frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 09 December 2024
DOI 10.3389/fgene.2024.1436947

Check for updates

OPEN ACCESS

EDITED BY
Sisi Ma,
University of Minnesota, United States

REVIEWED BY
Shuai Zeng,
University of Missouri, United States
Xinpeng Shen,
University of Minnesota Twin Cities, United States

*CORRESPONDENCE
Dominik Heider,
dominik.heider@uni-muenster.de

†These authors share last authorship

RECEIVED 22 May 2024
ACCEPTED 20 November 2024
PUBLISHED 09 December 2024

CITATION
Ribeiro AH, Crnkovic M, Pereira JL, Fisberg RM, Sarti FM, Rogero MM, Heider D and Cerqueira A (2024) AnchorFCI: harnessing genetic anchors for enhanced causal discovery of cardiometabolic disease pathways. *Front. Genet.* 15:1436947. doi: 10.3389/fgene.2024.1436947

COPYRIGHT
© 2024 Ribeiro, Crnkovic, Pereira, Fisberg, Sarti, Rogero, Heider and Cerqueira. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

AnchorFCI: harnessing genetic anchors for enhanced causal discovery of cardiometabolic disease pathways

Adèle H. Ribeiro[1], Milena Crnkovic[2], Jaqueline Lopes Pereira[3], Regina Mara Fisberg[3], Flavia Mori Sarti[4], Marcelo Macedo Rogero[3], Dominik Heider[1*†] and Andressa Cerqueira[2†]

[1]Institute of Medical Informatics, University of Münster, Münster, Germany, [2]Department of Statistics, Federal University of São Carlos (UFSCar), São Carlos, Brazil, [3]Department of Nutrition, School of Public Health, University of São Paulo, São Paulo, Brazil, [4]School of Arts, Sciences and Humanities (EACH), University of São Paulo, São Paulo, Brazil

Introduction: Cardiometabolic diseases, a major global health concern, stem from complex interactions of lifestyle, genetics, and biochemical markers. While extensive research has revealed strong associations between various risk factors and these diseases, latent confounding and limited causal discovery methods hinder understanding of their causal relationships, essential for mechanistic insights and developing effective prevention and intervention strategies.

Methods: We introduce anchorFCI, a novel adaptation of the conservative Really Fast Causal Inference (RFCI) algorithm, designed to enhance robustness and discovery power in causal learning by strategically selecting and integrating reliable anchor variables from a set of variables known not to be caused by the variables of interest. This approach is well-suited for studies of phenotypic, clinical, and sociodemographic data, using genetic variables that are recognized to be unaffected by these factors. We demonstrate the method's effectiveness through simulation studies and a comprehensive causal analysis of the 2015 ISA-Nutrition dataset, featuring both anchorFCI for causal discovery and state-of-the-art effect size identification tools from Judea Pearl's framework, showcasing a robust, fully data-driven causal inference pipeline.

Results: Our simulation studies reveal that anchorFCI effectively enhances robustness and discovery power while handles latent confounding by integrating reliable anchor variables and their non-ancestral relationships. The 2015 ISA-Nutrition dataset analysis not only supports many established causal relationships but also elucidates their interconnections, providing a clearer understanding of the complex dynamics and multifaceted nature of cardiometabolic risk.

Discussion: AnchorFCI holds significant potential for reliable causal discovery in complex, multidimensional datasets. By effectively integrating non-ancestral

knowledge and addressing latent confounding, it is well-suited for various applications requiring robust causal inference from observational studies, providing valuable insights in epidemiology, genetics, and public health.

KEYWORDS

causal discovery, explainability, RFCI, genetic anchors, unfaithfulness, partial ancestral graphs, causal effect identification, cardiometabolic risk factors

1 Introduction

Cardiometabolic diseases, including cardiovascular disease and metabolic syndrome, are major global health concerns. They significantly contribute to the global burden of disease and mortality, impacting not only individual health but also societies, health systems, and economies on a global scale (Miranda et al., 2019; Rao, 2018). Numerous studies have identified a range of risk factors that contribute to the development and progression of cardiometabolic diseases. These factors arise from a complex interplay of lifestyle choices, genetic predispositions, and pre-existing conditions such as obesity, hypertension, inflammation, insulin resistance, type 2 diabetes, and dyslipidemia (Valenzuela et al., 2023; Kälsch et al., 2015; Wilson and Meigs, 2008).

Understanding the interconnections among these factors in a causal manner is essential for gaining insights into the mechanisms at play and accurately assessing an individual's risk profile. This knowledge empowers clinicians to make informed decisions and aid researchers in the development of interventions and prevention strategies that are both effective and tailored to the specific needs of each patient.

Traditionally, causality has been established through experimental studies, following the Bradford-Hill considerations on causality (Höfler, 2005). However, conducting controlled experiments on humans presents significant challenges and ethical limitations, especially when addressing long-term effects and interventions that may necessitate lifestyle changes or carry health risks for participants. In this work, we aim to identify the causal relationships among various cardiometabolic risk factors observed in the Health Survey of São Paulo with a Focus on Nutrition Study (2015 ISA-Nutrition), a comprehensive epidemiological study conducted in São Paulo, Brazil (Fisberg et al., 2018). Observational studies are invaluable tools for gaining insights into the complexities of the real world, but they face critical challenges due to biases. Significant statistical associations among variables do not always indicate causality; they may, in fact, be entirely spurious, originating from confounding factors or other issues such as selection bias. This complexity emphasizes the need for rigorous methodologies to address biases in causal inference from observational data.

For our analysis, we consider Judea Pearl's framework of causation (Pearl, 2000a). This framework represents a seminal advancement in data analysis, enabling us to move beyond mere correlations and discern causal relationships directly from observational data, thus reflecting the rigor typically associated with controlled experiments. Causality is articulated in an intuitive manner: a variable X is considered a cause of another variable Y if an intervention on X (e.g., setting it to a specific value, $X = x$) results in an expected change in Y. Notably, certain causal discovery algorithms are capable of identifying, despite hidden confounding, the graphical structure of the class of causal models that may have generated the observed data (Zhang, 2008b). These algorithms, complemented by emerging tools for effect identification from their outputs (Maathuis and Colombo, 2015; Perkovi et al., 2018; Jaber et al., 2022), have paved the way for a powerful, data-driven approach to causal inference.

As our primary methodological contribution, we propose the anchor Fast Causal Inference (anchorFCI) algorithm, designed to robustly uncover causal relationships among a set of variables by leveraging an additional set comprising only non-ancestors (non-causes) of the variables of interest. This is achieved by strategically selecting reliable anchors and integrating knowledge of their non-ancestral relationships into the conservative Really Fast Causal Inference (RFCI) algorithm (Colombo et al., 2012), renowned for its effectiveness and robustness in scenarios with latent confounding and limited sample sizes. Specifically, anchorFCI identifies as reliable anchors those variables from the additional set that are significantly associated with the variables of interest and form unambiguous triplets during the conservative orientation phase of the algorithm. Then, it incorporates knowledge of their non-ancestral relationships as outlined in Section 2.4.2. AnchorFCI not only enforces the appropriate arrowheads but also utilizes an adapted skeleton phase that prevents conditional independence tests between anchor variables conditioned on non-anchor variables. This adaptation is vital for preserving the integrity of the encoding of the conditional independencies implied by the final graph while reducing the number of required conditional independence tests, thereby enhancing the algorithm's scalability and robustness.

Our proposed approach is particularly well-suited for causal discovery among phenotypes, clinical factors, and sociodemographic variables when information on genetic variants, such as single nucleotide polymorphisms (SNPs), is available. This is because we can leverage the well-established understanding that genotypes are not influenced by any of the non-genetic variables. It is crucial to emphasize that, while this approach shares some similarities with Mendelian randomization principles (Davey Smith and Hemani, 2014; Ribeiro et al., 2016), it does not assume that SNPs identified as anchors are valid instruments (de Leeuw et al., 2022). Rather, the anchorFCI algorithm solely relies on conditional independence tests to identify genetic anchor variables and uncover causal relationships.

In analyzing the 2015 ISA-Nutrition dataset, our approach has successfully unveiled the causal connections among all examined phenotype and clinical variables, by leveraging SNPs identified in Genome-Wide Association Studies (GWAS). Furthermore, by employing effect identification tools, we have successfully identified and estimated the causal effects associated to all uncovered causal relationships. The results corroborate numerous

well-established relationships, while also providing a deeper understanding of the intricate network of connections among various cardiometabolic risk factors. Additionally, they demonstrate the potential of robust data-driven causal inference methods in addressing complex and multifactorial diseases, paving the way for the development of more effective interventions and treatments.

2 Materials and methods

2.1 Study design and data collection

Our analysis focused on 681 individuals who were not genetically closely related, selected from a total of 841 participants in the *2015 Health Survey of São Paulo with a Focus on Nutrition Study* (2015 ISA-Nutrition) – a subset of the cross-sectional population-based *2015 Health Survey of São Paulo* (2015 ISA-Capital). The primary goal of the 2015 ISA-Nutrition study is to investigate the relationships between lifestyle choices, biochemical markers, and genetics in the development of cardiometabolic diseases among residents of São Paulo city, the largest Brazilian city located in the Southeastern region of the country. For more comprehensive information about the study, refer to Fisberg et al. (2018).

The survey targeted a probabilistic sample of individuals age ≥ 12, residing in permanent households within the urban area of São Paulo city, excluding those who were pregnant or lactating. The sampling process involved two stages with stratification by clusters (urban census tracts and households) to ensure representative coverage at the population level. During the year of 2015, ISA-Capital collected sociodemographic data, information regarding the use of health services, lifestyle, and other relevant information through a structured questionnaire administered in the households by trained interviewers (Alves et al., 2018). Anthropometric data, blood pressure measurements, and blood samples were collected by trained nurses during a second visit to the participant's household. Detailed protocols for these measurements are also available in Fisberg et al. (2018).

Genomic DNA was obtained from peripheral blood samples extracted by automated method. SNPs were assessed using the Axiom™ 2.0 Precision Medicine Research Array in the Thermo Fisher Scientific Laboratory (Affymetrix Inc., Santa Clara, CA).

This study has been conducted according to the principles expressed in the Declaration of Helsinki and was approved by the Ethics Committee on Research of the School of Public Health of the University of São Paulo (#30848914.7.0000.5421). All participants authorized their genotyping and signed a written informed consent/assent before entering the study and, if they were adolescents, also their proxies. The data and samples were anonymized after collection.

2.2 Sample selection and description of the variables

The 2015 ISA-Nutrition study incorporates genetic data from 841 individuals, with some being relatives. Notably, conducting causal analysis on samples from genetically or familial-related individuals necessitates careful consideration of the underlying dependence structure among them (Ribeiro and Soler, 2020). Since addressing this issue is beyond the scope of this study, we excluded individuals who might share a parental relationship, resulting to a final sample size of 681 individuals. Specifically, the genomic relationship matrix (GRM) was computed for all 841 individuals and our analysis was limited to individuals with a genetic distance of ≤ 0.125, indicating relationships beyond second-degree relatives.

The dataset comprises a diverse array of sociodemographic, clinical, and genetic data. Additionally, it includes principal components of global ancestry allowing for adjustment for population stratification effects. This adjustment is particularly crucial in highly admixed populations such as the Brazilian population. The principal components of global ancestry are estimated using a larger set of SNPs across the genome of the ISA-Nutrition participants and the 1,000 Genome Project reference data, utilizing the software PLINK 2.0 and R (`SNPRelate` package to control for population stratification). More details of genetic evaluation of ISA-Nutrition data are available in Pereira et al. (2024).

In our analysis, we regarded sex, age, and the first two components of global ancestry as standard covariates. Additionally, we selected variables representing lifestyle factors, biochemical markers, and health conditions acknowledged as pertinent by domain experts. Moreover, SNP data was utilized in identifying genetic anchors crucial for causal analysis among the phenotypic variables.

The variables referring to lifestyle factors, biochemical markers, and health conditions included in the analysis are described below:

2.2.1 Obesity

Obesity is defined based on the body mass index (BMI), given as weight (kg)/height $(m)^2$. An adult is obese if their BMI ≥ 30 kg/m^2 (WHO Consultationm, 2000). In adolescents, obesity is identified when their BMI-for-age surpasses two standard deviations (2SD) above the mean, which, for a 19-year-old, corresponds to a BMI of 30 kg/m^2 (Onis et al., 2007).

2.2.2 Type 2 diabetes (T2D)

It is considered positive if fasting blood glucose is > 126 mg/dL or if medication for T2D (indicated by the binary variable Med_T2D) is being used, which includes hypoglycemic agents and/or insulin therapy (Cobas et al., 2022).

2.2.3 Hypertension (HTN)

For adults, hypertension is diagnosed when systolic blood pressure is elevated (≥ 140 mmHg), diastolic blood pressure is elevated (≥ 90 mmHg), or if antihypertensive drugs (indicated by the binary variable Med_HTN) are being used (Barroso et al., 2021). For adolescents aged 12 and 13 years, the cutoff points for systolic or diastolic blood pressure are defined as the 95th percentile based on sex, age, and height. For individuals aged 14–19 years, the cutoffs were set at systolic blood pressure (SBP) ≥ 130 mmHg or diastolic blood pressure (DBP) ≥ 80 mmHg (Flynn et al., 2017).

2.2.4 C-reactive protein (CRP)

The concentration of C-reactive proteins in the blood, obtained through a blood test, serves as a biomarker of inflammation in the

body. It is measured in milligrams per deciliter (mg/dL), with the normal range typically falling below 1 mg/dL (Ridker, 2003).

2.2.5 Homeostatic model assessment of insulin resistance (HOMA-IR)

HOMA-IR is calculated using the formula: fasting plasma insulin (μU/mL) × fasting plasma glucose (mmol/L)/22.5. The cutoff point for HOMA-1R, proposed by Geloneze et al. (2009) and validated for the Brazilian population, is set at 2.71 (Golbert et al., 2019).

2.2.6 Triglycerides (TGL), LDL cholesterol (LDLc), and HDL cholesterol (HDLc)

They are all measured using a colorimetric enzymatic method with reagents from Cobas–Roche Diagnostics GmbH (Mannheim, Germany). The normal ranges are typically as follows: TGL ideally below 150 mg/dL (adults) and 130 mg/dL (adolescents), LDLc ideally below 160 mg/dL (adults) and 130 mg/dL (adolescents), and HDLc ideally above 40 mg/dL (men), 50 mg/dL (women), and 45 mg/dL (adolescents) (Précoma et al., 2019; Giuliano et al., 2005).

2.2.7 Dyslipidemia (DLP)

Dyslipidemia is classified as positive if the individual is taking lipid-lowering medication (indicated by Med_DLP) or if any of the following conditions are met: elevated LDLc levels ($\geq$ 160 mg/dL for adults and $\geq$ 130 mg/dL for adolescents), elevated TGL levels ($\geq$ 150 mg/dL for adults and $\geq$ 130 mg/dL for adolescents), or low HDLc levels (men $<$ 40 mg/dL, women $<$ 50 mg/dL, and $<$ 45 mg/dL for adolescents) (Précoma et al., 2019).

2.2.8 Physical activity

A binary variable indicating whether the individual meets the total physical activity recommendations across four domains–work, domestic activities, transportation, and leisure–as outlined in the *2010 Global Recommendations on Physical Activity for Health* (World Health Organization, 2010).

Additionally, variables indicating use of Medication for T2D (Med_T2D), Medication for HTN (Med_HTN), and Medication for DLP (Med_DLP) were included in the analysis.

2.3 Genome-wide association study (GWAS)

Within the database of 681 unrelated individuals, we conducted a quality control process for genotypes, excluding SNPs with a minor allele frequency (MAF) of less than 5% or a Hardy-Weinberg equilibrium p-value of less than 1×10^{-5}. This process led to the removal of a total of 474,649 SNPs, resulting in a final dataset of 330,656 SNPs.

To conduct the GWAS, we included age, sex, and the square of age as covariates to separate genetic effects from the influence of these individual characteristics. Additionally, we included ancestry as a covariate using the first two estimated principal components of global ancestry to control for population structure and reduce false positive associations. We conducted a single SNP-GWAS using additive logistic regression to evaluate the association between genetic predictors and the binary phenotypic traits of interest: obesity, HTN, T2D, and DLP. For the continuous variables HDLc and CRP, we applied linear regression on their log-transformed values. In all regression models, we used hypothesis tests to determine the significance of each SNP, with a null hypothesis stating that there is no association between the SNP and the studied trait (the regression parameter associated with the SNP is equal to zero), and an alternative hypothesis stating that there is an association. To account for the increase in type I error due to multiple testing, we selected SNPs with at least genome-wide suggestive association (p-value $\leq 10^{-5}$).

2.4 AnchorFCI: causal discovery with non-ancestral knowledge

Causal discovery algorithms are increasingly being employed to elucidate, from observational data, the nature of statistical associations among variables. They can distinguish between purely spurious associations, which arise from the shared influence of other variables (referred to as confounders), and relationships that are truly causal, sometimes even elucidating their direction.

Among the existing algorithms, the Fast Causal Inference (FCI) (Spirtes et al., 2001), combined with the complete set of orientation rules by Zhang (2008b), stands out for its rigorous foundational principles and minimal reliance on assumptions compared to other methods. Since the introduction of the FCI, a range of strategies and adaptations have emerged to tackle scalability and robustness in scenarios of limited data. These include the anytime FCI by Spirtes (2001), the RFCI by Colombo et al. (2012), and their conservative counterparts (Colombo and Maathuis, 2014). Crucially, these algorithms *do not require causal sufficiency*, demonstrating their effectiveness in managing latent confounding. This capability makes FCI-like algorithms highly suitable for analyzing real-world datasets.

Relying solely on a reliable conditional independence test, the FCI and its variants learn a Partial Ancestral Graph (PAG) representing the class of all causal models that entail the set of observed conditional independencies, referred to as the Markov equivalence class (MEC). In a PAG, tails and arrowheads represent, respectively, ancestral (causal) and non-ancestral (non-causal) relationships common to all models within the most plausible MEC. A circle ("o") denotes a non-invariant edge mark, indicating that within the MEC, there is a model where the edge mark is a tail and another model where the same edge mark is an arrowhead. Remarkably, the output is ensured to be asymptotically sound and complete, even in scenarios involving unobserved confounding and selection bias.

Scalability and integration of background knowledge are major challenges in causal discovery under latent confounding. While increasing the number of variables in the graph can boost discovery power, the reliability of the inferences often decreases due to statistical instabilities. Moreover, merely enforcing edge marks can lead to incorrect downstream orientations and violations of the equivalence class representation if the algorithm is not properly adapted. Currently, no complete strategy exists for integrating general background knowledge into the FCI framework. This integration has only been explored in specific contexts, such as studies with time series data (Gerhardus and Runge, 2020) or certain

types of knowledge, such as tired (known causal order) or local knowledge (Andrews et al., 2020; Wang et al., 2022).

In this work, we introduce *anchorFCI*, a novel adaptation of the conservative RFCI designed to strategically identify reliable anchors and effectively integrate them with their known non-ancestral relationships. As detailed in Section 2.4.2, anchorFCI operates on two variable sets: the first contains the variables of interest, while the second comprises variables that are not caused by any from the first. It begins by identifying *reliable anchors*, defined as variables V_i from the second set that are significantly associated with a variable V_j from the first set and form unambiguous triplets during the algorithm's conservative orientation phase. Next, it performs an adapted skeleton phase that ensures a consistent selection of d-separators while remaining order-independent, as it is based on the stable skeleton algorithm by Colombo and Maathuis (2014). Finally, it enforces arrowheads on the appropriate edges $V_i \circ\!\!\rightarrow V_j$ according to the established non-ancestral relationship between the two variable sets. Notably, strategies for managing conflicting orientations and limiting conditioning set sizes, as detailed in Section 2.4.1, are integral features of the RFCI and, thus, can be easily adjusted through parameters in the anchorFCI function. The algorithm is publicly available as an R package on GitHub at https://github.com/adele/anchorFCI and it leverages the RFCI implementation from the `pcalg` R package (Kalisch et al., 2024).

We apply the *anchorFCI* algorithm with strategies to enhance robustness to uncover causal relationships among the 14 phenotypic and clinical variables outlined in Section 2.2. The algorithm identifies anchors from SNPs that are significantly associated with phenotypes of interest through GWAS, and incorporates knowledge of their non-ancestral relationships. This approach not only improves robustness but also boosts discovery power by facilitating the identification of additional causal relationships. To account for confounding effects related to sex, age (both original and squared), and the first two principal components of global ancestry, these covariates will be included in all necessary tests of conditional independence.

2.4.1 Strategies for addressing conflicts and inconsistencies arising from unfaithfulness

Despite the significant potential of causal discovery algorithms, they present considerable challenges in terms of scalability to larger graphs and robustness to statistical errors. Firstly, the number of potential causal structures grows super-exponentially with the number of variables, rendering the problem computationally NP-hard (Chickering et al., 2004). Achieving scalability to large graphs, particularly in scenarios involving latent confounding, necessitates reliance on model assumptions. These may include enforcing sparsity and constraining the size of conditioning sets (Claassen et al., 2013).

Moreover, the majority of the existing algorithms, including the FCI and its variants, rely on the assumption of *faithfulness*, which posits that the independencies inferred from data accurately represent the true underlying model. Notably, any falsely identified independencies can lead to conflicting orientations and inconsistencies in the implied conditional independencies (Zhang and Spirtes, 2008; Zhalama et al., 2017). This poses a significant challenge for real-world applications, particularly when dealing with limited sample sizes, as statistical tests may lack sufficient power to accurately identify potentially weak or noisy associations. This issue becomes more pronounced when dealing with larger graphs (e.g., 10 or more variables), exemplifying the curse of dimensionality, as it leads to an increased number of conditional independence tests and a significant decrease in statistical power as the conditioning set size grows.

To minimize issues arising from unfaithfulness, we employ the following two strategies:

- Conservative edge orientations with majority rule: as proposed by Colombo and Maathuis (2014), edge orientations are strictly conducted conservatively, meaning that they exclusively rely on triplets that have been previously determined as unambiguous. To assess the ambiguity of an unshielded triplet (i.e., three variables where the first and second, and the second and third, are adjacent, but not the first and third), additional conditional independence tests are performed. These tests aim to detect errors in determining whether the second variable belongs or not to the set that renders the first and third variables independent of each other, potentially due to instances of unfaithfulness. With the majority rule approach, the decision is based on the majority of the involved conditional independence tests. In the event of a tie, the triplet is deemed ambiguous, and no orientation is conducted.
- Constrained conditioning set size: as pointed out by Spirtes (2001) and Colombo et al. (2012), the accuracy of the FCI significantly decreases when relying on independencies conditional on a large set of variables, as, in such cases, statistical errors due to the low power of the tests become inevitable, especially with small sample sizes. They also demonstrate that restricting the size of conditioning sets in the independence tests does not compromise the correctness of the output PAG; it may only lead to a less informative PAG. In light of this result, the RFCI not only accepts a constraint on the size of the conditioning sets, but also employs a procedure that requires fewer conditional independence tests than the FCI. The authors demonstrated the RFCI's consistency in sparse scenarios, showing that by avoiding statistical tests with the least power (i.e., those with relatively large conditioning sets), the results become notably more reliable, particularly for small sample sizes.

2.4.2 Identification of reliable anchors and integration of non-ancestral knowledge

Applying the RFCI combined with strategies aimed at enhancing robustness (e.g., conservative edge orientations) can contribute to more reliable results. However, these strategies often result in PAGs with numerous underdetermined edge marks (i.e., circles).

To enhance robustness and discovery power, we propose extending the RFCI algorithm to utilize two partially ordered variable sets: the first contains the variables of interest, while the second comprises variables known not to be caused by any variable in the first set. This structure is ideal for datasets with non-genetic variables, such as phenotypic, clinical, or sociodemographic data, in the first set, and genetic variables in the second. This novel algorithm, referred to as *anchorFCI*, introduces three key adaptations:

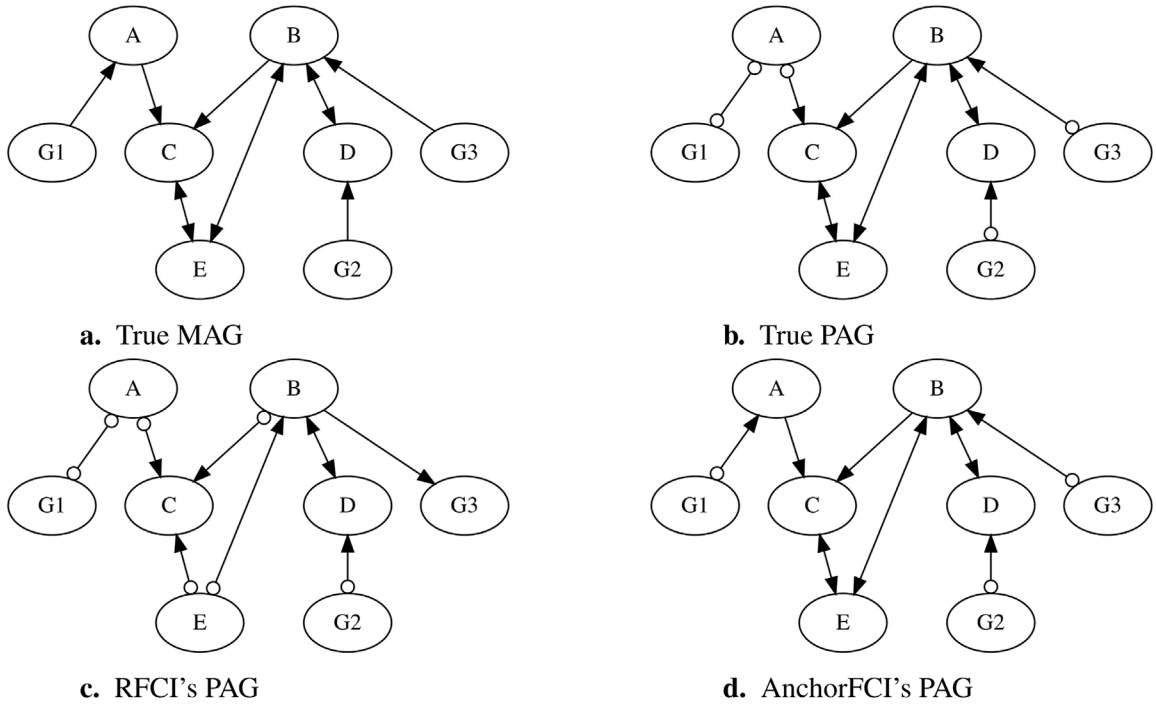

FIGURE 1
Comparison of PAGs inferred by RFCI and anchorFCI using a simulated dataset. The dataset was generated based on the MAG in **(A)**, structured as $G_1, G_2, G_3 \prec A, B, C, D, E$. The true PAG is presented in **(B)**, while the PAGs inferred by RFCI and anchorFCI are shown in **(C, D)**, respectively.

1. Identification of robust anchors: These are variables from the second set (e.g., SNPs) that are identified as significantly associated with variables from the first set (e.g., phenotypes) and, upon integration into the graph, form triplets that are unambiguous according to the majority rule.
2. Skeleton inference with an adapted search for minimal d-separating sets: In the initial phase of the algorithm, known as the adjacency phase, the model's skeleton is inferred through a series of conditional independence tests. Whenever a pair of variables is found to be conditionally independent given a set of other variables, such separating set is considered minimal and the edge connecting them is removed. In this phase, it is essential to prevent the algorithm from performing independence tests between two variables from the second set (e.g., two SNPs) conditional on any variable from the first set (e.g., a phenotype). Notably, this restriction does not compromise the correctness or completeness of the algorithm. This can be demonstrated through a direct application of (Tian et al., 1998, Theorem 2), which states that if a set $\mathbf{Z}$ d-separates two variables X and Y, we can get a smaller d-separator $\mathbf{Z}'$ by removing from $\mathbf{Z}$ all nodes that are not ancestors of X or Y. This implies that, since the first set consists entirely of non-ancestors of the second set, any faithful minimal separating set for a pair of variables from the second set will necessarily exclude variables from the first set.
3. Enforcing known arrowheads: During the orientation phase of the algorithm, edge-mark inference rules (R0 to R10, as detailed in Zhang (2008b)) are sequentially applied until they can no longer be utilized. Throughout this phase, we ensure that every edge connecting a variable V_i from the first set (e.g., a SNP) and a variable V_j from the second set (e.g., a phenotype) is oriented with an arrowhead pointing towards V_j. In other words, these edges will take the form $V_i * \!\!\rightarrow V_j$, where $*$ denotes any edge mark that the algorithm may learn. Incorporating this type of background knowledge is straightforward and can be achieved simply by applying the rules once the known edge marks are enforced and then checking the validity of the PAG as an ancestral graph (i.e., no cycles or almost cycles). Notably, the integration of non-ancestral knowledge in the context of time-series data (Gerhardus and Runge, 2020) showcases the potential to enhance both robustness and learning capabilities in causal discovery.

Notably, when applied to a dataset comprising SNPs and phenotype variables, the resultant PAG from the anchorFCI algorithm will exclusively include edges between an SNP and a phenotype falling into one of the following types: SNP $\leftrightarrow$ Phenotype (indicating purely spurious association), SNP $\rightarrow$ Phenotype (indicating a causal SNP), or SNP $\circ\!\!\rightarrow$ Phenotype (indicating there is not sufficient evidence to determine whether the SNP is a cause or only spuriously associated with the phenotype).

By integrating inherently unambiguous orientations, anchorFCI not only prevents discrepancies with established non-ancestral knowledge, potentially induced by violations of faithfulness, but also boosts discovery power by facilitating the conservative identification of additional causal relationships.

To further illustrate how anchorFCI can achieve greater robustness and accuracy compared to RFCI, particularly in

limited data settings where the faithfulness assumption is likely to be violated, we selected a dataset from our simulation study, as detailed in Section 2.4.4. The dataset consists of 1,000 samples, generated based on the Maximal Ancestral Graph (MAG) shown in Figure 1A. The variables $\{G_1, G_2, G_3\} \prec \{A, B, C, D, E\}$, with $\prec$ denoting precedence in the partial ordering. The corresponding true PAG, obtained without using the partial order information, is shown in Figure 1B.

When applying the conservative RFCI algorithm with the majority rule, we obtain the PAG in Figure 1C. The collider triplets $\langle A, C, B\rangle$, $\langle E, B, D\rangle$, $\langle A, C, E\rangle$, and $\langle G_2, D, B\rangle$ are identified as unambiguous and correctly oriented. The triplets $\langle G_1, A, C\rangle$ and $\langle G_3, B, C\rangle$ are identified as unambiguous non-collider triplets, however, without the partial order information, their orientations cannot be determined. The triplet $\langle C, B, D\rangle$ is marked as ambiguous. After applying the remaining orientation rules, RFCI incorrectly orients $B \rightarrow G_3$, violating the known, but unused, information that $G_3 \prec B$. The resulting PAG has a Structural Hamming Distance (SHD) of five from the true PAG, reflecting the number of incorrect orientations.

In contrast, when applying anchorFCI, we obtain the PAG in Figure 1D, which not only accurately captures all the oriented edges from the true PAG but also correctly identifies additional orientations from the original MAG. First, variables G_1, G_2, and G_3 are all identified as reliable anchors, as they form unambiguous triplets $\langle G_1, A, C\rangle$, $\langle G_2, D, B\rangle$, and $\langle G_3, B, C\rangle$. Then, after applying the adapted skeleton algorithm, anchorFCI integrates the known non-ancestral relationships by orienting $G_1 \circ\!\!\rightarrow A$, $G_2 \circ\!\!\rightarrow D$, and $G_3 \circ\!\!\rightarrow B$. This not only avoids the mistakes made by RFCI but also enables the accurate orientation of the edge $B \rightarrow C$ and of the collider $\langle C, E, B\rangle$. Notably, anchorFCI also infers the tail on $A \rightarrow C$, indicating that A is a definite cause of C, thereby learning a representation that goes beyond the MEC.

2.5 Conditional independence tests for mixed data

As stated before, the FCI and its variants identify ancestral and non-ancestral relationships by employing a combination of conditional independence tests and orientation rules.

To test conditional independence between binary (obesity, T2D, HTN, DLP, physical activity, and medication variables), continuous (CRP, HOMA-IR, TGL, LDLc, and HDLc), and multinomial (SNP) variables, we use the symmetric conditional independence test for mixed data proposed by Tsagris et al. (2018), available in the `MXM` R package (Lagani et al., 2017). This test utilizes linear or generalized linear regression, depending on the nature of the involved variables. Logistic regression is employed for binary outcomes, Gaussian linear regression for continuous outcomes, and multinomial log-linear regression for multinomial outcomes. To prevent departures from normality in tests involving continuous variables, we transformed them using the rank-based inverse normal transformation available in the `RNOmni` R package (McCaw, 2023). Missing values were imputed using *MissForest*, a non-parametric method for imputing mixed-type data sets by employing random forests (Stekhoven and Bühlmann, 2012). We utilized the implementation of the MissForest algorithm available in the `MissRanger` R package (Mayer and Mayer, 2019), setting the number of forests to 500.

Formally, Tsagris et al. (2018) test evaluates the conditional independence of two variables, V_i and V_j, given a set of variables $\mathbf{S}_{ij}$, by testing two null hypotheses: $H0_1$: $P(V_i|S_{ij}) = P(V_i|V_j, S_{ij})$ and $H0_2$: $P(V_j|S_{ij}) = P(V_j|V_i, S_{ij})$. The null hypothesis $H0_1$ is tested using a nested likelihood-ratio test comparing a reduced model (where V_i is regressed on $\mathbf{S}_{ij}$) against a full model (where V_i is regressed on both $\mathbf{S}_{ij}$ and V_j). Similarly, $H0_2$ is tested by reversing the roles of V_i and V_j. In general, the p-values p_1 and p_2 from the tests for $H0_1$ and $H0_2$, respectively, tend to be identical only asymptotically. To correct for any potential asymmetry in limited data scenarios, we adopt Tsagris et al. (2018) strategy of merging dependent p-values. Such method calculates the combined p-value as $\min(2 \times \min(p_1, p_2), \max(p_1, p_2))$ and has demonstrated superior learning accuracy when compared to alternative methods.

All tests include sex, age (original and squared), and the first two principal components of global ancestry in the conditioning set, ensuring comprehensive adjustment for these variables. Importantly, these variables stand out as special covariates because they are not caused by any other variables, and thus, conditioning on them can never introduce biases such as collider bias (Holmberg and Andersen, 2022).

Finally, all tests were conducted with a significance level set at 5%, without applying any correction for multiple comparisons. In contrast to association studies, which aim to validate statistical *dependencies*, causal discovery relies on establishing reliable statistical *independencies*. It is crucial to note that the non-rejection of the null hypothesis (indicating conditional independence) does not imply the acceptance of the alternative hypothesis (indicating conditional dependence). Any correction aimed at reducing the number of falsely identified associations may significantly increase the number of falsely identified independencies, thereby triggering a cascade of erroneous edge orientations. Conversely, as noted by Spirtes (2001); Colombo et al. (2012), false associations often prevent certain edge orientations, leading to a final PAG that, while less informative, still tends to uphold accuracy.

2.6 Comparative study of RFCI and AnchorFCI

To quantitatively assess the enhanced robustness and discovery power of anchorFCI compared to the conservative RFCI, we designed a simulation study. To capture diverse scenarios, we simulated 50 unique random MAGs with eight nodes, structured as $\{G_1, G_2, G_3\} \prec \{A, B, C, D, E\}$. Additionally, to account for varying degrees of unfaithfulness, we generated 30 datasets for each MAG, with sample sizes $N = 500, 1000, 5000, 10000$. The variables G_1, G_2, G_3 are modeled as discrete variables, each with three levels, following a multinomial distribution, while A, B, C, D, E are continuous variables, following a Gaussian distribution. The datasets were generated using the `simMixedDAG` R package (Lin, 2019), where each variable is modeled as a linear combination of its parents (including potential latent variables), with randomly assigned coefficients. The choice of distributions does not impact the comparison between the algorithms but was selected to reflect typical applications involving genotype and phenotype variables.

Since anchorFCI aims to enhance causal discovery among the variables of interest $\{A, B, C, D, E\}$ by selecting reliable anchors from $\{G_1, G_2, G_3\}$, we evaluate the algorithms using a score based on the Structural Hamming Distance (SHD) computed considering exclusively edges among the variables of interest. Given that anchorFCI can identify orientations beyond those in the true

MEC's PAG (i.e., the one obtained without using any partial order information), this score allows us to assess both the accuracy and informativeness of the inferred PAG. It is calculated as the difference between the SHD of the inferred PAG relative to the true MAG, and the SHD of the true MEC's PAG relative to the true MAG. Smaller scores indicate higher accuracy, with zero indicating the inferred PAG is as informative as the true MEC's PAG (i.e., the one obtained without using any partial order information), and negative values indicating the inferred PAG is more informative than the true MEC's PAG.

2.7 Causal effect identification and estimation

As previously discussed, a PAG represents a class of the most probable models based on the available observational data. Remarkably, all models in this class fit the data equally well, as they entail the same set of observed conditional independencies, rendering them statistically indistinguishable. Whenever ancestral and non-ancestral relationships are shared across all models within this equivalence class, they are represented as non-circle edge marks in the PAG. While this may suffice for a qualitative description of the relationships among the variables, a more comprehensive approach is required to quantify a causal effect.

For each directed edge inferred in the PAG, it is necessary to assess the identifiability of the corresponding causal effect. The identification of a causal effect is contingent upon its uniqueness–it is identifiable if and only if it is uniquely computable among all models within the equivalence class represented by the PAG, and utilizing the same expression solely based on observational (conditional) probabilities.

A necessary condition for identifying the causal effect of a variable X on a variable Y is that the edge $X \to Y$ is visible, indicating that X and Y do not share any latent causes This condition can be easily verified through a graphical criterion presented by Zhang (2008a).

The generalized backdoor criterion (Maathuis and Colombo, 2015) is regarded as one of the most straightforward effect identification graphical criteria. It states that the causal effect of a variable X on a variable Y can be identified from a PAG $\mathcal{P}$ if there exists a set $\mathbf{Z}$, comprising solely non-descendants of X, that blocks (in the sense of d-separation–see Pearl (2000b)) all definite-status backdoor (confounding) paths between X and Y in $\mathcal{P}$. In this case, the post-interventional probability distribution of a variable Y after an intervention that sets $X = x$, denoted $do(X = x)$, is expressed by:

$$P(Y = y|do(X = x)) = \int_{\mathbf{z}} P(Y = y|X = x, \mathbf{Z} = \mathbf{z})P(\mathbf{Z} = \mathbf{z})dz.$$

Such an integral can be estimated from N observational samples $\{y_i, x_i, \mathbf{z}_i\}_{i=1}^{N}$ as following:

$$\hat{P}(Y = y|do(X = x)) = \sum_{i=1}^{N} \hat{f}(x, \mathbf{z}_i),$$

where $\hat{f}(x, \mathbf{z}) = \hat{P}(Y = y|X = x, \mathbf{Z} = \mathbf{z})$ can be readily obtained through (generalized) regression analysis of Y on X and $\mathbf{Z}$. Note that if $\mathbf{Z} = \varnothing$ is admissible for backdoor adjustment, the interventional distribution simplifies to $P(Y = y|do(X = x)) = P(Y = y|X = x)$, which can also be readily obtained using (generalized) regression analysis of Y on X. If continuous response variables are (natural) log-transformed before regression analyses to normalize residual distributions, then estimates are subsequently back-transformed to the original scale using the improved Cox's method proposed by Olsson (2005). Effect sample sizes for each prediction are determined using the methodology by Thomassen et al. (2024).

In cases where the generalized backdoor criterion does not apply, alternative tools for causal effect identification and estimation from PAGs are available. These include the generalized adjustment criterion, introduced by Perkovi et al. (2018), and the complete causal calculus and (conditional) complete effect identification algorithm developed by Jaber et al. (2022).

3 Results

3.1 Descriptive analysis

In the database of 681 individuals, the proportions of gender are similar, with 53.59% male participants, and the mean age is 44.70 ± 23.37 years. The age groups also have similar proportions: 29.51% of adolescents (12–19 years old), 32.74% adults (20–59 years old), and 37.73% older adults (≥ 60 years old).

The proportions of binary health conditions vary: 23.2% of individuals are obese, 14.7% have T2D (with 70% of them taking medication for T2D), 45.8% have HTN (with 53.4% of them taking medication for HTN), and 66.7% have DLP (with 12.8% of them taking medication for DLP). Moreover, 68.8% of individuals fulfill the 2010 WHO recommendations for total physical activity (Bull et al., 2020). Summary statistics for the continuous variables in the study are provided in Table 1.

3.2 Genome-wide association study (GWAS)

We conducted a single SNP-GWAS to detect the association between each individual SNP and six phenotypic traits: obesity, HTN, T2D, DLP, CRP levels, and HDLc. Utilizing a p-value threshold indicative of genome-wide suggestion, set at 10^{-5}, we identified a total of ten SNPs.

Specifically, the Manhattan plots in Figure 2 reveal the following genetic associations: 1 SNP for obesity (rs41282114), 3 for HTN (rs726164, rs34500244, rs9354481), 1 for DLP (rs340643), 4 for HDLc (rs269029, rs17268691, rs6589567, rs4815295), and 1 for CRP levels (rs7577826). No genetic associations with T2D were found at this threshold. Table 2 presents a detailed summary of the SNPs identified for each of the phenotypic traits studied. Notably, no SNP was found to be common across all phenotypic traits. Observe that the SNPs associated with DLP, HDL, and CRP exhibit negative effects, indicating an association with a reduction in the respective studied traits. These identified SNPs are potential genetic markers for the respective conditions and may contribute to our understanding of the underlying genetic architecture. Additionally, they will undergo further analysis in the causal

TABLE 1 Summary statistics for the continuous variables in the study, including Minimum (Min.), Median, Mean, 1st and 3rd Quantiles (Qu.), Maximum, and total number of missing values (NA's).

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
CRP (mg/dL)	0.02	0.10	0.30	0.54	0.78	2.24	21
HOMA-IR	0.35	1.76	2.69	3.58	4.08	33.89	8
HDLc (mg/dL)	10.00	35.00	43.00	44.14	52.00	119.00	14
LDLc (mg/dL)	6.00	77.50	101.00	103.52	126.00	259.00	14
TGL (mg/dL)	10.00	73.00	100.00	120.22	139.50	1402.00	13

discovery phase to determine their potential as reliable anchors, thereby aiding in learning of causal relationships.

3.3 Causal discovery using the proposed AnchorFCI algorithm

3.3.1 Comparative study of RFCI and AnchorFCI

The results of the simulation study comparing the performance of RFCI and anchorFCI in learning the structure among the variables of interest are summarized in Table 3. The analysis reveals that anchorFCI consistently exhibits superior robustness and accuracy compared to RFCI. This is supported by statistically significantly lower scores for anchorFCI, with all p-values from a one-sided Wilcoxon test indicating very strong significance. Additionally, anchorFCI exhibits enhanced discovery power. Notably, it often achieves negative scores, indicating its ability to learn representations beyond the MEC even in scenarios with limited data. While both methods improve in performance with increasing sample sizes, anchorFCI consistently outperforms RFCI across all evaluated conditions.

For completeness, we provide a comparison of performance in learning the graphical structure among the variables of interest and the anchors in the Supplementary Material. As expected, anchorFCI significantly outperforms RFCI, thanks to its effective integration of partial order knowledge. Notably, while RFCI exhibits considerable instability in low-data regimes, anchorFCI consistently demonstrates superior robustness and accuracy.

3.3.2 Applying AnchorFCI to the 2015 ISA-nutrition dataset

Figure 3 shows the output PAG from our proposed anchorFCI algorithm. Conditional independence for mixed variables were conducted following the approach in Section 2.4.3, at a significance level of 5%. Furthermore, as outlined in Section 2.4.1, we enforced a limitation on the size of the conditioning sets, allowing for a maximum of 2 (two) variables. This choice is based on our exhaustive experimentation with the algorithm, which revealed that independencies conditional to larger sets led to the separation of certain pairs of dependent variables within the graph, thereby compromising the model's accuracy. Edge mark inference was conducted conservatively, meaning that edge orientations were established solely based on triplets identified as unambiguous according to the majority rule.

Among all SNPs identified as significantly associated with phenotypes (with p-value less than 10^{-5}), the anchorFCI algorithm exclusively selected those forming unambiguous triples in the inferred PAG, as determined by the majority rule. These include: rs41282114 (associated with Obesity), rs726164 and rs9354481 (associated with HTN), rs340643 (associated with DLP), and rs269029 and rs4815295 (associated with HDLc). Prior knowledge asserting those SNP variables cannot be caused by any of the other variables was incorporated in the learning process through the steps detailed in Section 2.4.2. Note that we excluded Med_DLP from the analysis due to numerous conditional independence tests raising errors, primarily because of the limited number of individuals who take medication for DLP in our sample.

In the resulting PAG, the relationships among all phenotypes and clinical variables have been fully characterized. Only a few edge marks in edges with SNPs remain undetermined (i.e., circles).

Directed edges represent definite ancestral (causal) relationships. Obesity emerges as a significant upstream causal factor for multiple conditions. It directly influences HTN, CRP levels–suggesting an impact on the body's inflammation levels –, and insulin resistance, as measured by HOMA-IR. Insulin resistance, inferred to be directly influenced by CRP, is considered a causal factor for T2D, TGL, and HDLc. TGL and HDLc are identified as causes of DLP. Furthermore, T2D appears to causally contribute to LDLc. The probabilities of undergoing medication for HTN (Med_HTN) and for T2D (Med_T2D) are, as expected, influenced by their respective conditions. Med_HTN appears to influence the likelihood of Med_T2D. Bidirected edges represent definite spurious associations, indicating no causal relationship in any direction. Notably, the model suggests that the relationship between LDLc and TGL exhibits this characteristic, thus solely attributable to the influence of latent confounders.

To assess the stability of the inferred relationships, we present the results from 50 bootstrap samples in the Supplementary Material. The following relationships were inferred in the respective percentages of bootstrap samples: Obesity → HOMA-IR (88%), TGL → DLP (84%), HTN → Medication for HTN (80%), TGL ↔ LDLc (64%), HDLc → DLP (58%), Medication for HTN → Medication for T2D (58%), HOMA-IR → T2D (54%). For certain relationships, however, the highest percentages correspond to non-adjacency, suggesting weak connections between CRP and TGL (52%), HOMA-IR and HDLc (42%), and Obesity and HTN (38%). For the remaining relationships, the edge types inferred with highest percentage correspond to those inferred from the original dataset, although such percentage did not exceed 50%.

It is essential to emphasize that the stability of any causal discovery algorithms is highly dependent on the number of variables in the graph, with performance significantly declining

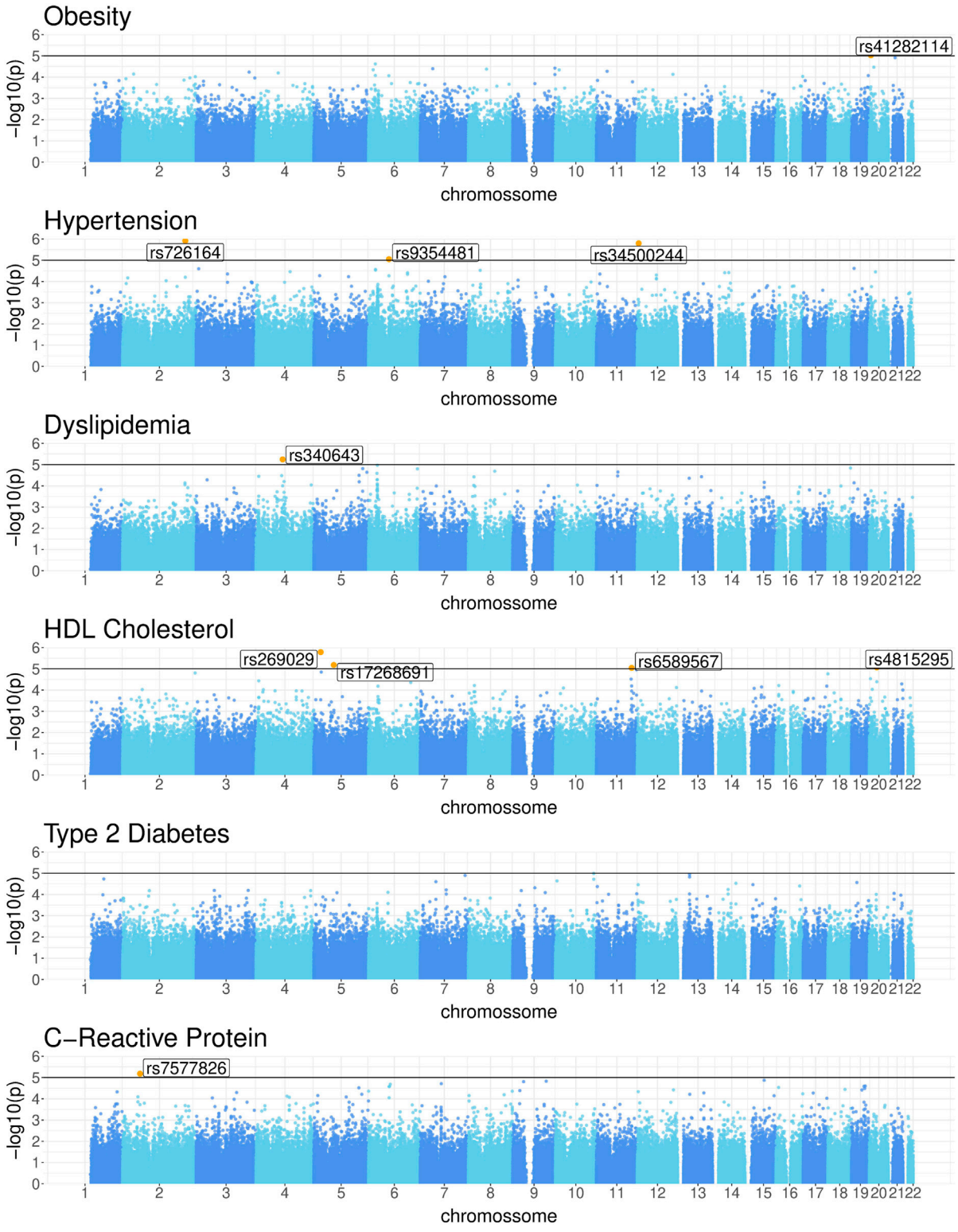

FIGURE 2
Manhattan plots for Obesity, Hypertension, Dyslipidemia, HDL cholesterol, Type 2 Diabetes, and C-Reactive Protein levels.

when the number of variables exceeds a handful. To illustrate this point, we present bootstrap results in the Supplementary Material based on only nine variables: Obesity, HOMA-IR, T2D, LDLc, HDLc, TGL, DLP, rs41282114, and rs4815295. In this scenario, the results are significantly more robust, as exemplified by the following relationships and respective percentages of bootstrap samples: Obesity → HOMA-IR (100%), TGL → DLP (90%), HOMA-IR → T2D (82%), and T2D → LDLc (54%).

TABLE 2 List of SNPs associated with Obesity, HTN, and DLP, HDL and CRP at a significance level of 10^{-5}, with their corresponding minor allele, MAF, chromosome (Chr.), linear regression coefficient (β), and p-value.

Trait	SNP	Minor allele	MAF	Chr.	β	p-value
Obesity	rs41282114	T	0.07	20	0.98	9.3×10^{-6}
HTN	rs726164	A	0.45	2	0.65	1.2×10^{-6}
	rs34500244	.	0.07	12	1.40	1.6×10^{-6}
	rs9354481	G	0.12	6	−0.62	9.0×10^{-6}
DLP	rs340643	A	0.36	4	−0.57	5.7×10^{-6}
HDLc	rs269029	G	0.29	5	−0.09	1.6×10^{-6}
	rs17268691	T	0.10	5	−0.13	6.6×10^{-6}
	rs6589567	A	0.12	11	−0.11	9.0×10^{-6}
	rs4815295	T	0.41	20	−0.08	8.5×10^{-6}
CRP	rs7577826	C	0.34	2	−0.30	6.6×10^{-6}

TABLE 3 Comparison of RFCI and anchorFCI performance in learning the structure among the variables of interest. Scores represent the difference between the SHD of the inferred PAG relative to the true MAG and the SHD of the true PAG relative to the true MAG.

N	RFCI		AnchorFCI			Diff	P-value
	[Min, max]	Mean ± SD	[Min, max]	Mean ± SD	#Anchors		
500	[2.00, 7.40]	5.33 ± 1.27	[−0.53, 6.30]	3.72 ± 1.39	1.84	1.62	8.76×10^{-193}
1,000	[1.47, 5.97]	4.27 ± 1.12	[−0.80, 4.43]	2.72 ± 1.28	2.19	1.55	4.94×10^{-204}
5,000	[0.80, 3.83]	2.65 ± 0.805	[−1.60, 2.83]	1.19 ± 1.10	2.64	1.47	9.63×10^{-227}
10,000	[0.60, 3.57]	2.21 ± 0.716	[−2.20, 2.40]	0.70 ± 0.97	2.73	1.50	8.07×10^{-236}

Values are averages across 50 randomly generated MAGs, each evaluated on 30 datasets with varying sample sizes N. #Anchors indicates the average number of selected reliable anchors by anchorFCI. Smaller scores indicate higher accuracy, with zero indicating performance equivalent to the true PAG and negative scores suggesting greater informativeness beyond the MEC. P-values are from a one-sided Wilcoxon test with alternative hypothesis that the average difference between RFCI and anchorFCI scores (Diff) is greater than 0.

Notably, if we had employed the original conservative RFCI with the SNP variables but without the adaptations proposed in the anchorFCI algorithm, not only would numerous orientations remain undetermined, but we would also see some phenotypes identified as causes of genotypes (e.g., DLP is learned as a cause of the SNP rs340643), contradicting the current understanding that genetic variables cannot be caused by phenotypes. For further details, please refer to the Supplementary Material.

3.4 Causal effect identification and estimation

All directed edges in Figure 3 of the PAG are free from latent confounding influence (as per the visibility graphical criterion by Zhang (2008a)), and their corresponding total causal effects are all identified using the generalized backdoor criterion (GBDC), by Maathuis and Colombo (2015). The estimation of interventional expectations and probabilities was conducted as detailed in Section 2.5. For coefficient estimates, associated statistics, and p-values of all regression models involved, please refer to the Supplementary Material.

3.4.1 Obesity → HTN

Figure 4A shows the point estimate and 95% confidence interval (CI) for the probability of HTN given an intervention on Obesity ($P(\text{HTN}|do(\text{Obesity}))$). The effect appears as unconfounded in the PAG, so the adjustment only considered the standard set of covariates, namely sex, age (original and squared), and the first two principal components of global ancestry. The non-overlapping intervals suggest a significant difference between these probabilities: for non-obese individuals, it is 0.38 with a 95% CI of [0.29, 0.5], while for obese individuals, it is 0.65 with a 95% CI of [0.52, 0.76].

3.4.2 Obesity → CRP

Figure 4B shows the point estimate and 95% CI for the expectation of CRP levels given an intervention on Obesity ($\mathbb{E}(\text{CRP}|do(\text{Obesity}))$). Since the effect appears unconfounded in the PAG, we adjusted solely for the standard set of covariates. The difference between these expectations is significant: for non-obese individuals, it is 0.35 with a 95% CI of [0.28, 0.44], while for obese individuals, it is 0.71 with a 95% CI of [0.53, 0.94].

3.4.3 Obesity → HOMA-IR

Figure 4C shows the point estimate and 95% CI for the expectation of HOMA-IR levels given an intervention on Obesity

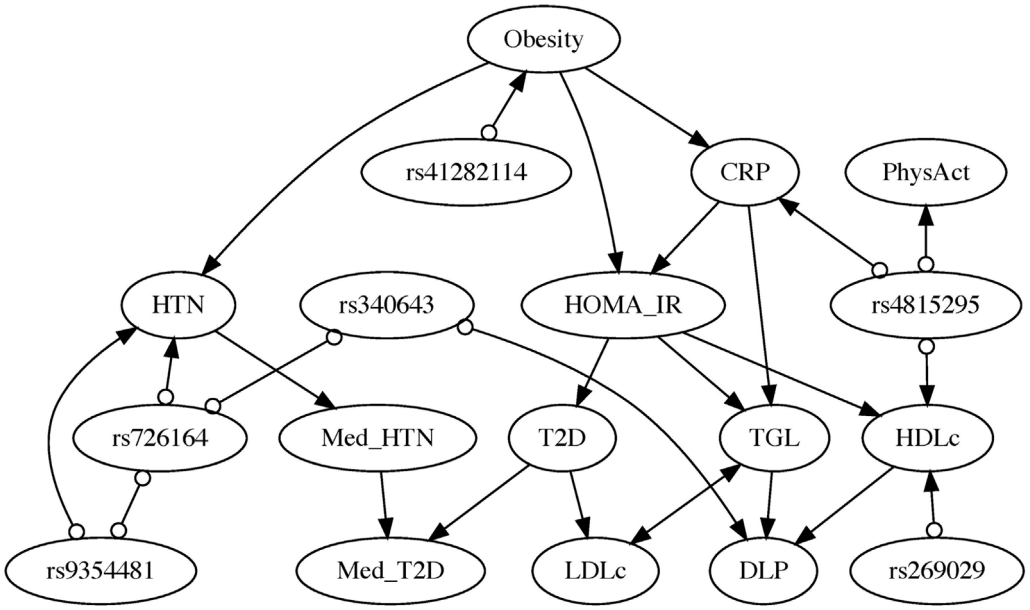

FIGURE 3
Partial Ancestral Graph inferred by the anchorFCI algorithm, using conservative orientations guided by the majority rule, and integrating the partial order SNPs ≺ phenotype variables. Directed edges imply ancestral (causal) relationships. Bidirected edges imply purely spurious associations. Circles indicate uncertain edge-marks, interchangeable with tail or arrowhead in equally probable models. Nodes representing various phenotypic traits include: Obesity, Type 2 Diabetes (T2D), Hypertension (HTN), Dyslipidemia (DLP), C-reactive protein (CRP), Homeostatic Model Assessment of Insulin Resistance (HOMA-IR), Triglycerides (TGL), Low-Density Lipoprotein Cholesterol (LDLc), High-Density Lipoprotein Cholesterol (HDLc), Medication for T2D (Med_T2D), Medication for HTN (Med_HTN), and Physical Activity (PhysAct).

$(\mathbb{E}(\text{HOMA} - \text{IR}|do(\text{Obesity})))$. Once more, the estimated expectations were adjusted solely by the standard set of covariates, and the difference between the expectations is statistically significant: for non-obese individuals, it is 2.6 with a 95% CI of [2.4, 3], while for obese individuals, it is 5.6 with a 95% CI of [4.8, 6.5].

3.4.4 T2D → LDLc

Figure 4D shows the point estimate and 95% CI for the expectation of LDLc given an intervention on T2D $(\mathbb{E}(\text{LDLc}|do(\text{T2D})))$. Given that the effect appears unconfounded in the PAG, it was adjusted solely by the standard set of covariates. For individuals without T2D, it is 100 with a 95% CI of [97, 110], while for individuals with T2D, it is 93 with a 95% CI of [85, 100]. The overlapping of the confidence intervals prevents us from concluding that the difference is significant.

3.4.5 Medication for HTN → medication for T2D

Figure 4E shows the point estimate and 95% CI for the probability of taking medication for T2D given an intervention on medication for HTN $(P(\text{Med\_T2D}|do(\text{Med\_HTN})))$. In accordance with the GBDC, the unbiased effect is achieved through adjustment for Obesity and standard covariates. If an individual is on medication for HTN, the probability of taking medication for T2D is 0.02, with a 95% CI of [0.0061, 0.064]. Conversely, if an individual is not on medication for HTN, the probability is 0.15, with a 95% CI of [0.05, 0.36]. Again, the overlapping of the confidence intervals prevents us from concluding that the difference is significant.

3.4.6 CRP → HOMA-IR

Figure 5A shows the average estimate and 95% confidence region for the expectation of HOMA-IR levels given different levels of CRP set by intervention $(\mathbb{E}(\text{HOMA} - \text{IR}|do(\text{CRP})))$. As per GBDC, estimates were obtained by adjusting for Obesity along with the standard set of covariates. While, on average, HOMA-IR levels tend to increase with CRP levels, the confidence region, which includes the constant line, suggests that this trend may not be statistically significant. For individuals with CRP = 0.3, the post-interventional expectation of HOMA-IR is 3.1, with a 95% CI of [2.7, 3.5]. In contrast, for individuals with CRP = 2, the expectation is 3.7, with a 95% CI of [3.1, 4.5].

3.4.7 CRP → TGL

Figure 5B shows the average estimate and 95% confidence region for the expectation of TGL given different levels of CRP set by intervention $(\mathbb{E}(\text{TGL}|do(\text{CRP})))$. As per GBDC, estimates were obtained by adjusting for Obesity along with the standard set of covariates. On average, TGL levels tend to rise with increasing CRP levels. However, the confidence region, which includes the constant line, suggests that this trend may not be statistically significant. Specifically, for individuals with CRP = 0.3, the post-interventional expectation of TGL is 110, with a 95% CI of [99, 120]. In contrast, for individuals with CRP = 2, the expectation is 130, with a 95% CI of [110, 150].

3.4.8 HOMA-IR → T2D

Figure 5C shows the average estimate and 95% confidence region for the probabilities of T2D, given different levels of HOMA-IR set by intervention $(P(\text{T2D}|do(\text{HOMA} - \text{IR})))$. Since

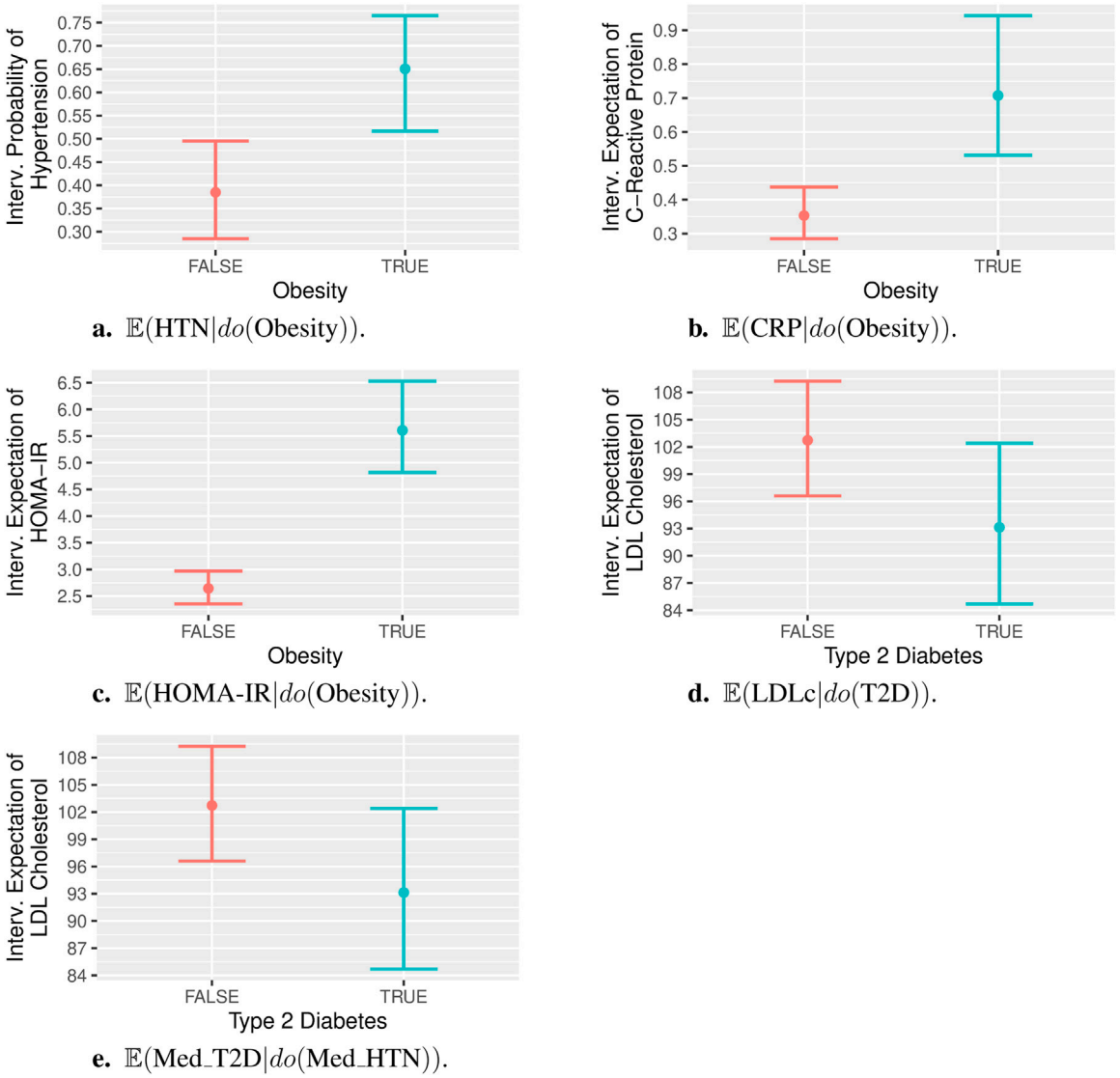

FIGURE 4
Post-interventional probabilities/expectations, with interventions on binary variables. Specifically, panels **(A)** to **(E)** show:$\mathbb{E}(\text{HTN}|do(\text{Obesity}))$, $\mathbb{E}(\text{CRP}|do(\text{Obesity}))$, $\mathbb{E}(\text{HOMA}-\text{IR}|do(\text{Obesity}))$, $\mathbb{E}(\text{LDLc}|do(\text{T2D}))$, and $\mathbb{E}(\text{Med\_T2D}|do(\text{Med\_HTN}))$. Dots represent point estimates, while the error bars indicate the 95% confidence intervals.

the effect is unconfounded in the PAG, it was adjusted solely using the standard set of covariates. The probability significantly increases with HOMA-IR levels. For instance, for individuals with HOMA-IR = 1, the probability of T2D is 0.02, with a 95% CI of [0.0062, 0.062]. In contrast, for individuals with HOMA-IR = 10.17, it rises to 0.16, with a 95% CI of [0.053, 0.39], and for those with HOMA-IR = 15.25, it further increases to 0.4, with a 95% CI of [0.14, 0.73].

3.4.9 HOMA-IR → TGL

Figure 5D shows the average estimate and 95% confidence region for the expectation of TGL given different levels of HOMA-IR set by intervention ($\mathbb{E}(\text{TGL}|do(\text{HOMA}-\text{IR}))$). Following the GBDC, unbiased estimates were derived through adjustment by CRP and standard covariates. TGL levels significantly increase with HOMA-IR levels. For individuals with HOMA-IR = 1 the post-interventional expectation of TGL is 100 with a 95% CI of [92, 110], while for individuals with HOMA-IR = 10.17, it is 130 with a 95% CI of [120, 150].

3.4.10 HOMA-IR → HDLc

Figure 5E shows the average estimate and 95% confidence region for the expectation of HDLc given different levels of HOMA-IR set by intervention ($\mathbb{E}(\text{HDLc}|do(\text{HOMA}-\text{IR}))$). As per GBDC, estimates were derived by adjusting for CRP and standard covariates. HDLc significantly decreases with HOMA-IR levels. For individuals with HOMA-IR = 1 the post-interventional expectation of HDLc is 45 with a 95% CI of [43, 48], while for individuals with HOMA-IR = 10.17, it is 38 with a 95% CI of [35, 41].

3.4.11 TGL → DLP

Figure 5F shows the average estimate and 95% confidence region for the probabilities of DLP, given different levels of TGL

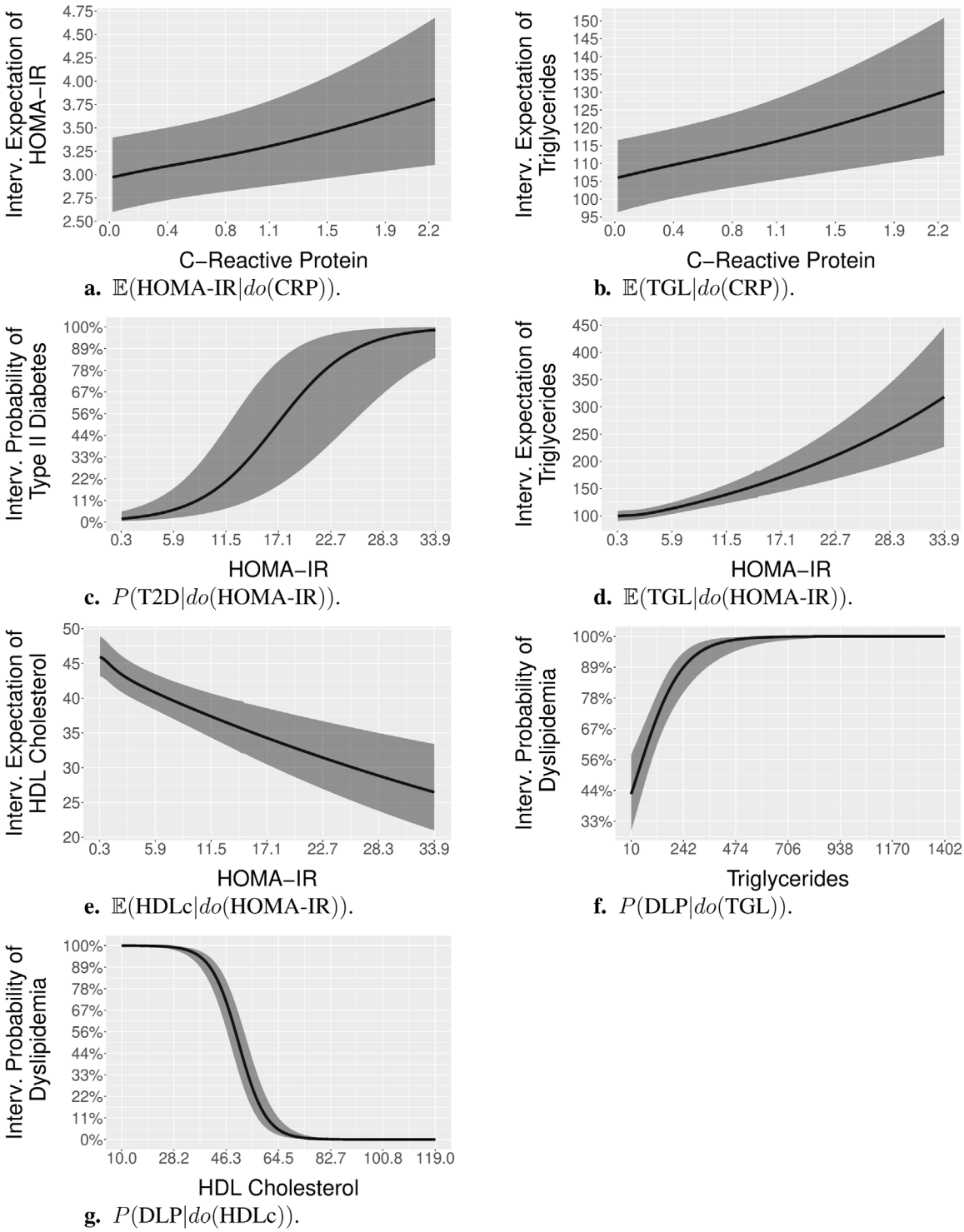

FIGURE 5
Post-interventional probabilities and expectations, with interventions on continuous variables. Specifically, panels (**A**). to (**G**). show: $\mathbb{E}(\text{HOMA} - \text{IR}|do(\text{CRP}))$, $\mathbb{E}(\text{TGL}|do(\text{CRP}))$, $P(\text{T2D}|do(\text{HOMA} - \text{IR}))$, $\mathbb{E}(\text{TGL}|do(\text{HOMA} - \text{IR}))$, $\mathbb{E}(\text{HDLc}|do(\text{HOMA} - \text{IR}))$, $P(\text{DLP}|do(\text{TLG}))$, and $P(\text{DLP}|do(\text{HDLc}))$. Full lines represent average estimates, while the shaded areas indicate the 95% confidence regions.

set by intervention ($P(\text{DLP}|do(\text{TLG}))$). The probability of DLP significantly increases with TGL levels. Following GBDC, unbiased estimates were derived through adjustment for CRP and the standard set of covariates. For individuals with TGL = 52.2, the post-interventional probability of DLP is 0.54 with a 95% CI of [0.42, 0.65], while for individuals with TGL = 150.6, it is 0.76 with a 95% CI of [0.66, 0.84].

3.4.12 HDLc → DLP

Figure 5G shows the average estimate and 95% confidence region for the probabilities of DLP, given different levels of HDLc set by intervention ($P(\text{DLP}|do(\text{HDLc}))$). Following the GBDC, unbiased estimates were obtained through adjustments for TGL and the standard set of covariates. The probability of DLP is notably elevated for low values of HDLc. With an HDLc of 40.82, the post-interventional probability of DLP reaches 0.89 with a 95% CI of [0.8, 0.94]. As HDLc levels increase, the probability decreases: for HDLc = 50.73, it drops to 0.49, with a 95% CI of [0.34, 0.64], and further decreases to 0.11 for HDLc = 60.64, with a 95% CI of [0.054, 0.2].

4 Discussion

Our data-driven approach not only uncovers a causal relationship from obesity to HTN, body inflammation (assessed by CRP levels), and insulin resistance (HOMA-IR), but also provides further evidence that obesity significantly increases the risk of developing these conditions. The analysis also indicates that body inflammation, as indicated by CRP, causally influences insulin resistance, as indicated by HOMA-IR. However, its effect size could not be obtained as statistically significant using our observed sample.

4.1 Obesity → HTN

Extensive research HTN established a strong association and a causal link from obesity to HTN, elucidated by complex underlying mechanisms (Jiang et al., 2016; Seravalle and Grassi, 2017; Powell-Wiley et al., 2021). The contribution of obesity to the development and progression of HTN can be attributed to a range of factors. These include the overactivation of the renin-angiotensin-aldosterone system and the sympathetic nervous system, the overstimulation of pro-inflammatory adipokines–such as tumor necrosis factor-α (TNF-α), leptin, and plasminogen activator inhibitor type 1 –, insulin resistance, immune dysfunction, and structural and functional alterations in renal, cardiac, and adipose tissues (Kotsis et al., 2010; Shams et al., 2022). Unraveling the precise interplay between these factors remains a significant ongoing challenge in the field of medical research.

4.2 Obesity → CRP → HOMA-IR → T2D

Numerous studies consistently show that individuals with obesity exhibit higher CRP levels—a marker of low-grade inflammation—compared to those with a healthy weight (Visser et al., 1999; Yudkin et al., 1999; Choi et al., 2013). Obesity primarily contributes to chronic low-grade inflammation by releasing a surge of signaling molecules, including the adipokines leptin and resistin, the cytokine interleukin-6 (IL-6), and the chemokine monocyte chemoattractant protein-1 (MCP-1). These molecules create a pro-inflammatory environment, attracting immune cells called monocytes into adipose tissue. This inflammatory cascade is implicated in obesity-related metabolic dysfunctions, potentially culminating in insulin resistance and the onset of T2D (Chen et al., 2015; Wu and Ballantyne, 2020).

4.3 Obesity → HOMA-IR

The contribution of obesity to insulin resistance is not solely mediated through inflammation. Besides inflammation, obesity-driven factors such as elevated levels of free fatty acids (FFA) in muscle and liver tissues can directly disrupt insulin signaling pathways, in a process called lipotoxicity (Ahmed et al., 2021; Filipović et al., 2021). Additionally, adipose tissue dysfunction in obesity leads to the release of proinflammatory cytokines such as IL-6 and TNF-α that can directly interfere with insulin signaling within cells, further exacerbating insulin resistance (Fève and Bastard, 2009).

Our causal analysis reveals that high levels of HOMA-IR significantly contributes not only to the onset of T2D but also to elevated TGL and decreased HDLc levels. Furthermore, in accordance with the definition of DLP, our analysis reveals that both HDLc and TGL causally contribute to DLP, with low levels of HDLc and high levels of TGL significantly increasing the likelihood of DLP.

4.4 HOMA-IR → TGL and HDLc → DLP

Various population-based studies have consistently established a correlation between insulin resistance and increased TGL alongside decreased levels of HDLc. For a comprehensive list of references, see, for instance, Howard (1999). Individuals with insulin resistance often exhibit a condition known as diabetic dyslipidemia or dyslipidemia of insulin resistance, marked by elevated TGL levels, decreased levels of HDLc, and normal or slightly elevated levels of LDLc (Bahiru et al., 2021). The potential of a higher TGL/HDLc ratio as a marker for insulin resistance has been explored in several studies (Giannini et al., 2011; Behiry et al., 2019). In insulin resistance, the typical suppression of VLDL production by the liver, a lipoprotein rich in triglycerides, does not occur as expected. Consequently, there is an increase in VLDL production, leading to elevated triglyceride levels in the bloodstream. Additionally, various factors in insulin resistance may contribute to reduced HDL, such as cholesterol ester exchange between HDLc and VLDL triglycerides, increased hepatic lipase activity, and altered hepatic function affecting production of apo AI (the main apoprotein of HDL) and secretion of nascent HDLc (Howard, 1999; Ginsberg et al., 2005).

Certain identified causal relationships exhibited effect sizes that did not reach statistical significance. These included the effects of CRP on HOMA-IR and TGL, as well as the effect of T2D on LDLc. The corresponding effect of the causal relationship from Med_HTN to Med_T2D was also not significant.

4.5 CRP → HOMA-IR and TGL

Numerous studies indicate that systemic inflammation significantly influences insulin resistance (Wu and Ballantyne,

2020; Bulcão et al., 2006) and is associated with elevated TGL levels Feingold and Grunfeld (2022); Ronti et al. (2006). The lack of statistical significance in our analysis could be attributed to substantial variability within our dataset, indicating a need for a larger sample size in further investigations.

4.6 T2D → LDL

The lack of statistical significance in the effect of T2D on LDLc may indicate that this relationship involves a more complex interplay than a mere quantitative change in LDLc levels. Within our dataset, many diabetic patients have normal or even low LDLc levels, which is consistent with existing studies (Howard et al., 2000). However, T2D can qualitatively alter LDL particles, leading to the formation of smaller, denser particles that are more prone to oxidation and arterial infiltration. These changes increase the risk of atherosclerosis and cardiovascular complications in diabetic individuals (Taskinen, 1992).

4.7 Med_HTN → Med_T2D

The causal edge from Med_HTN to Med_T2D appears to be false, given the non-significance of the corresponding effect and the absence of scientific evidence supporting the claim that medication for HTN has a direct effect on medication for T2D. However, some studies have indicated that Med_HTN can increase the risk of T2D Bangalore et al. (2007); Arnold et al. (2020). In this scenario, the causal edge from Med_HTN to Med_T2D may be accounting for an effect that is mediated by T2D but was not accurately represented in the graph. This discrepancy could have occurred because the edge between Med_HTN and T2D may have been mistakenly removed by the algorithm, possibly due to a false identification of conditional independencies from weak correlations. Further exploration is necessary to gain a better understanding of the nature of this relationship and its implications.

5 Conclusion, limitations, and final remarks

The conservative RFCI is a classical causal discovery algorithm tailored for scenarios featuring latent confounding. It distinguishes itself by providing superior computational efficiency while maintaining consistency in sparse models, along with enhanced robustness through an approach that avoids tests susceptible to low statistical power (Colombo et al., 2012). However, it tends to produce rather uninformative models when faced with conflicting orientations.

To address this limitation, this paper introduces anchorFCI, a novel adaptation of the conservative RFCI. Given two sets of variables, where the first includes the variables of interest and the second consists solely of those known not to be caused by the first, the algorithm strategically identifies reliable anchor variables and seamlessly integrates their known non-ancestral relationships in the learning process. Reliable anchor variables are defined as those from the second set that are significantly associated with a variable from the first set and, when integrated into the graph, form unambiguous triples based on the majority rule. By effectively identifying and integrating these variables, anchorFCI significantly enhances its capability to orient edges within the conservative framework, thereby increasing both robustness and overall discovery power, particularly in real-world scenarios marked by latent confounding and limited sample sizes.

By employing the anchorFCI, followed by state-of-the-art effect identification tools (Jaber et al., 2022), we conducted a fully data-driven causal analysis of various cardiometabolic risk factors and related SNP variables included in the 2015 ISA-Nutrition dataset (Fisberg et al., 2018). The approach proved effective, uncovering the causal relationships among all the phenotype and clinical variables under consideration. Importantly, many of the identified orientations are strongly supported by existing literature, thereby enhancing the credibility of our findings.

While our investigation has yielded interesting results, it also highlights the critical need for more robust methods for causal discovery from observational data in practical scenarios. A plethora of causal discovery methods have been recently proposed, with many emphasizing scalability (Zheng et al., 2018; Glymour et al., 2019; Vowels et al., 2022). However, they often focus on scenarios that are unrealistic due to the imposition of strong assumptions, including not only faithfulness but also causal sufficiency and various parametric or distributional assumptions. Only a few have been proposed for non-parametric causal discovery in scenarios involving latent confounding (Colombo et al., 2012; Colombo and Maathuis, 2014; Hyttinen et al., 2014; Magliacane et al., 2016). Nevertheless, they still exhibit high vulnerability to violations of the faithfulness assumption, which is commonly observed in real-world settings. Any falsely identified conditional dependence or independence may lead to findings that deviate from the truth and our current understanding. Hence, there is an urgent demand for methods that demonstrate greater resilience in situations with limited data, capable of adeptly managing conflicting orientations, accommodating uncertainty in the learning process, and integrating background knowledge.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: The dataset utilized in this study is not publicly available. Requests to access these datasets should be directed to Regina Mara Fisberg. Further details can be found at www.gac-usp.com.br.

Ethics statement

The studies involving humans were approved by the Ethics Committee on Research of the School of Public Health of the University of São Paulo (\#30848914.7.0000.5421). The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

AR: Conceptualization, Formal Analysis, Investigation, Methodology, Software, Supervision, Visualization, Writing–original draft, Writing–review and editing. MC: Data curation, Formal Analysis, Visualization, Writing–review and editing. JP: Data curation, Formal Analysis, Writing–review and editing. RF: Resources, Writing–review and editing. FS: Resources, Writing–review and editing. MR: Resources, Writing–review and editing. DH: Conceptualization, Funding acquisition, Resources, Supervision, Writing–review and editing. AC: Conceptualization, Formal Analysis, Supervision, Visualization, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. We acknowledge support from the Open Access Publication Fund of the University of Münster. Also, this work has been funded in part by the Bundesministerium für Bildung und Forschung (Federal Ministry of Education and Research: Deep Insight 031L0267A) and the LOEWE program of the State of Hesse (Germany) in the research cluster Diffusible Signals (LOEWE/2/13/519/03/06.001 (0002)/74) to DH. This research was also supported by the São Paulo Research Foundation (FAPESP) grant #2022/03420–0 to MC and grant #2023/05857–9 to AC, as well as by São Paulo Municipal Health Department grant #2013–0.235.936–0; FAPESP grant #2017/05125–7; and National Council for Scientific and Technological Development (CNPq grant #402674/2016–2) to RMF. The funding sources were not involved in the study design, data collection, data analysis and interpretation, writing of the report, or the decision to submit the paper for publication.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2024.1436947/full#supplementary-material

References

Ahmed, B., Sultana, R., and Greene, M. W. (2021). Adipose tissue and insulin resistance in obese. *Biomed. and Pharmacother.* 137, 111315. doi:10.1016/j.biopha.2021.111315

Alves, M. C. G. P., Escuder, M. M. L., Goldbaum, M., Barros, M. B. D. A., Fisberg, R. M., and Cesar, C. L. G. (2018). Sampling plan in health surveys, city of são paulo, Brazil, 2015. *Rev. Saúde Pública* 52, 81. doi:10.11606/S1518-8787.2018052000471

Andrews, B., Spirtes, P., and Cooper, G. F. (2020). "On the completeness of causal discovery in the presence of latent confounding with tiered background knowledge," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics* (PMLR), 4002–4011.

Arnold, S. V., Bhatt, D. L., Barsness, G. W., Beatty, A. L., Deedwania, P. C., Inzucchi, S. E., et al. (2020). Clinical management of stable coronary artery disease in patients with type 2 diabetes mellitus: a scientific statement from the american heart association. *Circulation* 141, e779–e806. doi:10.1161/CIR.0000000000000766

Bahiru, E., Hsiao, R., Phillipson, D., and Watson, K. E. (2021). Mechanisms and treatment of dyslipidemia in diabetes. *Curr. Cardiol. Rep.* 23, 26–6. doi:10.1007/s11886-021-01455-w

Bangalore, S., Parkar, S., Grossman, E., and Messerli, F. H. (2007). A meta-analysis of 94,492 patients with hypertension treated with beta blockers to determine the risk of new-onset diabetes mellitus. *Am. J. Cardiol.* 100, 1254–1262. doi:10.1016/j.amjcard.2007.05.057

Barroso, W. K. S., Rodrigues, C. I. S., Bortolotto, L. A., Mota-Gomes, M. A., Brandão, A. A., Feitosa, A. D. D. M., et al. (2021). Brazilian guidelines of hypertension - 2020. *Arq. Bras. Cardiol.* 116, 516–658. doi:10.36660/abc.20201238

Behiry, E. G., El Nady, N. M., AbdEl Haie, O. M., Mattar, M. K., and Magdy, A. (2019). Evaluation of tg-hdl ratio instead of homa ratio as insulin resistance marker in overweight and children with obesity. *Endocr. Metabolic and Immune Disorders-Drug Targets Formerly Curr. Drug Targets-Immune, Endocr. and Metabolic Disord.* 19, 676–682. doi:10.2174/1871530319666190121123535

Bulcão, C., Ferreira, S. R., Giuffrida, F., and Ribeiro-Filho, F. F. (2006). The new adipose tissue and adipocytokines. *Curr. diabetes Rev.* 2, 19–28. doi:10.2174/157339906775473617

Bull, F. C., Al-Ansari, S. S., Biddle, S., Borodulin, K., Buman, M. P., Cardon, G., et al. (2020). World health organization 2020 guidelines on physical activity and sedentary behaviour. *Br. J. sports Med.* 54, 1451–1462. doi:10.1136/bjsports-2020-102955

Chen, L., Chen, R., Wang, H., and Liang, F. (2015). Mechanisms linking inflammation to insulin resistance. *Int. J. Endocrinol.* 2015, 508409. doi:10.1155/2015/508409

Chickering, M., Heckerman, D., and Meek, C. (2004). Large-sample learning of bayesian networks is np-hard. *J. Mach. Learn. Res.* 5, 1287–1330.

Choi, J., Joseph, L., and Pilote, L. (2013). Obesity and c-reactive protein in various populations: a systematic review and meta-analysis. *Obes. Rev.* 14, 232–244. doi:10.1111/obr.12003

Claassen, T., Mooij, J., and Heskes, T. (2013). "Learning sparse causal models is not np-hard," in Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, UAI 2013, Arlington, VA, USA, August 11-15, 2013 (AUAI Press).

Cobas, R., Rodacki, M., Giacaglia, L., Calliari, L., Noronha, R., Valerio, C., et al. (2022). *Diagnóstico do diabetes e rastreamento do diabetes tipo 2*. Diretriz Oficial da Sociedade Brasileira de Diabetes, 557753–562022.

Colombo, D., and Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.* 15, 3741–3782.

Colombo, D., Maathuis, M. H., Kalisch, M., and Richardson, T. S. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann. Statistics* 40, 294–321. doi:10.1214/11-aos940

Davey Smith, G., and Hemani, G. (2014). Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* 23, R89–R98. doi:10.1093/hmg/ddu328

de Leeuw, C., Savage, J., Bucur, I. G., Heskes, T., and Posthuma, D. (2022). Understanding the assumptions underlying mendelian randomization. *Eur. J. Hum. Genet.* 30, 653–660. doi:10.1038/s41431-022-01038-5

Feingold, K. R., and Grunfeld, C. (2022). *The effect of inflammation and infection on lipids and lipoproteins*. South Dartmouth, MA, USA: MDText.com, Inc.

Fève, B., and Bastard, J.-P. (2009). The role of interleukins in insulin resistance and type 2 diabetes mellitus. *Nat. Rev. Endocrinol.* 5, 305–311. doi:10.1038/nrendo.2009.62

Filipović, B., Marušić, M., Paić, M., Knobloch, M., and Liberati Pršo, A.-M. (2021). Nafld, insulin resistance, and diabetes mellitus type 2. *Can. J. Gastroenterology Hepatology* 2021, 6613827. doi:10.1155/2021/6613827

Fisberg, R. M., Sales, C. H., Fontanelli, M. D. M., Pereira, J. L., Alves, M. C. G. P., Escuder, M. M. L., et al. (2018). 2015 health survey of são paulo with focus in nutrition: rationale, design, and procedures. *Nutrients* 10, 169. doi:10.3390/nu10020169

Flynn, J. T., Kaelber, D. C., Baker-Smith, C. M., Blowey, D., Carroll, A. E., Daniels, S. R., et al. (2017). Clinical practice guideline for screening and management of high blood pressure in children and adolescents. *Pediatrics* 140, e20171904. doi:10.1542/peds.2017-1904

Geloneze, B., Vasques, A. C. J., Stabe, C. F. C., Pareja, J. C., Rosado, L. E. F. P. D. L., Queiroz, E. C. D., et al. (2009). Homa1-ir and homa2-ir indexes in identifying insulin resistance and metabolic syndrome: Brazilian metabolic syndrome study (brams). *Arquivos Brasileiros de Endocrinol. and Metabologia* 53, 281–287. doi:10.1590/s0004-27302009000200020

Gerhardus, A., and Runge, J. (2020). "High-recall causal discovery for autocorrelated time series with latent confounders," in *Advances in neural information processing systems*. Editors H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Curran Associates, Inc.), 33, 12615–12625.

Giannini, C., Santoro, N., Caprio, S., Kim, G., Lartaud, D., Shaw, M., et al. (2011). The triglyceride-to-hdl cholesterol ratio: association with insulin resistance in obese youths of different ethnic backgrounds. *Diabetes care* 34, 1869–1874. doi:10.2337/dc10-2234

Ginsberg, H. N., Zhang, Y.-L., and Hernandez-Ono, A. (2005). Regulation of plasma triglycerides in insulin resistance and diabetes. *Archives Med. Res.* 36, 232–240. doi:10.1016/j.arcmed.2005.01.005

Giuliano, I. D. C. B., Caramelli, B., Pellanda, L. C., Duncan, B. B., Mattos, S., and Fonseca, F. A. H. (2005). I diretriz de prevenção da aterosclerose na infância e na adolescência. *Arq. Bras. Cardiol.* 85 (6), 1–36. doi:10.1590/S0066-782X2005002500001

Glymour, C., Zhang, K., and Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Front. Genet.* 10, 524. doi:10.3389/fgene.2019.00524

Golbert, A., Vasques, A. C. J., Faria, A., Lottenberg, A. M. P., Joaquim, A. G., Vianna, A. G. D., et al. (2019). *Diretrizes da sociedade brasileira de diabetes 2019-2020*. São Paulo: Clannad, 1–491.

Höfler, M. (2005). The bradford hill considerations on causality: a counterfactual perspective. *Emerg. themes Epidemiol.* 2, 11–19. doi:10.1186/1742-7622-2-11

Holmberg, M. J., and Andersen, L. W. (2022). Collider bias. *Jama* 327, 1282–1283. doi:10.1001/jama.2022.1820

Howard, B. V. (1999). Insulin resistance and lipid metabolism. *Am. J. Cardiol.* 84, 28–32. doi:10.1016/s0002-9149(99)00355-0

Howard, B. V., Robbins, D. C., Sievers, M. L., Lee, E. T., Rhoades, D., Devereux, R. B., et al. (2000). Ldl cholesterol as a strong predictor of coronary heart disease in diabetic individuals with insulin resistance and low ldl: the strong heart study. *Arteriosclerosis, thrombosis, Vasc. Biol.* 20, 830–835. doi:10.1161/01.atv.20.3.830

Hyttinen, A., Eberhardt, F., and Järvisalo, M. (2014). Constraint-based causal discovery: conflict resolution with answer set programming. *Uncertain. Artif. Intell. (UAI)*. doi:10.5555/3020751.3020787

Jaber, A., Ribeiro, A. H., Zhang, J., and Bareinboim, E. (2022). Causal identification under markov equivalence: calculus, algorithm, and completeness. *Adv. Neural Inf. Process. Syst.* 35, 3679–3690.

Jiang, S.-Z., Lu, W., Zong, X.-F., Ruan, H.-Y., and Liu, Y. (2016). Obesity and hypertension. *Exp. Ther. Med.* 12, 2395–2399. doi:10.3892/etm.2016.3667

Kalisch, M., Hauser, A., Maechler, M., Colombo, D., Entner, D., Hoyer, P., et al. (2024). Package pcalg

Kälsch, J., Bechmann, L. P., Heider, D., Best, J., Manka, P., Kälsch, H., et al. (2015). Normal liver enzymes are correlated with severity of metabolic syndrome in a large population based cohort. *Sci. Rep.* 5, 13058. doi:10.1038/srep13058

Kotsis, V., Stabouli, S., Papakatsika, S., Rizos, Z., and Parati, G. (2010). Mechanisms of obesity-induced hypertension. *Hypertens. Res.* 33, 386–393. doi:10.1038/hr.2010.9

Lagani, V., Athineou, G., Farcomeni, A., Tsagris, M., and Tsamardinos, I. (2017). Feature selection with the R package MXM: discovering statistically equivalent feature subsets. *J. Stat. Softw.* 80. doi:10.18637/jss.v080.i07

Lin, I. (2019). simMixedDAG: simulate mixed data type datasets from DAG models. *R. package version 1.0, Commit. 11f720c104faca4834cf76b3dcb376c67b864e8a*.

Maathuis, M. H., and Colombo, D. (2015). A generalized back-door criterion. *Ann. Statistics* 43, 1060–1088. doi:10.1214/14-AOS1295

Magliacane, S., Claassen, T., and Mooij, J. M. (2016). Ancestral causal inference. *Adv. Neural Inf. Process. Syst. (NeurIPS)*.

Mayer, M., and Mayer, M. M. (2019). Package 'missranger. *R. Package*.

McCaw, Z. (2023). RNOmni: rank normal transformation omnibus test. R package version 1.0.1.2

Miranda, J. J., Barrientos-Gutierrez, T., Corvalan, C., Hyder, A. A., Lazo-Porras, M., Oni, T., et al. (2019). Understanding the rise of cardiometabolic diseases in low-and middle-income countries. *Nat. Med.* 25, 1667–1679. doi:10.1038/s41591-019-0644-7

Olsson, U. (2005). Confidence intervals for the mean of a log-normal distribution. *J. Statistics Educ.* 13. doi:10.1080/10691898.2005.11910638

Onis, M. D., Onyango, A. W., Borghi, E., Siyam, A., Nishida, C., and Siekmann, J. (2007). Development of a who growth reference for school-aged children and adolescents. *Bull. World health Organ.* 85, 660–667. doi:10.2471/blt.07.043497

Pearl, J. (2000a). *Causality: models, reasoning, and inference*. 2nd edn. NY, USA: Cambridge University Press.

Pearl, J. (2000b). Causal inference without counterfactuals: comment. *J. Am. Stat. Assoc.* 95, 428–431. doi:10.2307/2669380

Pereira, J. L., de Souza, C. A., Neyra, J. E. M., Leite, J. M. R. S., Cerqueira, A., Mingroni-Netto, R. C., et al. (2024). Genetic ancestry and self-reported "skin color/race" in the urban admixed population of são paulo city, Brazil. *Genes* 15, 917. doi:10.3390/genes15070917

Perkovi, E., Textor, J., Kalisch, M., and Maathuis, M. H. (2018). Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs. *J. Mach. Learn. Res.* 18, 1–62. doi:10.5555/3122009.3242077

Powell-Wiley, T. M., Poirier, P., Burke, L. E., Després, J.-P., Gordon-Larsen, P., Lavie, C. J., et al. (2021). Obesity and cardiovascular disease: a scientific statement from the american heart association. *Circulation* 143, e984–e1010. doi:10.1161/CIR.0000000000000973

Précoma, D. B., Oliveira, G. M. M. D., Simão, A. F., Dutra, O. P., Coelho, O. R., Izar, M. C. D. O., et al. (2019). Updated cardiovascular prevention guideline of the Brazilian society of cardiology - 2019. *Arq. Bras. Cardiol.* 113, 787–891. doi:10.5935/abc.20190204

Rao, G. (2018). Cardiometabolic diseases: a global perspective. *J. Cardiol. Cardiovasc Ther.* 12, 555834. doi:10.19080/jocct.2018.12.555834

Ribeiro, A. H., Soler, J. M., Neto, E. C., and Fujita, A. (2016). Causal inference and structure learning of genotype–phenotype networks using genetic variation. *Big Data Anal. Genomics*, 89–143. doi:10.1007/978-3-319-41279-5_3

Ribeiro, A. H., and Soler, J. M. P. (2020). Learning genetic and environmental graphical models from family data. *Statistics Med.* 39, 2403–2422. doi:10.1002/sim.8545

Ridker, P. M. (2003). Clinical application of c-reactive protein for cardiovascular disease detection and prevention. *Circulation* 107, 363–369. doi:10.1161/01.cir.0000053730.47739.3c

Ronti, T., Lupattelli, G., and Mannarino, E. (2006). The endocrine function of adipose tissue: an update. *Clin. Endocrinol.* 64, 355–365. doi:10.1111/j.1365-2265.2006.02474.x

Seravalle, G., and Grassi, G. (2017). Obesity and hypertension. *Pharmacol. Res.* 122, 1–7. doi:10.1016/j.phrs.2017.05.013

Shams, E., Kamalumpundi, V., Peterson, J., Gismondi, R. A., Oigman, W., and de Gusmão Correia, M. L. (2022). Highlights of mechanisms and treatment of obesity-related hypertension. *J. Hum. Hypertens.* 36, 785–793. doi:10.1038/s41371-021-00644-y

Spirtes, P. (2001). "An anytime algorithm for causal inference," in *Proceedings of the eighth international workshop on artificial intelligence and statistics*. Editors T. S. Richardson and T. S. Jaakkola, 278–285.

Spirtes, P., Glymour, C., and Scheines, R. (2001). *Causation, prediction, and search*. 2nd edn. Cambridge, MA: MIT Press.

Stekhoven, D. J., and Bühlmann, P. (2012). Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28, 112–118. doi:10.1093/bioinformatics/btr597

Taskinen, M.-R. (1992). Quantitative and qualitative lipoprotein abnormalities in diabetes mellitus. *Diabetes* 41, 12–17. doi:10.2337/diab.41.2.s12

Thomassen, D., le Cessie, S., van Houwelingen, H. C., and Steyerberg, E. W. (2024). Effective sample size: a measure of individual uncertainty in predictions. *Statistics Med.* 43, 1384–1396. doi:10.1002/sim.10018

Tian, J., Paz, A., and Pearl, J. (1998). *Finding minimal d-separators* (Citeseer)

Tsagris, M., Borboudakis, G., Lagani, V., and Tsamardinos, I. (2018). Constraint-based causal discovery with mixed data. *Int. J. data Sci. Anal.* 6, 19–30. doi:10.1007/s41060-018-0097-y

Valenzuela, P. L., Carrera-Bastos, P., Castillo-García, A., Lieberman, D. E., Santos-Lozano, A., and Lucia, A. (2023). Obesity and the risk of cardiometabolic diseases. *Nat. Rev. Cardiol.* 20, 475–494. doi:10.1038/s41569-023-00847-5

Visser, M., Bouter, L. M., McQuillan, G. M., Wener, M. H., and Harris, T. B. (1999). Elevated c-reactive protein levels in overweight and obese adults. *Jama* 282, 2131–2135. doi:10.1001/jama.282.22.2131

Vowels, M. J., Camgoz, N. C., and Bowden, R. (2022). D'ya like dags? a survey on structure learning and causal discovery. *ACM Comput. Surv.* 55, 1–36. doi:10.1145/3527154

Wang, T.-Z., Qin, T., and Zhou, Z.-H. (2022). "Sound and complete causal identification with latent variables given local background knowledge," in *Advances in neural information processing systems*. Editors A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, 29.

WHO Consultation (2000). *Obesity: preventing and managing the global epidemic: report of a who consultation*. Geneva: WHO.

Wilson, P., and Meigs, J. (2008). Cardiometabolic risk: a framingham perspective. *Int. J. Obes.* 32, S17–S20. doi:10.1038/ijo.2008.30

World Health Organization (2010). *Global recommendations on physical activity for health*. Schweiz: Genf. Tech. rep.

Wu, H., and Ballantyne, C. M. (2020). Metabolic inflammation and insulin resistance in obesity. *Circulation Res.* 126, 1549–1564. doi:10.1161/CIRCRESAHA.119.315896

Yudkin, J. S., Stehouwer, C., Emeis, J., and Coppack, S. (1999). C-reactive protein in healthy subjects: associations with obesity, insulin resistance, and endothelial dysfunction: a potential role for cytokines originating from adipose tissue? *Arteriosclerosis, thrombosis, Vasc. Biol.* 19, 972–978. doi:10.1161/01.atv.19.4.972

Zhalama, Z., Zhang, J., and Mayer, W. (2017). Weakening faithfulness: some heuristic causal discovery algorithms. *Int. J. Data Sci. Anal.* 3, 93–104. doi:10.1007/s41060-016-0033-y

Zhang, J. (2008a). Causal reasoning with ancestral graphs. *J. Mach. Learn. Res.* 9, 1437–1474. doi:10.5555/1390681.1442780

Zhang, J. (2008b). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif. Intell.* 172, 1873–1896. doi:10.1016/j.artint.2008.08.001

Zhang, J., and Spirtes, P. (2008). Detection of unfaithfulness and robust causal inference. *Minds Mach.* 18, 239–271. doi:10.1007/s11023-008-9096-4

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. (2018). Dags with no tears: continuous optimization for structure learning. *Adv. neural Inf. Process. Syst.* 31.