the entire cluster. We then iteratively select k disparity explanations. At the first iteration, we pick a random explanation from the cluster with the highest disparity score. At the $j{-}th$ iteration (for $1 < j \leq k$), we select the explanation ϕ^* such that:

$$\phi^* = \underset{\phi \in \Phi \wedge \text{SIM}(\phi,\phi') < \tau}{\arg\max} \Delta(\phi), \quad \text{for } \phi' \in \Phi_{j-1}.$$

where Φ is the set of candidate explanations consists of a random explanation from each cluster, and Φ_{j-1} is the set of explanations selected up to iteration j.

Complexity analysis. The maximum number of disparity explanations in a database D with attributes $\mathbb{A}$ is bounded by $|D|^{|\mathbb{A}|}$ (considering both subpopulation and treatment patterns), which is polynomial in terms of data complexity, assuming a fixed schema [65]. Our greedy search is also polynomial in the number of explanations considered. Additional operations, such as calculating CATE values, are polynomial in D, leading to worst-case polynomial data complexity. As we demonstrate in Section 6.4, ExDis is capable of efficiently handling large, high-dimensional datasets in practice.

6 EXPERIMENTAL EVALUATION

In this section, we present an experimental evaluation of the effectiveness of ExDis in practical settings. We aim to address the following questions:

- **Q1**: How does the quality of ExDis-generated disparity explanations compare to that of existing methods? (Section 6.2)

Table 2: Details of the datasets for experiments and case studies.

Dataset	#Tuples	$\|\mathbf{I}\|$	$\|\mathbf{M}\|$	g_1	g_2	$\|g_1\|$	$\|g_2\|$	O	AVG_{g_1}	AVG_{g_2}
Stack Overflow (SO) [1]	47,702	4	6	Data/business analysts	Back-end developers	4,088	28,987	Total Compensation (TC)	$106K	$96K
American Community (ACS) [64]	1,420,652	15	10	Manual labor	Overall data	99,790	1,420,652	Likelihood of having a health insurance	78.6%	91.5%
Medical Expenditure Panel (MEPS) [2]	13,528	7	5	Males	Non-males	5,731	7,797	Likelihood of feeling nervous frequently	41.6%	46.9%

- **Q2**: How is the quality of ExDis-generated explanations affected by various parameters and how to tune the system parameters? (Section 6.3)
- **Q3**: How efficient and scalable is ExDis? (Section 6.4)
- **Q4**: How do our proposed optimizations affect ExDis' runtime performance? (Section 6.5)

6.1 Experimental Setup

All experiments were performed on a Windows computer, Intel CPU, with 16 GB memory. We implemented ExDis in Python 3 and used DoWhy library [58] to compute the CATE values. Our source code is publicly available [6].

6.1.1 Datasets, causal DAGs, and preprocessing. We used three popular datasets (Table 2) and obtained the corresponding causal DAGs given in prior work [77]. To process continuous numerical attributes, we applied equal-width binning across 10 bins.

SO: The Stack Overflow Developer Survey [1] dataset contains responses from developers worldwide, covering topics such as professional experience, education, technologies used, and employment-related information, such as annual total compensation (TC).

ACS: The American Community Survey (ACS) [64] is a nationwide survey conducted by the U.S. Census Bureau, with demographic, social, economic, and housing data. We focused on 7 states: California, Texas, Florida, New York, Pennsylvania, Illinois, and Ohio.

MEPS: The Medical Expenditure Panel Survey (MEPS) [2] dataset provides information on healthcare utilization, expenditures, insurance coverage, and demographics of individuals in the U.S.

6.1.2 System parameters. Unless otherwise specified, we used the following default parameters:

- The desired size of the explanation set $k = 5$.
- The minimum support threshold for the support constraint: $\sigma = 0.05$. This means that we consider only groups that account for at least 5% of the data.
- The maximum similarity threshold for diversity constraint: $\tau = 0.55$, using Jaccard similarity.
- The number of clusters for the greedy search phase = 10.

6.1.3 Use cases. Throughout our experimental evaluation, we examine three scenarios that represent different use cases of ExDis (as mentioned in Section 1). The description of the groups, the outcome variables, and relevant statistics are given in Table 2.

(1) Investigating a surprising fact. In tech, developers generally earn more than analysts. However, an analysis of the SO dataset revealed a surprising trend: data or business analysts (g_1) earn more on average than backend developers (g_2). To analyze this, we fix the outcome O as total compensation (TC). We aim to identify which subpopulation significantly contribute to the observed disparity and the underlying explanations in terms of treatments that favor g_1 over g_2.

(2) Fairness debugging. In the ACS dataset, we observe people in certain type of occupation to have lower than average rate of health-insurance coverage. Specifically, we fix the outcome O to indicate whether an individual holds a health insurance; g_1 represents individuals employed in manual-labor occupations (cleaning, maintenance, farming, fishing, construction, etc.), where the health insurance coverage rate is only 78.6%; and g_2 represents the entire dataset across all occupations, with a coverage rate of 91.5%. Our objective is to investigate the underlying factors contributing to this discrepancy by identifying subpopulations for which the average insurance coverage rate of g_1 is significantly lower than that of g_2, along with a causal explanation specific to each subpopulation.

(3) Finding reverse trends. We investigate an intriguing observation in the MEPS dataset: while typically males (g_1) have a lower likelihood of feeling nervous frequently than non-males (g_2), our analysis reveals subpopulations where a reverse trend holds. To dig deeper, we fix the outcome O to indicate whether an individual feels nervous frequently. Our goal is to pinpoint subpopulations where the relative trend involving g_1 and g_2 is reverse compared to the global trend and explore the underlying causes.

6.1.4 Baselines. We consider the following baselines:

Brute Force. We employ an exhaustive Brute Force algorithm, which considers all possible k-sized disparity explanation sets, and returns the best one according to Problem 1. Note that in the absence of an absolute ground-truth, the results obtained from this Brute Force technique can be treated as ground truths for Problem 1.

Top-k. This baseline ignores the diversity constraint and simply returns the top-k explanations ranked by their disparity scores.

XInsight. XInsight [36] is designed to identify both causal and non-causal patterns to explain disparities between two groups in aggregate SQL queries. Unlike ExDis, which provides local (possibly different) explanations for each subpopulation, XInsight provides a single global explanation for the entire data. Since XInsight includes a causal discovery phase, we ensure a fair comparison as follows: for each treatment pattern (i.e., explanation) identified by ExDis, we report its causal effect over the entire dataset (in the "Global" column in Tables 4–6). We aim to empirically demonstrate that subpopulation-specific explanations may not be valid globally.

DivExplorer. DivExplorer [43] is designed to analyze the behavior of classification models with the primary goal of identifying data regions where a performance metric (e.g., false positive/negative rates) deviates significantly from its value over the entire dataset. Given a divergence metric, DivExplorer identifies subsets of the data, defined by patterns, where the metric shows a significant disparity compared to the overall population. We use DivExplorer as a baseline, setting the divergence function to the disparity score. To adapt DivExplorer to our setting, we use a fixed treatment across all the subpopulations. This treatment is selected as the one that yields the highest disparity score between the two groups of interest in the

Table 3: Overall disparity scores (Δ), runtimes, and diversity visualizations for explanations generated by various baselines and ExDis across three use cases. We report the disparity scores w.r.t the Brute Force baseline since it gives the ground truths. In the heat-map, the diagonal is black, denoting 0 self distance from an explanation to itself. Lighter colors denote less similar pair of explanations, offering diversity.

Approach \ Dataset	SO				ACS				MEPS			
	Δ (%)	Runtime (s)	#Explanations	Diversity	Δ (%)	Runtime (s)	#Explanations	Diversity	Δ (%)	Runtime (s)	#Explanations	Diversity
Brute Force	100	181	5		100	3130	5		100	19	5	
Top-K	118	180	5		121	1514	5		166	14	5	
DivExplorer	N/A	33	0	N/A	86	1055	5		N/A	4	0	N/A
FairDebugger	N/A	2258	0	N/A	26	717	2		N/A	103	0	N/A
ExDis (this paper)	55	62	5		84	1170	5		100	17	5	

overall data. We then compute the disparity scores for each subpopulation and return the top-k subpopulations with the highest scores.

FairDebugger. FairDebugger [61] identifies training data subsets that contribute to fairness violations in random forest models by evaluating how their removal affects model outcomes. To adapt FairDebugger to our setting, we fix the treatment pattern to the one exhibiting most disparity across the entire dataset. We then assess the influence of each subpopulation by measuring the change in disparity after removing it. The difference in disparity scores quantifies the subpopulation's contribution to the overall disparity. FairDebugger returns the top-k highest contributing subpopulations.

6.2 Explanation Quality

Table 3 shows a quantitative comparison contrasting the explanations generated by ExDis with those of the baselines over three use cases. Details of the explanations generated by each of these approaches are in Section B.

6.2.1 Investigating a surprising fact (SO). The explanations generated by ExDis are shown in Table 4. Notably, ExDis identifies subpopulations where the average salary of analysts is higher than that of back-end developers, oftentimes with this disparity being more pronounced than in the overall population (where the average salary for analysts is $106,542 and for back-end developers it is $96,609). For almost all cases, ExDis provides a different causal explanation to highlight the factor contributing to the disparity within the specific subpopulation. In all five disparity explanations, the causal effect of the chosen treatment on the entire population is not statistically significant as shown in the "Global" column, emphasizing the importance of providing "local" explanations for each subpopulation. *This observation highlights the distinction between our approach and XInsight (that provides a global explanation, as shown in the Global column), demonstrating that subpopulation-level explanations differ from those provided at the entire population level.*

In this scenario, ExDis shares only 1 out of 5 explanations with the optimal solution by Brute Force. However, ExDis selects a highly diverse set of explanations, achieving 55% of the optimal disparity score produced by Brute Force. In contrast, although the Top-k baseline achieves high disparity scores, the selected explanations are highly similar to one another (as indicated by the heatmap in Table 3), highlighting the importance of incorporating similarity-awareness when selecting explanations.

In this scenario, both DivExplorer and FairDebugger failed to identify any valid solution. This is because they relied on a fixed treatment (as discussed in Section 6.1.4). In every subpopulation they identified, either the average income of backend developers exceeded that of analysts, or the treatment disproportionately benefited backend developers. According to our problem definition, such cases do not qualify as valid solutions.

6.2.2 Fairness debugging (ACS). The disparity explanations generated by ExDis for the second use case are presented in Table 5. ExDis identified subpopulations in which the percentage of manual-labor workers with health insurance was lower than in the overall population. Notably, all explanations (i.e., treatments) involved personal earnings, underscoring the strong causal relationship between income level and the likelihood of having health insurance in the U.S. Interestingly, in this case, a single explanation was sufficient to account for the disparity across all identified subpopulations. However, this explanation behaves differently when applied to the entire dataset. Specifically, for the overall population, not having personal earnings decreases the likelihood of having health insurance across all occupations. In contrast, for each of the reported subpopulations, not having personal earnings actually increases the probability of having health insurance when considering all individuals within those subpopulations. We note that even in the optimal solution found by Brute Force (as well as in the Top-K solution), only two distinct treatments were selected. This suggests that, in this case, the space of relevant treatments is limited, indicating few plausible explanations for the observed disparity.

Compared to the Brute-Force baseline, ExDis recovered 2 out of the 5 disparity explanations while exploring a significantly smaller search space. Top-K overlapped with Brute-Force in just one explanation and exhibited higher similarity among its selected explanations compared to ExDis. Both DivExplorer and FairDebugger identified different subpopulations than ExDis; FairDebugger detected only two (highly overlapping) subpopulations, while DivExplorer found five, though with high similarity among them.

6.2.3 Finding reverse trends (MEPS). Table 6 shows the disparity explanations for MEPS. ExDis identifies subpopulations where the

Table 4: Disparity explanations discovered by ExDis for the SO dataset. Patterns that form a subpopulation are orange, while treatment patterns are blue. We highlight the two groups of interest using yellow and pink. The first explanation highlights that for the subpopulation White heterosexual individuals whose parents attended secondary school, the average TC for `analysts` observes an increase of $154,024 when the treatment hoping to become a manager in the next 5 years is applied. In contrast, the same treatment yields only an increase of $31,354 for `back-end developers`. We observe that this treatment is not a good explanation globally due to not being statistically significant.

Disparity Explanation	Support	Total Compensation (TC)				Δ
		Subpopulation (ExDis)		Global (XInsight)		
		Average	CATE	Average	CATE	
For White heterosexual individuals whose parents attended secondary school, TC growth is more influenced by hoping to become a manager in the next 5 years for analysts compared to back-end developers.	11.14%	$129,749 $110,026	$154,024 ↑ $31,354 ↑	not statistically significant		0.061
For White males aged between 18-24 years old, TC growth is more influenced by have 3-5 years in professional coding for analysts compared to back-end developers.	10.64%	$90,909 $67,754	$75,692 ↑ $24,164 ↑	not statistically significant		0.025
For males aged between 25-34 years old whose parents hold a bachelor degree, TC growth is more influenced by working in a company size between 100 - 499 workers for analysts compared to back-end developers.	13.21%	$105,694 $96,085	$70,069 ↑ $19,807 ↑	not statistically significant		0.025
For White heterosexual males aged between 25-34 years old whose parents hold a Master degree, TC growth is more influenced by not having coding as a hobby for analysts compared to back-end developers.	7.56%	$117,879 $103,957	$84,648 ↑ $43,292 ↑	not statistically significant		0.020
For White individuals aged between 25–34, TC growth is more influenced by working 6-8 years in professional coding for analysts compared to back-end developers.	34.69%	$115,777 $105,988	$44,058 ↑ $10,552 ↑	not statistically significant		0.016

Table 5: Disparity explanations discovered by ExDis for the ACS dataset.

Disparity Explanation	Support	Likelihood of having a health insurance				Δ
		Subpopulation (ExDis)		Global (XInsight)		
		Average	CATE	Average	CATE	
For White individuals from the Southern region who speak Spanish, the likelihood of having a health insurance decreases when they have no personal earnings for manual labor occupations, whereas it increases for all occupations.	8.18%	52.42% 75.17%	7.74% ↓ 4.87% ↑	78.61% 91.58%	8.02% ↓ 3.63% ↓	0.126
For individuals from the Southern region who were born in USA, the likelihood of having a health insurance decreases when they have no personal earnings for manual labor occupations, whereas it increases for all occupations.	25.09%	72.77% 81.24%	6.38% ↓ 1.83% ↑	78.61% 91.58%	8.02% ↓ 3.63% ↓	0.082
For White natives individuals from Texas who were born in USA, the likelihood of having a health insurance decreases when they have no personal earnings for manual labor occupations, whereas it increases for all occupations.	12.21%	71.86% 80.52%	5.20% ↓ 2.83% ↑	78.61% 91.58%	8.02% ↓ 3.63% ↓	0.080
For White males from the Southern region, the likelihood of having a health insurance decreases when they have no personal earnings for manual labor occupations, whereas it increases for all occupations.	18.00%	67.56% 74.20%	4.15% ↓ 2.69% ↑	78.61% 91.58%	8.02% ↓ 3.63% ↓	0.068
For individuals who were born in USA, the likelihood of having a health insurance decreases when they have no personal earnings for manual labor occupations, whereas it increases for all occupations.	75.67%	84.11% 88.97%	3.46% ↓ 1.36% ↑	78.61% 91.58%	8.02% ↓ 3.63% ↓	0.048

average likelihood of experiencing nervous attacks very frequently for males (g_1) is higher than that of non-males (g_2), in contrast to the opposite trend observed in the overall population. ExDis generated a solution identical to that of the Brute-Force approach, while Top-k identified explanations with substantial overlap among themselves. In this scenario, both DivExplorer and FairDebugger failed to produce meaningful results. The subpopulations they identified either did not align with the use case. Specifically, they did not exhibit a trend that reversed the one observed in the overall dataset, or, in the few relevant subpopulations they did find, the fixed treatment yielded a disparity score of zero, rendering the explanation ineffective for explaining the observed disparity.

Table 6: Disparity explanations discovered by ExDis for the MEPS dataset.

Disparity Explanation	Support	Likelihood of feeling nervous frequently				Δ
		Subpopulation (ExDis)		Global (XInsight)		
		Average	CATE	Average	CATE	
For individuals who were never married, are from the Southern region, don't have a doctor's recommendation to exercise, and aren't diagnosed with Diabetes, the likelihood of feeling nervous frequently decreases less for males who do not currently smoke compared to non-males.	5.78%	46.08% 41.71%	13.06% ↓ 20.73% ↓	37.58% 45.10%	3.39% ↓ 3.94% ↓	0.076
For individuals who are White, were never married, are under 29, and do not have asthma or Diabetes, the likelihood of feeling nervous frequently decreases less for males than non-males when they are uninsured for the whole year.	8.41%	49.20% 48.95%	14.16% ↓ 18.48% ↓	37.58% 45.10%	7.02% ↓ 4.86% ↓	0.043
For individuals who were never married, do not have a recommendation from the doctor to exercise, and were born in USA, the likelihood of feeling nervous frequently increases more for males who have private insurance compared to non-males.	14.25%	45.90% 45.74%	10.84% ↑ 7.51% ↑	37.58% 45.10%	4.63% ↑ 5.76% ↑	0.033
For individuals who are White, were never married, and don't have a doctor's recommendation to exercise, and don't Asthma diagnosis, the likelihood of feeling nervous frequently decreases less for males who were uninsured the whole year compared to non-males.	9.74%	47.25% 46.64%	12.52% ↓ 14.09% ↓	37.58% 45.10%	7.02% ↓ 4.86% ↓	0.015
For individuals aged between 30-42 years and don't have Diabetes, the likelihood of feeling nervous frequently increases more for males who have health insurance compared to non-males.	19.64%	43.93% 43.57%	7.47% ↑ 6.47% ↑	37.58% 45.10%	2.34% ↑ 4.54% ↑	0.010

Remark. Note that DivExplorer and FairDebugger serve as baselines only for identifying subpopulations that exhibit significant disparities between the two groups of interest. However, unlike ExDis, they do not offer any causal explanations. To enable comparison, we use a fixed treatment while using these two baselines, which is not necessarily the optimal treatment that maximizes subpopulation-level disparity scores. Therefore, these approaches result in a low total disparity score (which is expected) when the subpopulation-level disparity scores are added to obtain the score for the objective function (Problem 1). Nonetheless, the comparison allows for a qualitative contrast between the subpopulations identified by the baselines and those discovered by ExDis.

Results Summary

- The top-k baseline results in overlapping disparity explanations, demonstrating the need to consider the similarity among selected explanations.
- The output of ExDis closely matches that of Brute Force, demonstrating that our solution prioritizes efficiency without compromising quality.
- Explanations for disparity at the entire population level (as generated by XInsight) do not necessarily account for disparities within subpopulations, highlighting the need to find a specific local explanation for each subpopulation.
- DivExplorer and FairDebugger were less successful in identifying subpopulations with high disparity between g_1 and g_2, resulting in low-score explanations.

6.3 Parameters Sensitivity

In this section, We investigate the impact of various parameters on our objective function, namely disparity score Δ. Our goal is to gain insights into effective default parameter settings.

Robustness to similarity threshold τ: Figure 2 (a) shows how our objective function, i.e., the disparity score Δ changes with varying similarity threshold τ across three datasets: SO, ACS, and MEPS. As τ increases, all datasets exhibit a rising trend in disparity. This is expected because low similarity threshold significantly restricts the feasible solution space. For a fixed budget k (we used $k = 11$ for this experiment, to observe the impact of τ without any other restriction), relaxing τ allows for the inclusion of a larger number of disparity explanations, which can cumulatively increase the overall disparity score. This is because more similar subpopulations are permitted to coexist, potentially capturing less diverse and more overlapping causes of disparity. However, setting τ too high undermines the goal of maintaining diversity among the explanations, as excessive overlap can dilute the quality and reduce the distinctiveness of the explanation set. Thus, τ must be carefully chosen to balance comprehensiveness and diversity.

Robustness to number of clusters: Figure 2 (b) shows how the disparity score Δ varies with the number of clusters across three datasets: SO, ACS, and MEPS. As the number of clusters increases, all datasets exhibit an initial rise in disparity, which eventually plateaus. Notably, the ACS dataset shows the most significant increase, with the disparity growing rapidly up to 5 clusters before stabilizing around 0.4. In contrast, SO and MEPS show more modest increases, leveling off at lower disparity scores. These results suggest that finer-grained clustering help improve the disparity scores, but has diminishing return.

Robustness to the Causal DAG: The quality of the solution may depend on the quality of the underlying causal DAG. To evaluate this, we assess the impact of using different causal DAGs, generated by commonly used causal discovery algorithms. We consider the following causal DAGs: (1) **DEFAULT**, a default two-layer causal DAG in which all immutable attributes have edges to both the outcome and all mutable attributes, and all mutable attributes have

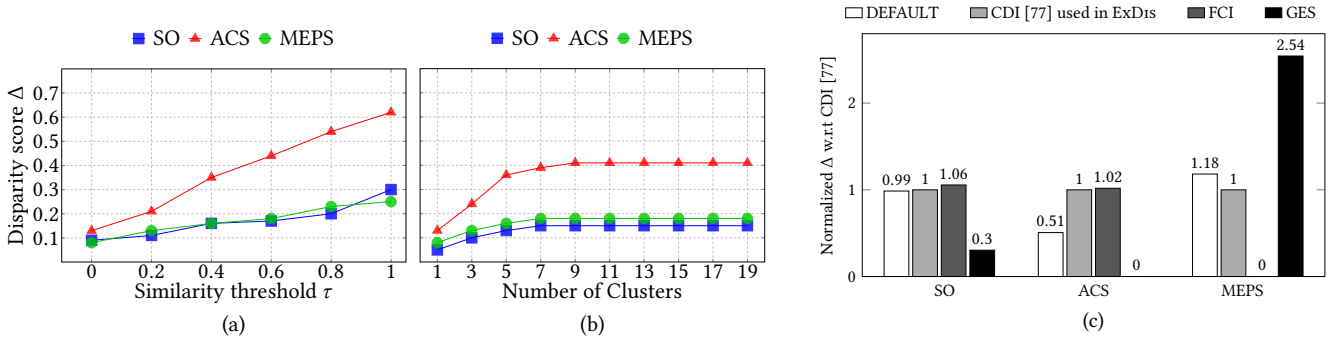


Figure 2: Effect of various system parameters on the disparity score. (a) & (b) The absolute disparity scores are reported here to show direct impact of the similarity threshold τ and the number of clusters on the disparity score Δ. (c) Effect of casual DAG modification. Disparity scores here are shown as a relative value w.r.t the disparity score of CDI [77] (which ExDis uses).

edges to the outcome. (2) Causal Data Integration (**CDI**) [77], which ExDis uses, (3) **FCI** [59], and (4) **GES** [11].

The results are depicted in Figure 2 (c). We report the relative disparity scores (Δ) computed using the different causal DAGs with respect to the disparity score of CDI. We observe that for the default naive causal DAG, the results closely resemble those obtained using the CDI causal DAG employed by ExDis across all datasets. In contrast, the FCI and GES causal DAGs produce more varied results across different use cases. This variability is likely due to the challenges that causal discovery algorithms face when applied to real-world data [20, 42], often leading to noisier and less reliable DAGs. Nonetheless, the disparity explanations generated using the FCI and GES DAGs were largely consistent with those derived from the CDI DAG, with similar patterns selected for both treatments and subpopulations. This suggests that ExDis remains robust and capable of producing meaningful explanations even when the underlying causal DAG is noisy or imperfect.

Results Summary

- A higher τ can increase the disparity score by allowing overlapping explanations at the cost of compromising diversity.
- Finer-grained clustering helps improve the disparity scores, but has diminishing return.
- Even with imperfect causal DAGs, ExDis is able to produce meaningful and robust disparity explanations.

6.4 Efficiency & Scalability

In this section, we present results demonstrating the efficiency of various components of ExDis, analyze the impact of different parameters on its runtime, and evaluate its scalability as the dataset grows both vertically (in tuples) and horizontally (in attributes).

Step-by-step breakdown of runtime: We present a step-wise breakdown of runtime in Table 7. Not surprisingly, The step "explanation miner", which focuses on identifying the causal explanation for each subpopulation, is the most computationally expensive one, accounting for over 80% of the total runtime in all examined scenarios. Nevertheless, ExDis generates the solution within a reasonable time, even for large, high-dimensional datasets like ACS.

Effect of various parameters on runtime: Next, we analyze how various parameters impact runtime. Since parameter variations involve sampling, we repeat each experiment 5 times and report the average runtime across all runs.

Solution size k. Figure 3 (left) shows the impact of the solution size k on the runtime. Note that k only affects the last step (fast greedy search), which selects the explanations from the candidates mined in the previous steps. Recall that this step evaluates the pairwise intersection between the subpopulations corresponding to the disparity explanations to account for the diversity constraint. As k increases, the number of pairwise comparisons grows and so does the runtime. The effect of k on runtime is negligible for the smaller datasets (MEPS and SO) compared to the larger dataset ACS, where computing intersections among subpopulations takes longer due to the dataset's size.

Number of attributes. Figure 3 (center) shows the impact of the number of attributes on runtime. In this experiment, we randomly sampled subsets of attributes to retain and removed the rest from the dataset. The number of attributes influences the search-space size, as more attributes result in a larger set of subpopulations and treatment patterns to consider. Theoretically, runtime should grow exponentially with the number of attributes. However, this worst-case behavior was not always observed in practice, as several factors influence computation—such as the structure of the underlying causal DAG, the choice of mutable and immutable attributes, and other system parameters. Nevertheless, we observe that, as expected, the number of attributes in the dataset significantly affects runtime. For the largest dataset, ACS, we do in fact observe exponential growth in runtime as the number of attributes increases.

Dataset cardinality. Figure 3 (right) illustrates the impact of the dataset cardinality (in number of tuples) on runtime. In this experiment, we varied the dataset size using specific-sized horizontal dataset slices. We find that the runtime is linearly influenced by the number of tuples. For the ACS dataset, we applied the sampling optimization (during the Explanation Miner phase, as explained in Section 5.2). We found that the growth in runtime is more moderate up to around 30% of the ACS data (which is 500,000 tuples). Beyond this point, the Explanation Miner module operates on a random fixed-sized sample of 500,000 tuples.

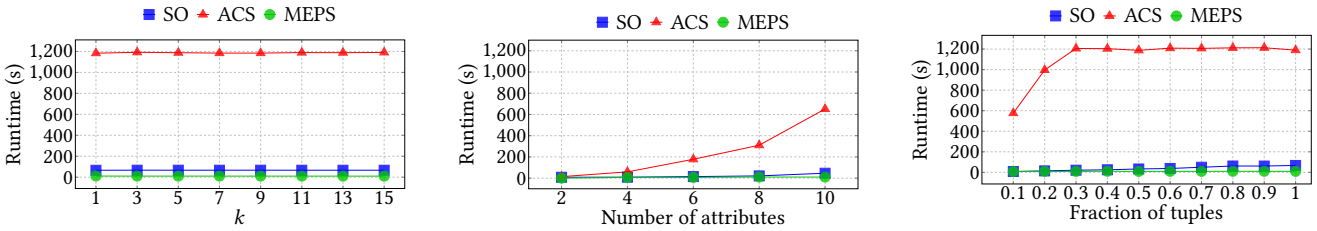


Figure 3: Effects of various parameters on runtime: (left) the budget parameter k, (center) number of attributes, and (right) fraction of data.

Table 7: Breakdown of runtime by steps (seconds).

Dataset	Subpopulation Miner	Explanation Miner	Fast Greedy Search
SO	0.4	70.1	0.5
ACS	3.0	1294.6	13.0
MEPS	0.3	11.3	0.1

Results Summary

- The explanation miner step of ExDis, which focuses on identifying the causal explanation for each subpopulation, takes the longest time, accounting for over 80% of the total runtime.
- The runtime grows (almost linearly) with the solution size k because of the pair-wise similarity computation.
- The runtime is greatly influenced by the number of attributes in the dataset, as it affects the size of the search space.
- The runtime grows linearly with the number of dataset tuples.

6.5 Ablation Study

To assess the impact of our proposed optimizations on the runtime of ExDis, we compare 5 variants: (1) None, implying no optimization was applied, (2) No Parallel, denoting the setting where parallelization was removed, (3) No Cache, denoting the setting where caching was removed, (4) No Clustering, denoting the setting where clustering was removed, (5) All, denoting the ExDis setting where all optimizations were applied.

Figure 4 illustrates how the removal of individual optimization techniques affects runtime performance across three datasets: SO, ACS, and MEPS. The y-axis is plotted on a logarithmic scale to clearly illustrate differences in runtimes. As expected, without any optimization, we observe the highest runtimes across all three datasets. We also observe that ExDis, with all optimizations, yield the best runtime performance. Removing clustering increases runtimes moderately, indicating clustering provides a noticeable but relatively modest speedup. Eliminating caching also causes a moderate performance degradation. While caching was introduced to avoid redundant computations, it offers relatively smaller gains. This is because when causal DAGs are small, modifying them may require less overhead than reading and writing them from cache. However, we observe substantial reduction in runtime using caching for the larger datasets SO and ACS. The absence of parallelization significantly worsens performance, especially on SO and ACS datasets, indicating its significant contribution in boosting the runtime performance across the board.

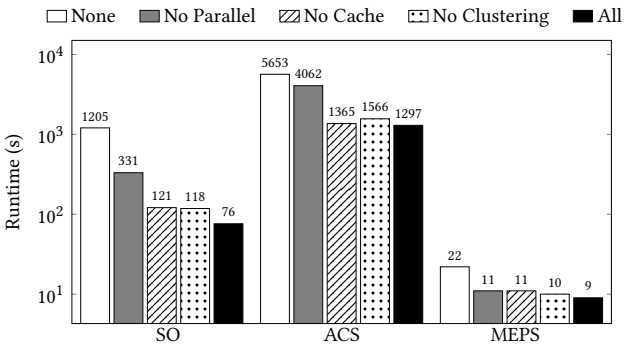


Figure 4: Effect of various settings of using optimization techniques on runtime across three datasets. Note that the y-axis is in log scale.

Overall, while all optimizations contribute to performance gains, parallelization yields the most substantial runtime improvements. Clustering also helps, though to a lesser extent. This experiment underscores the critical value of each optimization, highlighting how removal of any of them can degrade runtime efficiency.

7 LIMITATIONS AND FUTURE WORK

We have presented ExDis, a framework for discovering causal explanations for disparities between two groups of interest. ExDis identifies data regions where disparities are most pronounced (or reversed), and associates specific factors that causally contribute to the disparity. We acknowledge that several factors can influence the quality of the disparity explanations, including data quality, the quality of the underlying causal model, and system parameters. In Section 6.3, we showed that meaningful results can still be obtained even with imperfect causal DAGs. We also provided insights on tuning system parameters to achieve satisfactory results across different use cases and datasets.

ExDis currently operates on single-relation databases, assuming no dependencies between tuples. As explained in [75], while treatment and grouping patterns are straightforward in single-table scenarios, extending these concepts to multi-table databases introduces significant challenges. Supporting multi-relation datasets with dependencies among tuples remains an important avenue for future research. Notably, most prior works that apply causal inference for explanations [36, 52, 75, 76] are also limited to single-relation databases. Future work will focus on extending the framework to support comparisons among multiple groups, enabling more comprehensive analyses.

REFERENCES

[1] 2021. 2021 Stackoverflow Developer Survey. https://insights.stackoverflow.com/survey/2021.

[2] Agency for Healthcare Research and Quality (AHRQ). 2024. Medical Expenditure Panel Survey (MEPS) - Data Overview. https://meps.ahrq.gov/mepsweb/data_stats/data_overview.jsp Accessed: 2024-01-30.

[3] Shunit Agmon, Amir Gilad, Brit Youngmann, Shahar Zoarets, and Benny Kimelfeld. 2024. Finding Convincing Views to Endorse a Claim. *Proceedings of the VLDB Endowment* 18, 2 (2024), 439–452.

[4] Rakesh Agrawal, Ramakrishnan Srikant, et al. 1994. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, Vol. 1215. Santiago, Chile, 487–499.

[5] Abolfazl Asudeh, Zhongjun Jin, and HV Jagadish. 2019. Assessing and remedying coverage for a given dataset. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 554–565.

[6] Anonymous Authors. 2024. Causal Explanation for Disparity. https://anonymous.4open.science/r/CausalExplanationforDisparity-257A/ Anonymous GitHub repository.

[7] Nicole Bidoit, Melanie Herschel, and Katerina Tzompanaki. 2014. Query-based why-not provenance with nedexplain. In *Extending database technology (EDBT)*.

[8] Pierre Bourhis, Daniel Deutch, and Yuval Moskovitch. 2020. Equivalence-Invariant Algebraic Provenance for Hyperplane Update Queries. In *Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020, online conference [Portland, OR, USA], June 14-19, 2020*. ACM, 415–429. https://doi.org/10.1145/3318464.3380578

[9] Ángel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. 2019. FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning. In *14th IEEE Conference on Visual Analytics Science and Technology, IEEE VAST 2019, Vancouver, BC, Canada, October 20-25, 2019*, Remco Chang, Daniel A. Keim, and Ross Maciejewski (Eds.). IEEE, 46–56. https://doi.org/10.1109/VAST47406.2019.8986948

[10] Adriane Chapman and HV Jagadish. 2009. Why not?. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*. 523–534.

[11] D.M Chickering. 2002. Optimal structure identification with greedy search. *JMLR* 3, Nov (2002), 507–554.

[12] Yeounoh Chung, Tim Kraska, Neoklis Polyzotis, Ki Hyun Tae, and Steven Euijong Whang. 2019. Automated data slicing for model validation: A big data-ai integration approach. *IEEE Transactions on Knowledge and Data Engineering* 32, 12 (2019), 2284–2296.

[13] Daniel Deutch, Nave Frost, and Amir Gilad. 2020. Explaining Natural Language query results. *VLDB J.* 29, 1 (2020), 485–508.

[14] Daniel Deutch, Amir Gilad, Tova Milo, Amit Mualem, and Amit Somech. 2022. FEDEX: An Explainability Framework for Data Exploration Steps. *Proc. VLDB Endow.* 15, 13 (2022), 3854–3868. https://www.vldb.org/pvldb/vol15/p3854-gilad.pdf

[15] Daniel Deutch, Amir Gilad, and Yuval Moskovitch. 2015. Selective Provenance for Datalog Programs Using Top-K Queries. *Proc. VLDB Endow.* 8, 12 (2015), 1394–1405. https://doi.org/10.14778/2824032.2824039

[16] Kareem El Gebaly, Parag Agrawal, Lukasz Golab, Flip Korn, and Divesh Srivastava. 2014. Interpretable and informative explanations of outcomes. *Proceedings of the VLDB Endowment* 8, 1 (2014), 61–72.

[17] Sainyam Galhotra, Amir Gilad, Sudeepa Roy, and Babak Salimi. 2022. Hyper: Hypothetical reasoning with what-if and how-to queries using a probabilistic causal approach. In *Proceedings of the 2022 International Conference on Management of Data*. 1598–1611.

[18] Sainyam Galhotra, Romila Pradhan, and Babak Salimi. 2021. Explaining Black-Box Algorithms Using Probabilistic Contrastive Counterfactuals. In *SIGMOD*. ACM, 577–590.

[19] Floris Geerts, Bart Goethals, and Taneli Mielikäinen. 2004. Tiling databases. In *Discovery Science: 7th International Conference, DS 2004, Padova, Italy, October 2-5, 2004. Proceedings 7*. Springer, 278–289.

[20] Clark Glymour, Kun Zhang, and Peter Spirtes. 2019. Review of causal discovery methods based on graphical models. *Frontiers in genetics* 10 (2019), 524.

[21] Ricky Charles Godbolt. 2011. *Black and Blue: African Americans, Blue-Collar Bias, and the Construction Industry in Prince George's County, Maryland*. Ph. D. Dissertation. University of Phoenix.

[22] Jochen Hipp, Ulrich Güntzer, and Gholamreza Nakhaeizadeh. 2000. Algorithms for association rule mining—a general survey and comparison. *ACM sigkdd explorations newsletter* 2, 1 (2000), 58–64.

[23] Paul W Holland. 1986. Statistics and causal inference. *Journal of the American statistical Association* 81, 396 (1986), 945–960.

[24] Manas Joglekar, Hector Garcia-Molina, and Aditya G. Parameswaran. 2019. Interactive Data Exploration with Smart Drill-Down. *IEEE Trans. Knowl. Data Eng.* 31, 1 (2019), 46–60. https://doi.org/10.1109/TKDE.2017.2685998

[25] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2023. A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations. *Comput. Surveys* 55, 5 (2023), 95:1–95:29.

[26] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. 2021. Algorithmic Recourse: from Counterfactual Explanations to Interventions. In *FAccT*. ACM, 353–362.

[27] Been Kim, Cynthia Rudin, and Julie A Shah. 2014. The bayesian case model: A generative approach for case-based reasoning and prototype classification. *Advances in neural information processing systems* 27 (2014).

[28] Trupti A Kumbhare and Santosh V Chobe. 2014. An overview of association rule mining algorithms. *International Journal of Computer Science and Information Technologies* 5, 1 (2014), 927–930.

[29] Laks VS Lakshmanan, Jian Pei, and Jiawei Han. 2002. Quotient cube: How to summarize the semantics of a data cube. In *VLDB'02: Proceedings of the 28th International Conference on Very Large Databases*. Elsevier, 778–789.

[30] Seokki Lee, Bertram Ludäscher, and Boris Glavic. 2020. Approximate Summaries for Why and Why-not Provenance. *Proceedings of the VLDB Endowment* 13, 6 (2020).

[31] Benton Li, Nativ Levy, Brit Youngmann, Sainyam Galhotra, and Sudeepa Roy. 2025. Fair and Actionable Causal Prescription Ruleset. *Proceedings of the ACM on Management of Data* 3, 3 (2025), 1–28.

[32] Chenjie Li, Zhengjie Miao, Qitian Zeng, Boris Glavic, and Sudeepa Roy. 2021. Putting Things into Context: Rich Explanations for Query Answers using Join Graphs. In *Proceedings of the 2021 International Conference on Management of Data*. 1051–1063.

[33] Jinyang Li, Yuval Moskovitch, and H. V. Jagadish. 2023. Detection of Groups with Biased Representation in Ranking. In *39th IEEE International Conference on Data Engineering, ICDE 2023, Anaheim, CA, USA, April 3-7, 2023*. IEEE, 2167–2179. https://doi.org/10.1109/ICDE55515.2023.00168

[34] Yin Lin, Brit Youngmann, Yuval Moskovitch, HV Jagadish, and Tova Milo. 2021. On detecting cherry-picked generalizations. *Proceedings of the VLDB Endowment* 15, 1 (2021), 59–71.

[35] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. 2013. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 623–631.

[36] Pingchuan Ma, Rui Ding, Shuai Wang, Shi Han, and Dongmei Zhang. 2023. XInsight: EXplainable Data Analysis Through The Lens of Causality. *Proc. ACM Manag. Data*, Article 156 (jun 2023), 27 pages.

[37] Alexandra Meliou, Wolfgang Gatterbauer, Katherine F Moore, and Dan Suciu. 2009. Why so? or why no? functional causality for explaining query answers. *arXiv preprint arXiv:0912.5340* (2009).

[38] Alexandra Meliou, Wolfgang Gatterbauer, Katherine F Moore, and Dan Suciu. 2010. The Complexity of Causality and Responsibility for Query Answers and non-Answers. *Proceedings of the VLDB Endowment* 4, 1 (2010).

[39] Zhengjie Miao, Qitian Zeng, Boris Glavic, and Sudeepa Roy. 2019. Going beyond provenance: Explaining query answers with pattern-based counterbalances. In *Proceedings of the 2019 International Conference on Management of Data*. 485–502.

[40] Yuval Moskovitch, Jinyang Li, and H. V. Jagadish. 2023. Dexer: Detecting and Explaining Biased Representation in Ranking. In *Companion of the 2023 International Conference on Management of Data, SIGMOD/PODS 2023, Seattle, WA, USA, June 18-23, 2023*. ACM, 159–162. https://doi.org/10.1145/3555041.3589725

[41] Zafeiria Moumoulidou, Andrew McGregor, and Alexandra Meliou. 2021. Diverse Data Selection under Fairness Constraints. In *24th International Conference on Database Theory, ICDT 2021, March 23-26, 2021, Nicosia, Cyprus (LIPIcs, Vol. 186)*, Ke Yi and Zhewei Wei (Eds.). Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 13:1–13:25. https://doi.org/10.4230/LIPICS.ICDT.2021.13

[42] RT O'donnell, Ann E Nicholson, B Han, Kevin B Korb, MJ Alam, and LR Hope. 2006. Incorporating expert elicited structural information in the CaMML causal discovery program. In *Proceedings of the 19th Australian Joint Conference on Artificial Intelligence: Advances in Artificial Intelligence*. 1–16.

[43] Eliana Pastor, Luca De Alfaro, and Elena Baralis. 2021. Looking for trouble: Analyzing classifier behavior via pattern divergence. In *Proceedings of the 2021 International Conference on Management of Data*. 1400–1412.

[44] Judea Pearl. 2009. Causal inference in statistics: An overview. (2009).

[45] Drago Plecko and Elias Bareinboim. 2023. Causal fairness for outcome control. *Advances in Neural Information Processing Systems* 36 (2023), 47575–47597.

[46] Drago Plecko and Elias Bareinboim. 2024. Causal fairness analysis. *ECAI* (2024).

[47] Sudeepa Roy, Laurel Orr, and Dan Suciu. 2015. Explaining query answers with explanation-ready databases. *Proceedings of the VLDB Endowment* 9, 4 (2015), 348–359.

[48] Sudeepa Roy and Dan Suciu. 2014. A formal approach to finding explanations for database queries. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*. 1579–1590.

[49] Donald Bruce Rubin. 1971. *The use of matched sampling and regression adjustment in observational studies*. Ph. D. Dissertation. Harvard University.

[50] Donald B Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* 100, 469 (2005), 322–331.

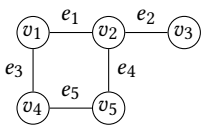
[51] Omer Sagi and Lior Rokach. 2021. Approximating XGBoost with an interpretable decision tree. *Information Sciences* 572 (2021), 522–542.

[52] Babak Salimi, Johannes Gehrke, and Dan Suciu. 2018. Bias in olap queries: Detection, explanation, and removal. In *Proceedings of the 2018 International*

Conference on Management of Data. 1021–1035.
[53] Sunita Sarawagi. 2000. User-adaptive exploration of multidimensional data. In *VLDB*. ResearchGate GmbH, 307–316.
[54] Sunita Sarawagi. 2001. User-cognizant multidimensional analysis. *The VLDB Journal* 10 (2001), 224–239.
[55] Sunita Sarawagi, Rakesh Agrawal, and Nimrod Megiddo. 1998. Discovery-driven exploration of OLAP data cubes. In *Advances in Database Technology—EDBT'98: 6th International Conference on Extending Database Technology Valencia, Spain, March 23–27, 1998 Proceedings 6*. Springer, 168–182.
[56] Gayatri Sathe and Sunita Sarawagi. 2001. Intelligent rollups in multidimensional OLAP data. In *VLDB*. 307–316.
[57] Holger Schielzeth. 2010. Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution* 1, 2 (2010), 103–113.
[58] Amit Sharma and Emre Kiciman. 2020. DoWhy: An End-to-End Library for Causal Inference. *arXiv preprint arXiv:2011.04216* (2020).
[59] P. Spirtes et al. 2000. *Causation, prediction, and search*. MIT press.
[60] Hao Sun, Evan Munro, Georgy Kalashnov, Shuyang Du, and Stefan Wager. 2021. Treatment allocation under uncertain costs. *arXiv preprint arXiv:2103.11066* (2021).
[61] Tanmay Surve and Romila Pradhan. 2024. Example-based Explanations for Random Forests using Machine Unlearning. *CoRR* abs/2402.05007 (2024).
[62] Yuchao Tao, Amir Gilad, Ashwin Machanavajjhala, and Sudeepa Roy. 2022. DPX-Plain: Privately Explaining Aggregate Query Answers. *Proc. VLDB Endow.* 16, 1 (2022), 113–126. https://www.vldb.org/pvldb/vol16/p113-tao.pdf
[63] Balder ten Cate, Cristina Civili, Evgeny Sherkhonov, and Wang-Chiew Tan. 2015. High-level why-not explanations using ontologies. In *Proceedings of the 34th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*. 31–43.
[64] U.S. Census Bureau. 2024. American Community Survey (ACS) - Data. https://www.census.gov/programs-surveys/acs/data.html Accessed: 2024-01-30.
[65] Moshe Y. Vardi. 1982. The Complexity of Relational Query Languages (Extended Abstract). In *Proceedings of the Fourteenth Annual ACM Symposium on Theory of Computing* (San Francisco, California, USA) *(STOC '82)*. ACM, New York, NY, USA, 137–146. https://doi.org/10.1145/800070.802186
[66] Manasi Vartak, Sajjadur Rahman, Samuel Madden, Aditya Parameswaran, and Neoklis Polyzotis. 2015. Seedb: Efficient data-driven visualization recommendations to support visual analytics. In *VLDB*, Vol. 8. NIH Public Access, 2182.
[67] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods* 17, 3 (2020), 261–272.
[68] Clifford H Wagner. 1982. Simpson's paradox in real life. *The American Statistician* 36, 1 (1982), 46–48.
[69] Tong Wang and Cynthia Rudin. 2022. Causal rule sets for identifying subgroups with enhanced treatment effects. *INFORMS Journal on Computing* 34, 3 (2022), 1626–1643.
[70] Yue Wang, Alexandra Meliou, and Gerome Miklau. 2018. RC-Index: Diversifying Answers to Range Queries. *Proc. VLDB Endow.* 11, 7 (2018), 773–786. https://doi.org/10.14778/3192965.3192969
[71] Yuhao Wen, Xiaodan Zhu, Sudeepa Roy, and Jun Yang. 2018. Interactive summarization and exploration of top aggregate query answers. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, Vol. 11. NIH Public Access, 2196.
[72] Eugene Wu and Samuel Madden. 2013. Scorpion: Explaining away outliers in aggregate queries. (2013).
[73] Yu Xie, Jennie E Brand, and Ben Jann. 2012. Estimating heterogeneous treatment effects with observational data. *Sociological methodology* 42, 1 (2012), 314–347.
[74] Brit Youngmann, Sihem Amer-Yahia, and Aurélien Personnaz. 2022. Guided Exploration of Data Summaries. *Proc. VLDB Endow.* 15, 9 (2022).
[75] Brit Youngmann, Michael Cafarella, Amir Gilad, and Sudeepa Roy. 2024. Summarized Causal Explanations For Aggregate Views. *Proceedings of the ACM on Management of Data* 2, 1 (2024), 1–27.
[76] Brit Youngmann, Michael Cafarella, Yuval Moskovitch, and Babak Salimi. 2023. On Explaining Confounding Bias. *2023 IEEE 39th International Conference on Data Engineering (ICDE)* (2023).
[77] Brit Youngmann, Michael Cafarella, Babak Salimi, and Anna Zeng. 2023. Causal Data Integration. *Proceedings of the VLDB Endowment* 16, 10 (2023), 2659–2665.
[78] Cong Yu, Laks Lakshmanan, and Sihem Amer-Yahia. 2009. It takes variety to make a world: diversification in recommender systems. In *Proceedings of the 12th international conference on extending database technology: Advances in database technology*. 368–378.
[79] Xiaozhong Zhang, Xiaoyu Ge, Panos K Chrysanthis, and Mohamed A Sharaf. 2021. Viewseeker: An interactive view recommendation framework. *Big Data Research* 25 (2021), 100238.

Appendix A NP-HARDNESS PROOF

In this part, we give the proof for showing the hardness of the decision problem defined in Section 4 (proposition 4.1).



(a) Input instance to the IS problem

	v_1	v_2	v_3	v_4	v_5	g_1	g_1	T	O
e_1	1	1	0	0	0	★	★	★	★
e_2	0	1	1	0	0	★	★	★	★
e_3	1	0	0	1	0	★	★	★	★
e_4	0	1	0	0	1	★	★	★	★
e_5	0	0	0	1	1	★	★	★	★

(b) The database D of the reduction output

Figure 5: Reduction Example

Proof of Proposition 4.1. We show a polynomial reduction from the Independent Set (IS) problem. Recall that $\langle G = (V, E), k\rangle \in IS$ if there exists $I \subseteq V$ s.t. $|I| = k$ and $\forall v_i, v_j \in I$ there is no edge between v_i and v_j in G. Given an input instance $\langle G = (V, E), k\rangle$ to the IS problem, we create an input instance to our decision problem as follows. D is a dataset with $|V|+4$ attributes and $|E|$ tuples, where there is an attribute v_i for each $v_i \in V$, the attributes g_1, g_2 and an attribute O. There is a tuple e_i for each $e_i \in E$ where $e_i[v_j] = 1 \iff v_j \in e_i$. The values in the g_1, g_2, T and O attributes can be assigned arbitrarily, as they do not affect the reduction correctness. Figure 5 shows a simple example of the resulting database D. We set S to be 1, τ to be 0, B to be 0, and use the same k value. We define the set of possible disparity explanations to be $\{(v_i = 1, \star) \mid v_i \in V\}$. I.e. the set of possible disparity explanations consist of $|V|$ disparity explanation, for each one, the grouping pattern ψ_g is defined as $v_i = 1$ whereas the treatment pattern ψ_e can be assigned arbitrarily. Note that for $\phi_i = (v_i = 1, \star)$ and $\phi_j = (v_j = 1, \star)$ we have that $\text{SIM}(\phi_i, \phi_j) = 0$ if and only if there is no edge between v_i and v_j. Finally, the causal model $\mathcal{G}_D$ can be defined arbitrarily. Clearly, this reduction is polynomial. We next show that $\langle G = (V, E), k\rangle \in IS \iff$ there exists a subset $\Phi \subseteq \{(v_i = 1, \star) \mid v_i \in V\}$, such that $|\Phi| \leq k$, $\forall \phi_i \in \Phi$, $support(\phi_i) \geq S$, $\forall \phi_i, \phi_j \in \Phi$, $\text{SIM}(\phi_i, \phi_j) \leq \tau$ and $\sum_{\phi \in \Phi} \Delta(\phi) \geq B$.

First, assume that $\langle G = (V, E), k\rangle \in IS$. Thus, there is a set $I \subseteq V$ such that $|I| = k$ and $\forall v_i, v_j \in I$ there is no edge between v_i and v_j. We define $\Phi = \{(v_i = 1, \star) \mid v_i \in I\}$. Clearly, $\forall \phi_i \in \Phi$, $support(\phi_i) \geq 1$. Note that since I is an IS, $\forall \phi_i, \phi_j \in \Phi$, $\text{SIM}(\phi_i, \phi_j) = 0$, and since $\Delta(\phi) \geq 0$ for every ϕ we have that $\sum_{\phi \in \Phi} \Delta(\phi) \geq 0 = B$.

Now assume there exists a subset $\Phi \subseteq \{(v_i = 1, \star) \mid v_i \in V\}$, such that $|\Phi| \leq k$, $\forall \phi_i \in \Phi$, $support(\phi_i) \geq 1$, $\forall \phi_i, \phi_j \in \Phi$, $\text{SIM}(\phi_i, \phi_j) = 0$ and $\sum_{\phi \in \Phi} \Delta(\phi) \geq B = 0$. From the construction $\text{SIM}(\phi_i, \phi_j) = 0$ where $\phi_i = (v_i = 1, \star)$ and $\phi_j = (v_j = 1, \star)$ only when here is no edge between v_i and v_j. Therefore, the set $I = \{v_i \mid (v_i = 1, \star) \in \Phi\}$ is an IS of size k in G.

□

Appendix B RESULTS FROM THE BASELINES

DivExplorer and FairDebugger only produced explanations for the ACS dataset, which are shown in Table 8 and Table 9, respectively.

Top-k.

- The results for the SO dataset are in Table 10.
- The results for the ACS dataset are in Table 11.
- The results for the MEPS dataset are in Table 12.

Brute-Force.

- The results for the SO dataset are in Table 13.
- The results for the ACS dataset are in Table 14.
- The results for the MEPS dataset are in Table 15.

Disparity Explanation	Support	Likelihood of having a health insurance				Δ
		Subpopulation		Global (XInsight)		
		Average	CATE	Average	CATE	
For White individuals from the Southern region who speak Spanish, the likelihood of having health insurance decreases when they have no personal earnings for manual labor occupations whereas it increases for all occupations.	8.18%	52.12% 61.03%	7.74% ↓ 4.80% ↑	78.61% 91.58%	2.65% ↓ 1.98% ↑	0.126
Among White natives from Texas who born in USA, the likelihood of having health insurance decreases when they have no personal earnings for manual labor occupations whereas it increases for all occupations.	12.21%	71.86% 80.52%	5.20% ↓ 2.83% ↑	78.61% 91.58%	2.65% ↓ 1.98% ↑	0.08
For individuals from the Southern region, the likelihood of having health insurance decreases when they have no personal earnings for manual labor occupations whereas it increases for all occupations.	34.1%	65.80% 75.78%	5.25% ↓ 2.45% ↑	78.61% 91.58%	2.65% ↓ 1.98% ↑	0.077
For White males from the Southern region, the likelihood of having health insurance decreases when they have no personal earnings for manual labor occupations whereas it increases for all occupations.	18.00%	67.56% 74.20%	4.22% ↓ 2.68% ↑	78.61% 91.58%	2.65% ↓ 1.98% ↑	0.069
For White natives from the South region without disabilities, the likelihood of having health insurance decreases when they have no personal earnings for manual labor occupations whereas it increases for all occupations.	18.17%	73.05% 81.50%	4.17% ↓ 2.11% ↑	78.61% 91.58%	2.65% ↓ 1.98% ↑	0.063

Table 8: DivExplorer results for the ACS dataset.

Disparity Explanation	Support	Likelihood of having a health insurance				Δ
		Subpopulation		Global (XInsight)		
		Average	CATE	Average	CATE	
For individuals from the Southern region, the likelihood of having health insurance decreases when they have no personal earnings for manual labor occupations whereas it increases for all occupations.	34.10%	65.85% 75.78%	5.25% ↓ 2.43% ↑	78.61% 91.58%	2.65% ↓ 1.98% ↑	0.076
Among individuals who born in USA, the likelihood of having health insurance decreases when they have no personal earnings for manual labor occupations whereas it increases for all occupations.	75.67%	84.11% 88.97%	3.46% ↓ 1.36% ↑	78.61% 91.58%	2.65% ↓ 1.98% ↑	0.048

Table 9: FairDebugger results for the ACS dataset.

Disparity Explanation	Support	Total Compensation (TC)				Δ
		Subpopulation		Global (XInsight)		
		Average	CATE	Average	CATE	
For White males aged between 18-24, TC growth is more influenced by having coding as a hobby and learned in undergrad major computer science for analysts compared to back-end developers.	10.64%	$90,909 $67,754	$219,469 ↑ $55,233 ↑	not statistically significant		0.082
For heterosexual individuals whose parents attended to secondary school, TC growth is more influenced by the desire become manager and not being a student for analysts compared to back-end developers.	14.41%	$112,896 $99,126	$176,698 ↑ $53,814 ↑	not statistically significant		0.061
For White heterosexual individuals whose parents attended to secondary school, TC growth is more influenced by the desire become manager for analysts compared to back-end developers.	11.14%	$129,749 $110,026	$154,024 ↑ $31,354 ↑	not statistically significant		0.061
For White individuals aged between 25-34, TC increases by having 3-5 years of professional coding experience and spending over 12 hours on computer daily for analysts whereas it decreases for back-end developers.	34.69%	$115,777 $105,988	$93,826 ↑ $25,951 ↓	$106,542 $96,609	$55,485 ↑ $14,854 ↓	0.059
For White heterosexual males whose parents attended to secondary school, TC growth is more influenced by the desire become manager for analysts compared to back-end developers.	10.77%	$125,581 $109,837	$147,514 ↑ $32,445 ↑	not statistically significant		0.057

Table 10: Top-k results for the SO dataset.

Disparity Explanation	Support	Likelihood of having a health insurance				Δ
		Subpopulation		Global (XInsight)		
		Average	CATE	Average	CATE	
For natives from the Southern region who were born in USA, the likelihood of having health insurance increases when they didn't report absence from work and didn't attend school in the last 3 months for manual labor occupations, whereas it decreases for all occupations.	25.09%	72.77% 81.24%	9.58% ↑ 4.22% ↓	78.61% 91.58%	5.47% ↑ 0.50% ↓	0.138
Among individuals from the Southern region who were born in USA, the likelihood of having health insurance increases when they didn't report absence from work and didn't attend school in the last 3 months for manual labor occupations, whereas it decreases for all occupations.	25.09%	72.77% 81.24%	9.58% ↑ 4.22% ↓	78.61% 91.58%	5.47% ↑ 0.50% ↓	0.138
For White individuals from the Southern region who speak Spanish, the likelihood of having health insurance decreases when they have no personal earnings for manual labor occupations, whereas it increases for all occupations.	8.18%	52.12% 61.03%	7.74% ↓ 4.87% ↑	78.61% 91.58%	2.65% ↓ 1.98% ↑	0.126
For individuals from the Southern region who speak Spanish, the likelihood of having health insurance decreases when they have no personal earnings for manual labor occupations, whereas it increases for all occupations.	10.30%	51.48% 59.88%	6.63% ↓ 3.87% ↑	78.61% 91.58%	2.65% ↓ 1.98% ↑	0.105
For White individuals from Texas, Southern region who were born in USA, the likelihood of having health insurance decreases when they have no personal earnings for manual labor occupations, whereas it increases for all occupations.	12.21%	71.86% 80.52%	5.20% ↓ 2.83% ↑	78.61% 91.58%	2.65% ↓ 1.98% ↑	0.080

Table 11: Top-k results for the ACS dataset.

Disparity Explanation	Support	Likelihood of feeling nervous frequently				Δ
		Subpopulation		Global (XInsight)		
		Average	CATE	Average	CATE	
For individuals who never married, are from South region, don't have doctor's recommendation to exercise, and don't have Diabetes, the likelihood of feeling nervous frequently decreases less for males if they don't smoke currently compared to non-males.	5.78%	46.08% 41.71%	13.06% ↓ 20.73% ↓	37.58% 45.10%	3.39% ↓ 3.94% ↓	0.076
Among individuals who never married, are from South region, don't have doctor's recommendation to exercise, and don't have Diabetes or Asthma, the likelihood of feeling nervous frequently decreases less for males than non-males when they don't smoke currently.	5.3%	46.52% 41.33%	13.12% ↓ 19.36% ↓	37.58% 45.10%	3.39% ↓ 3.94% ↓	0.062
For individuals who never married, are from South region, don't have doctor's recommendation to exercise, the likelihood of feeling nervous frequently decreases less for males if they don't smoke currently compared to non-males.	5.89%	46.00% 41.24%	12.56% ↓ 18.40% ↓	37.58% 45.10%	3.39% ↓ 3.94% ↓	0.058
For individuals who never married, are from South region, don't have doctor's recommendation to exercise, and don't have Asthma, the likelihood of feeling nervous frequently decreases less for males who do not currently smoke compared to non-males.	5.38%	46.44% 40.89%	12.66% ↓ 18.38% ↓	37.58% 45.10%	3.39% ↓ 3.94% ↓	0.057
For White individuals who never married, aged below 29 years, and don't have Asthma or Diabetes, the likelihood of feeling nervous frequently decreases less for males who haven't health insurance compared to non-males.	8.41%	49.21% 48.95%	14.16% ↓ 18.48% ↓	37.58% 45.10%	6.77% ↓ 4.93% ↓	0.043

Table 12: Top-k results for the MEPS dataset.

Disparity Explanation	Support	Total Compensation (TC)				Δ
		Subpopulation		Global (XInsight)		
		Average	CATE	Average	CATE	
For White males aged between 18-24, TC growth is more influenced by having coding as a hobby and learned in undergrad major computer science for analysts compared to back-end developers.	10.64%	$90,909 $67,754	$219,469 ↑ $55,233 ↑	not statistically significant		0.082
For heterosexual individuals whose parents attended to secondary school, TC growth is more influenced by the desire become manager and not being a student for analysts compared to back-end developers.	14.41%	$112,896 $99,126	$176,698 ↑ $53,814 ↑	not statistically significant		0.061
For White individuals aged between 25-34, TC growth increases by having 3-5 years of professional coding experience and spending over 12 hours on computer daily for analysts whereas it decreases for back-end developers.	34.69%	$115,777 $105,988	$93,826 ↑ $25,951 ↓	$106,542 $96,609	$55,485 ↑ $14,854 ↓	0.059
For White individuals, TC increases by having 3-5 professional coding years and spending over 12 hours on computer daily for analysts whereas it decreases for back-end developers.	67.49%	$122,765 $108,953	$63,311 ↑ $23,068 ↓	$106,542 $96,609	$55,485 ↑ $14,854 ↓	0.043
For males between the age 25-34 whose parents hold a bachelor's degree, TC growth is more influenced by working in a company size between 100 - 499 workers for analysts compared to back-end developers.	13.21%	$105,694 $96,085	$70,069 ↑ $19,807 ↑	not statistically significant		0.025

Table 13: Brute-Force results for the SO dataset.

Disparity Explanation	Support	Likelihood of having a health insurance				Δ
		Subpopulation		Global (XInsight)		
		Average	CATE	Average	CATE	
For natives from the Southern region who were born in USA, the likelihood of having health insurance increases when they didn't report absence from work and didn't attend school in the last 3 months for manual labor occupations, whereas it decreases for all occupations.	25.09%	72.77% 81.24%	9.58% ↑ 4.22% ↓	78.61% 91.58%	5.47% ↑ 0.50% ↓	0.138
For White individuals from the Southern region who speak Spanish, the likelihood of having health insurance decreases when they have no personal earnings for manual labor occupations, whereas it increases for all occupations.	8.18%	52.12% 61.03%	7.74% ↓ 4.87% ↑	78.61% 91.58%	2.65% ↓ 1.98% ↑	0.126%
For White individuals from Texas or Southern region, who were born in USA, the likelihood of having health insurance decreases when they have no personal earnings for manual labor occupations, whereas it increases for all occupations.	12.21%	71.86% 80.52%	5.20% ↓ 2.8% ↑	78.61% 91.58%	2.65% ↓ 1.98% ↑	0.080
Among White males from the Southern region, the likelihood of having health insurance decreases when they have no personal earnings for manual labor occupations, whereas it increases for all occupations.	18.00%	67.56% 74.20%	4.19% ↓ 2.68% ↑	78.61% 91.58%	2.65% ↓ 1.98% ↑	0.068
Among natives who were born in USA, the likelihood of having health insurance increases when they didn't report absence from work and didn't attend school in the last 3 months for manual labor occupations, whereas it decreases for all occupations.	75.67%	84.11% 88.97%	4.95% ↑ 1.91% ↓	78.61% 91.58%	5.47% ↑ 0.50% ↓	0.068

Table 14: Brute-Force results for the ACS dataset.

Disparity Explanation	Support	Likelihood of feeling nervous frequently				Δ
		Subpopulation		Global (XInsight)		
		Average	CATE	Average	CATE	
For individuals who never married, are from the Southern region, don't have a doctor's recommendation to exercise, and aren't diagnosed with Diabetes, the likelihood of feeling nervous frequently decreases less for males compared to non-males if they do not smoke currently.	5.78%	46.08% 41.71%	13.06% ↓ 20.73% ↓	37.58% 45.10%	3.39% ↓ 3.94% ↓	0.076
Among White individuals who never married, are under 29, and don't have Asthma or Diabetes, the likelihood of feeling nervous frequently decreases less for males than non-males if they are uninsured for the whole year.	8.41%	49.20% 48.90%	14.16% ↓ 18.48% ↓	37.58% 45.10%	7.02% ↓ 4.86% ↓	0.043
For individuals who never married, don't have a recommendation from the doctor to exercise, and were born in USA, the likelihood of feeling nervous frequently increases more for males compared to non-males if they have private insurance.	14.25%	45.90% 45.70%	10.84% ↑ 7.51% ↑	37.58% 45.10%	4.63% ↑ 5.76% ↑	0.033
For white individuals who never married, don't have doctor's recommendation to exercise, and aren't diagnosed with Asthma, the likelihood of feeling nervous frequently decreases less for males compared to non-males if they are uninsured the whole year .	9.74%	47.25% 46.64%	12.52% ↓ 14.09% ↓	37.58% 45.10%	7.02% ↓ 4.86% ↓	0.015
For individuals aged between 30–42, and aren't diagnosed with Diabetes, the likelihood of feeling nervous frequently increases more for males compared to non-males if they have health insurance.	19.64%	43.93% 43.57%	7.47% ↑ 6.47% ↑	37.58% 45.10%	2.34% ↑ 4.54% ↑	0.010

Table 15: Brute-Force results for the MEPS dataset.