Direct causal variable discovery leveraging the invariance principle: application in biomedical studies

Liangying Yin[1,2], Menghui Liu[3], Yujia Shi[1], Jinghong Qiu[1], Hon-Cheong So[1,4-9*]

[1]School of Biomedical Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong
[2]Eric and Wendy Schmidt Center, The Broad Institute of MIT and Harvard, USA
[3]Department of Epidemiology, University of Florid, USA
[4]CUHK Shenzhen Research Institute, Shenzhen, China
[5]Department of Psychiatry, The Chinese University of Hong Kong, Hong Kong
[6]KIZ-CUHK Joint Laboratory of Bioresources and Molecular Research of Common Diseases, Kunming Institute of Zoology and The Chinese University of Hong Kong, China
[7]Margaret K.L. Cheung Research Centre for Management of Parkinsonism, The Chinese University of Hong Kong, Shatin, Hong Kong
[8]Brain and Mind Institute, The Chinese University of Hong Kong, Hong Kong SAR, China
[9]Hong Kong Branch of the Chinese Academy of Sciences Center for Excellence in Animal Evolution and Genetics, The Chinese University of Hong Kong, Hong Kong SAR, China

Correspondence to: Hon-Cheong So, Lo Kwee-Seong Integrated Biomedical Sciences Building, The Chinese University of Hong Kong, Shatin, Hong Kong. Tel: +852 3943 9255; E-mail: hcso@cuhk.edu.hk

Abstract

Accurate identification of direct causal (parental) variables for a target is of primary interest in many applications, especially in biomedical sciences. It could promote our understanding of disease mechanisms, and facilitate the discovery of new biomarkers and therapeutic targets for clinical traits. However, standard machine learning approaches often identify spurious associations, while existing causal inference methods for direct causal variables can be computationally infeasible for high-dimensional biomedical data.

Here, we proposed a novel and efficient two-stage approach (I-GCM) to discover direct causal variables (including genetic and clinical variables) for clinical outcomes. The method first employs the PC-simple algorithm for feature screening, then leverages the principle of causal invariance across different environments. Causal relationships are robustly identified by testing for changes in the generalized covariance measure (GCM), calculated using flexible gradient-boosted tree models.

We first verified the proposed method through extensive simulations. I-GCM constantly yielded high precision (positive predictive value) and specificity while maintaining satisfactory sensitivity in general, and consistently outperformed a standard method. Notably, the precision was larger than 90% in our simulated scenarios, even in high-dimensional settings.

We then applied I-GCM to the UK-Biobank, analyzing genetic and clinical data to identify causal factors for COVID-19 infection/severity and lipid traits (HDL, Triglycerides). The analysis successfully recovered many known clinical risk factors, validating the method's real-world performance, and uncovered novel putative causal genes and biological pathways supported by existing literature.

Importantly, our work pioneers the application of the invariance principle for causal inference in biomedical/clinical studies, and suggests a new avenue for causal discovery in these settings.

Introduction

Accurate identification of direct causal(parental) variables for a target variable is crucial in many applications. It could enhance our understanding of the biological and pathophysiological mechanisms underlying various diseases/traits. More importantly, it may lead to the discovery of new biomarkers and therapeutic targets for studied clinical outcomes. In recent years, causal inference has gained increasing

attention in different areas[1-5]., e.g., economics, social science, biomedicine, etc. Despite this trend, many researchers still rely on machine learning methods, e.g., linear regression, random forest, and deep neural networks, to decipher outcome-associated variables. However, these methods are association-based, and association does not necessarily imply causation. Many identified variables may be subject to the influence of confounders and prove to be irrelevant to the target.

Causal inference could serve as a valuable tool to eliminate spuriously associated variables. However, fewer studies have investigated how to identify the direct(parental) causal variables for the target variable. In earlier work, Buhlmann et al.[6] presented a method (PC-simple), which utilized partial correlation screening, to infer the linear causal relations between covariates and the target from observational data. It's an efficient method in high-dimensional settings but tended to have a high false discovery rate. Besides, it was not designed to incorporate multiple data sources for causal inference. Also, a prior assumption of a linear model was required. A previous work by Peters et al.[7] proposed to exploit the invariance property of a causal linear model (ICP) from different experimental settings/environments (i.e., observational and interventional data) to uncover the direct causal variables for a target. This method first enumerated all variable sets and then examined whether each set is a candidate parental set. This is done by testing whether the residuals derived from the linear model fitted on the current variable set are equally distributed across different environments. The ultimate causal variable set is the intersection of all candidates. Heinze-Deml et al.[8] extended the work by Peters et al to accommodate nonlinear causal relations (nonlinear ICP). However, the computational complexity of the above grows super-exponentially with variable dimension (2^q). Moreover, it may be challenging to achieve high power if the actual parental set size is >2. Consequently, these methods may not be practical or suitable for high-dimensional settings.

In this study, we introduced a novel approach, invariant generalized covariance measure (I-GCM), designed to accurately identify direct causal variables for a given variable of interest. The method operates in two stages. The first stage involves variable selection using the PC-simple method. The second stage exploits the invariant property of causal relations indicated by the generalized covariance measure (GCM) to reliably discover causal variables from multiple data sources. We proposed to use gradient-boosted trees to compute the generalized covariances between variables, boosting the capability to discern both linear and non-linear relations between variables. The initial face of variable selection substantially reduces the feature dimension, leading to improved computational efficiency. We hypothesized that leveraging the heterogeneity inherent across varied data sources can substantially reduce the false positive rate in causal variable discovery. This reduction in false discovery rate is particularly desirable in many applications, especially in the field of biomedicine and medicine, where precision is paramount given that experimental validation often entails considerable time and resources. To assess the feasibility and validity of the proposed I-GCM approach, we conducted extensive simulations and applied the method to uncover direct causal variables for various traits/diseases with clinical significance.

Our contributions can be summarized as follows:

(1) We combined machine learning approaches and the principles of ICP to identify (direct) causal variables in the presence of a large number of covariates. Previous ICP applications or method primarily focus on the case when only a limited of covariates are present; however, in many applications including biomedical studies, high-dimensional data (such as omics data) is very common and the conventional ICP approach may be difficult to be applied.
(2) While the ICP principle is a very useful, theoretically sound, and novel approach for causal inference, it has observed very limited use in biomedical or clinical studies, especially more substantive applications in large datasets. We are also unaware of ICP being applied to genetic epidemiology studies. Here we applied the proposed approach to the UK-Biobank and identified potential (direct) causal genes leading to various clinical traits/disorders, which to our knowledge is the first study employing ICP in genetic epidemiology.
(3) Related to the above, the finding of specific genes that may be directly causal to diseases such as COVID-19 and traits such as lipid levels are of scientific and clinical importance. The findings may potentially inform drug development and genetic risk prediction, for example.

To the best of our knowledge, our study pioneers the exploration of invariant causal prediction within the realm of genetic epidemiology and human genomics studies. It demonstrates the potential of this innovative method in reliably identifying causal variable sets under high dimensional data settings. It is also the first to discover both genetic and clinical causal risk factors for several important clinical outcomes, including COVID infection, severe COVID, and lipid traits such as high-density lipoprotein (HDL) and triglycerides (TG).

Method

In this study, we proposed the invariant generalized covariance measure(I-GCM), a novel two-stage framework designed to accurately identify causal variables for target variables. This framework comprises two steps, i.e., variable selection, and direct causal variable set identification. For variable selection, we proposed to employ the PC-simple[6] algorithm, which utilizes ordered independence test to screen potential causal variables for the outcome under study. Following this, we employ our newly proposed metric, i.e., invariant generalized covariance measure, to infer the ultimate causal variable set for our studied trait from the preselected variables. This metric leverages the invariance property of causal relations across different experimental settings (e.g., observational and interventional) to identify direct causal variables. In the following sections, we will delve into the details of this method.

Variable selection via PC-simple

As mentioned earlier, we utilized the PC-simple algorithm to first perform variable selection. It employs partial correlation screening to eliminate irrelevant variables for the target variable. We first review the original PC-simple algorithm. Let $X = [X^1, X^2, \dots, X^p]$ be a $n \times p$ matrix for n observations with p variables, Y be a vector for n observations. Suppose Y is defined by the following generative model:

$$Y = \sum_{j=1}^{p} \beta^j X^j + \varepsilon \tag{1}$$

Here, ε indicates the noise item that is independent of X^j and follows a multivariate normal distribution ($\epsilon \sim N(0, \sum)$). Variable with a non-zero β^j is the true direct causal variable for Y. If X^j is independent of Y, i.e., $\beta^j = 0$, then we have:

$$\exists s \subseteq S, \rho(Y, X^j | X^s) = 0 \tag{2}$$

Here, $S = \{1,2, \dots, j-1, j+1, \dots, p\}$ defines the set including variables excluding X^j, s is a subset of S, $\rho(Y, X^j | X^S)$ denotes the partial correlation between X^j and Y. Since the distribution for the correlation coefficient is highly skewed, it's difficult to directly test whether the partial correlation is zero. We could address this issue by employing the Fisher's Z-transform:

$$Z(Y, X^j | X^s) = \frac{1}{2}\{\frac{1 + \hat{\rho}(Y, X^j | X^s)}{1 - \hat{\rho}(Y, X^j | X^s)}\} \tag{3}$$

$\hat{\rho}(Y, X^j | X^s)$ indicates the estimated partial correlation from data. The null hypothesis of independence would be rejected if

$$(n - |s| - 3)^{1/2} \left| Z(Y, X^j | X^S) \right| > \phi^{-1}(1 - \alpha/2) \tag{4}$$

n denotes the sample size, $|s|$ is the cardinality of the set, ϕ indicates the inverse cumulative function for normal distribution. Recursively performing partial correlation screening with increased order could exclude irrelevant variables from previous candidate variables set until they do not vary anymore. The original candidate variable set is S.

In this study, we set the maximum order of $|s|$ to 3 and α to 0.05 for the feature screening process. This implies that variables which survive the 3-order partial correlation screening will be retained for further analysis. For more details about this method, please refer to [6]. The feature dimension q will be dramatically reduced after employing the PC-simple algorithm for feature selection.

Causal variable set identification

While the PC-simple algorithm outperforms commonly used feature selection methods (e.g., Lasso, elastic net, etc.) in selecting causally relevant variables for biomedical data, it still struggles with high false positive rates in high-dimensional data and non-linear relations. To address these challenges, we proposed a novel framework that leverages the invariance property of causal models to identify a direct causal variable set. Causal relations are universal and robust across different experimental settings. The key idea underlying ICP is that the conditional distribution of the target variable *Y* given its direct causes (*Xs*) remains invariant, if we intervene on other variables in the model, except the target itself. Put it in another way, given the complete causal set, the conditional distributions for the target variable across various experimental settings should be identical. This also implies predictions from a causal model will remain consistent across different environments.

In biomedical studies, 'omics' measurements are often made under a combination of interventional and observational settings, or under different interventions, for example knockout of different genes, different drug treatments etc. The ICP approach is a natural choice for causal inference in such scenarios, but it is also useful in observational settings with distinct 'environments'. Intuitively, the causal structure or components should be consistent across different sub-populations, while non-causal components may vary.

In a previous work, Rajen et al.[9] proposed to use generalized covariance measure (GCM) to detect the causal relationships between variables. Notably, GCM We will extend this concept to reliably identify causal variable set in high-dimensional data settings.

Generalized covariance measure (GCM)-based causal variable identification

For a given distribution of random variables (X, Y, Z), we can always decompose the distribution into the following equations:

$$\begin{aligned} X &= f(Z) + \varepsilon_X \\ Y &= g(Z) + \varepsilon_Y \end{aligned} \tag{5}$$

Notably, Z can either be a single variable or a set of variables. We employed the XGBoost (gradient boost trees)[10] to build the prediction models $f(Z)$ and $g(Z)$. Note that there is no restriction on the type of regression or machine learning models for $f(Z)$ and $g(Z)$. In brief, XGBoost is a supervised learning method attempting to predict the target by combining the estimates from a sequential of simpler and weaker tree models. Each new tree is trained based on the residuals from the previous tree using gradient descent to minimize the loss. Given n observations for the variables (x_i, y_i, z_i) $(i = 1,2, \dots n)$, we could calculate the product between residuals(R) from the prediction functions for each observation, i.e.,

$$R_i = (x_i - f(z_i))(y_i - g(z_i)) \tag{6}$$

The normalized sum of R_i (T^n)(a.k.a., generalized covariance measure (GCM)) could be represented as follows:

$$T^n = \frac{\sqrt{n}.\frac{1}{n}\sum_{i=1}^{n} R_i}{(\frac{1}{n}\sum_{i=1}^{n} R_i^2 - (\frac{1}{n}\sum_{r=1}^{n} R_r)^2)^{1/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n} R_i}{[(\frac{1}{n}\sum_{i=1}^{n} R_i^2 - (\frac{1}{n}\sum_{r=1}^{n} R_r)^2)^{1/2}]/\sqrt{n}} \tag{7}$$

T^n could be utilized to test the null hypothesis that X and Y are conditional independent given variable(s) Z. Rajen et al.[9] proved that T^n is asymptotically standard normal. The null hypothesis would be rejected if $|T^n|$ has a larger value, i.e.,

$$|T^n_{X_j^e}| > \phi^{-1}(1 - \alpha/2) \tag{8}$$

The parameter α is set to a default value of 0.05.

This method can be used to detect whether a single variable is causally relevant to the target variable. If the conditional set *Z* encompasses all direct cause for the target variable, then *X* and *Y* are conditionally independent given variable Z. In our case, *X* will be replaced by the environmental variable *E*, i.e., we will test if the target *Y* is independent of the environment *E*, given a set of covariates *Z*[8].

If *Z* contains all direct causal variables, *Y* should be invariant across or independent of *E*. It's worth noting that even if *Z* includes some irrelevant variables, the above-mentioned null hypothesis remains valid.

Direct application of this method for detecting reliable causal variable set may result in the identification of a set with superfluous irrelevant variables. To mitigate this, we proposed to extend this concept and combine it with the invariance property of causal relations to identify reliable causal variable set for the target.

As discussed above, for a given environment variable E, the computed GCM between E and the target should be constant or similarly close to zero when conditioned on the full set of direct causes, even if some other (non-directly causal) variables are included. However, if we exclude some direct causes from the conditional set, the calculated GCM is expected to divert away from the center of the normal distribution. In other words, significant change on the calculated GCM would be observed between the reduced conditional set and full conditional set. If we exclude some irrelevant variables in the conditional set, the calculated GCM should remain stable. Notably, a direct causal variable for the target can also act as an environment variable[7]. If this is the case, the aforementioned statements still hold; under this scenario, the computed GCM between E and the target will be non-zero, but the relationship should be constant, if conditioned on other direct causes. The "full set of direct causes" does not include the environment variable in this case.

We employed backward feature selection to sequentially drop variables in the conditional set and calculate the GCM for each conditional set. The variables were ordered by the derived Zmin value from the PC-simple algorithm in descending order (i.e. the *least* likely causal variables are ranked last, and will be removed from the set first) before performing the backward feature selection. Here Zmin refers to the minimum *z*-statistic in the series of conditional tests from PC-simple; a higher Zmin in general reflects stronger causal relationships. Here we propose a backward elimination procedure as it is practically impossible to enumerate all combinations of all covariates as direct causal variables. This may be considered a 'greedy' approach, but it can greatly reduce the computational burden to make modeling of high-dimensional data possible.

The change in the distance of GCM from zero between two consecutive conditional sets (ΔT^n) could be represented as follows:

$$\Delta T^n_{S_{j-1},S_j} = |T^n_{S_{j-1}}| - |T^n_{S_j}| \tag{9}$$

Here we used the absolute value as we are mainly concerned about the distance of GCM from zero, which reflects the degree of independence of E and Y (GCM=0 indicates complete independence), conditioned on other estimated direct causes. The variance of the distance change of GCM ($\mathrm{Var}(\Delta T^n_{S_{j-1},S_j})$) under the null can be estimated from the following:

$$\begin{aligned}\mathrm{Var}(\Delta T^n_{S_{j-1},S_j}) &= Var\left(\left|\left|T^n_{S_{j-1}}\right|\right|\right) + Var\left(\left|T^n_{S_j}\right|\right) - 2\left[E\left(\left|T^n_{S_{j-1}} * T^n_{S_j}\right|\right) - E\left(\left|T^n_{S_{j-1}}\right|\right) * E\left(\left|T^n_{S_j}\right|\right)\right] \\ &= 2 - \frac{4}{\pi}(\rho * arcsin\rho + \sqrt{1-\rho^2}) \ \ (\text{ref}^{11,12})\end{aligned} \tag{10}$$

where $\rho = Cor(T^n_{S_{j-1}}, T^n_{S_j})$ indicates the correlation between product of residuals calculated from conditional set S_{j-1} and S_j, S_{j-1} and S_j respectively denote two consecutive conditional set with $j-1$ and j variables. Since both $T^n_{S_{j-1}}$ and $T^n_{S_j}$ follow standard normal distribution under the null, $Cor\left(T^n_{S_{j-1}}, T^n_{S_j}\right) = Cov\left(T^n_{S_{j-1}}, T^n_{S_j}\right)$ holds. After replacing both $T^n_{S_{j-1}}$ and $T^n_{S_j}$ with equation 7, we have:

$$Cor\left(T^n_{S_{j-1}}, T^n_{S_j}\right) = Cov\left(T^n_{S_{j-1}}, T^n_{S_j}\right) = \frac{Cov(R^n_{S_{j-1}}, R^n_{S_j})/N}{se\left(R^n_{S_{j-1}}\right) * se\left(R^n_{S_j}\right)} \tag{11}$$

$Cov(R^n_{S_{j-1}}, R^n_{S_j})$ indicates the covariance between product of residuals calculated from conditional set S_{j-1} and S_j, $se\left(R^n_{S_{j-1}}\right)$ and $se(R^n_{S_j})$ respectively indicate the standard error for product of residuals calculated from conditional set S_{j-1} and S_j, and N is the sample size of the target dataset. For more details, please refer to the supplementary text. The null hypothesis of no significant distance change of GCM between two consecutive conditional set would be rejected if:

$$\frac{\Delta T^n}{sqrt\left(\mathrm{Var}(\Delta T^n)\right)} > \phi^{-1}(1-\alpha) \tag{12}$$

Here ϕ^{-1} denotes the cumulative function for normal distribution. It should be noted that the above does not account for multiple testing, as typically multiple hypothesis tests need to be performed. We recommend a principled approach for formal statistical inference by stability selection, as described below.

Algorithm 1 summarizes how we could identify a reliable direct causal variable set by I-GCM. Notably, α and T may vary between different data settings, where T refers to the threshold for ΔT^n. For low-dimensional settings, we may simply set α to 0.05 and T to 0. A more stringent criteria is recommended for high-dimensional settings. We recommend adopting a flexible detection rule based on the number of remaining variables after feature selection. For example, if the investigator wishes to experimentally follow up a limited number of genes, the thresholds may be adjusted such that the number of genes left matches with the number planned for further follow-up studies. As detailed below, a formal statistical approach (stability selection) can be employed to control the expected number of false positive findings.

Algorithm 1: I-GCM for causal variable set identification	
Input: $\boldsymbol{X} \in \boldsymbol{R}^{n\times q}$, $\boldsymbol{Y} \in \boldsymbol{R}^{n}$, $E \in \boldsymbol{R}^{n}$, $\boldsymbol{\alpha}$, T	
For j = **q**: **2**, set $\boldsymbol{S_j} = \{\mathbf{1}, \mathbf{2}, \ldots j\}$, $\boldsymbol{S_{j-1}} = \{\mathbf{1}, \mathbf{2}, \ldots \boldsymbol{j}-\mathbf{1}\}$	
1.	Calculate the GCM between $\boldsymbol{E}$ and Y by conditioning on $\boldsymbol{S_j}$ ($\boldsymbol{T^n_{S_j}}$) and $\boldsymbol{S_{j-1}}$($\boldsymbol{T^n_{S_{j-1}}}$) respectively based on equations 6-7
2.	Compute the distance change between $\boldsymbol{T^n_{S_{j-1}}}$ $\boldsymbol{and}$ $\boldsymbol{T^n_{S_j}}$, and the corresponding variance based on equations 9-10
3.	Test the null hypothesis of no significant distance change. If inequality 12 holds and $\Delta\boldsymbol{T^n} > \boldsymbol{T}$ stop the testing
end	
Output: causal variable set $\boldsymbol{S_j}$	

Note: $\boldsymbol{n}$ indicates number of all observations in environments set E, T denotes the predefined test statistic change

Stability selection for statistical inference

We propose the stability selection method [13] as a principled approach to formally control the expected number of falsely selected variables. In brief, this approach involves repeated subsampling of the data, and variables that are consistently selected across subsamples are more likely to be true positives. Meinshausen and Bühlmann[13] established a formula to control for $E(V)$, the expected number of false positives.

$$E(V) \leq \frac{1}{2\pi_{thr} - 1}\frac{q^2}{p} \tag{13}$$

π_{thr} defines the selection probability cutoff (selection probability refers to the proportion of times a certain feature is selected across all sub-samples), q denotes the average number of identified variables, while p indicates the original number of variables.

In addition, stability selection can help identify the optimal parameters (α and T) for analysis, as different parameters will result in different $E(V)$. The researcher may adjust α and T accordingly to achieve a desired level of $E(V)$. Alternatively, the $E(V)$ can be controlled given a combination of parameters (α and T). Empirically, we set α to 5e-03 and T(change of GCM statistics) to 0.2 for high-dimensional data settings in this study (the remaining feature number is >160 after feature selection in our data applications).

We conducted a further pilot study to illustrate the application and usefulness of stability selection, under a moderate data dimension: $p = 1000, n = 50000, t_e = direct\ cause$. We set the iteration number to 20, subsampling $4n/5$ observations for each iteration. We applied a selection probability threshold (0.9) to determine which variables were consistently selected across iterations. As suggested by Meinshausen et al., π_{thr} has only minor effect on the results when it falls into a reasonable range [0.6, 0.9]

Simulation

To verify the validity of our proposed framework, we simulate different scenarios with varying total variable number (p), sample sizes (n), environment variable types (t_e). Specifically, for each scenario, we randomly generated a directed acyclic graph (DAG) with the nodes representing the variables and the causal effects randomly generated from a given range following a uniform distribution. We employed the function "rmvDAG" in the R package "pcalg" [14] to realize this. The original function is designed to generate multivariate data with dependency structures specified by a given DAG. To make it adaptive to binary variables, we employed a predefined liability threshold(i.e.,0.3) to convert the simulated continuous variable into a binary one. As the threshold was solely used for variable type transformation, it does not affect the validity of our proposed method. The number of direct causes (p_c) in each scenario was determined by graph density. To align with real-

world application scenarios, we set the graph density to 0.05. We considered the following combination of settings:

$$p \in \{200,400,800,1000,2000,3000\}$$
$$n \in \{30000,50000\}$$
$$t_e \in \{\text{ direct cause, ancestor}\}$$

We generated data from randomly chosen DAG following linear Gaussian structural equation models. Also, we generated scenarios with a mixture of linear and non-linear relations with $p \in \{2000, 3000\}$. The simulated datasets were then utilized to identify the reliable causal variable set for the chosen target utilizing our proposed method. For all simulated scenarios, the last variable was chosen as the target variable. The environment variable was randomly selected from the direct causes or ancestors of the target.

We evaluated the efficacy of our proposed method in identifying the direct causal variables for the target variable. Specifically, we utilized 4 different metrics to evaluate the performance, i.e., positive predictive value (PPV, a.k.a., precision), negative predictive value (NPV), sensitivity, specificity. PPV represents the proportion of identified direct causal variables that are true ones while NPV denoted the same for non-direct causal variables. Sensitivity measures the ability in detecting the true direct causal variables while specificity gauges the ability to identify actual non-direct causal variables.

In addition, we compared the performance of our proposed method with the PC-simple algorithm in terms of recovering the direct causes for the target. For our proposed method, we divided the simulated dataset into different subsets based on the selected environment variable. For continuous environment variable, we firstly rank the samples based on the environment variable in ascending order, then split the original simulated dataset into 2 subsets based on the predefined subset size. If the environment variable is discrete, we divide the dataset according to the defined categories.

Real data application

We applied our proposed method to the UK-Biobank(UKBB) dataset, which contains GWAS data and clinical variables, to identify (direct) causal variable set for different phenotypes, i.e., COVID-19 infection, severe COVID-19, high density lipoprotein (HDL) and triglycerides (TG). Here, severe COVID-19 is defined by a combination of hospitalized and fatal cases. Notably, our input data comprises both clinical variables and imputed gene expression data. In this study, we employed "PrediXcan"[15] to impute tissue-specific gene expression levels from the genotypic data of the UK-Biobank subjects. In brief, PrediXcan was based on a prediction model for gene expression levels using elastic-net based regression model on GTEx (a reference dataset with available genotype and gene expression). The genotypic data of the UK-Biobank subjects were then used to "impute" the tissue-specific gene expression levels. The continuous clinical covariates were directly extracted from the UK-Biobank. The binary covariates were defined by ICD10-coded disease in the UK-Biobank. We incorporated imputed gene expression levels from whole blood and the lung for the analysis. Table 1 summarizes the studied phenotypes. For all application scenarios using UK Biobank, we chose sex as the environment variable. This choice is predicated on the well-established physiological and metabolic differences between males and females[16, 17], which can create distinct biological contexts. We hypothesized that while these sex-specific environments might alter the distribution of covariates or the apparent effects of non-causal variables, the conditional distribution of the outcome given its true direct causal variables should remain invariant across sexes, as per the principles underlying ICP. This approach allows us to leverage naturally occurring heterogeneity within the population to refine the identification of direct causal factors.

Table 1 summary for studied phenotypes in UK-Biobank

Phenotype	Sample size	No. of covariates	Tissue
COVID infection	154749	6366	Whole blood
Severe COVID	154749	6366	Whole blood
COVID infection	154749	8037	Lung
Severe COVID	154749	8037	Lung
HDL-C	328002	6349	Whole blood
Triglycerides (TG)	328002	6349	Whole blood

To further validate the reliability of our proposed method in identifying (direct) causal variables, we compared its performance with the PC-simple algorithm in identifying potential targets for COVID. Specifically,

we assessed whether the gene sets identified by our method had an equivalent chance of being listed as targets in Open Targets. The targets associated with COVID were downloaded from Open Targets Platform, which provides measures of relevance (ranging between 0 and 1) between potential targets and COVID based on various factors such as associations with known drugs, hits in relevant GWAS, etc.

To gain a deeper understanding of the biological mechanism underlying the identified causal genes for the target outcome, we performed pathway enrichment on the identified gene set. More specifically, an over-representation analysis was conducted on the identified causal genes using the web-based tool "ConsensuspathDB" [18, 19]. Furthermore, drug enrichment analyses were carried out to identify drugs related to COVID infection and severe COVID.

Results

Simulation results

As mentioned above, stringent detection criteria are desired for high-dimensional data. Table 2 summarizes the simulation results for different scenarios. Our proposed method demonstrated robust performance in identifying direct causal variables for the target, constantly achieving high precision (a.k.a., positive predictive value (PPV)) and specificity (Table 3). Besides, it exhibited satisfactory power (sensitivity) in identifying actual direct causal variables (Fig. 1). Notably, the PPVs exceeded 90% in all scenarios, making it particular useful in selecting variables for follow-up studies. The performance remained stable even with an increase in variable numbers. The F1 and F0.5 scores, which consider both PPV and sensitivity, are also generally higher for the proposed I-GCM method. As expected, the power of our proposed method improved with larger sample sizes. Given that current dataset typically comprises a very large sample size, and the advent of large biobanks like the UK-Biobank has further boosted the availability of large-scale datasets, we believe our proposed method is highly effective in uncovering (direct) causal variables for the target variables in high-dimensional settings.

Table 2 Simulation results for our proposed I-GCM and the PC-simple algorithm

Overall sample size	**Environment variable type**	**No. of input variables**	**No. of overall true causal**	**PC-simple**			**I-GCM**		
				total no.	true causal	false causal	total no.	true causal	false causal
50000	Direct cause	200	11	13	9	4	7	7	0
50000	Direct cause	400	21	27	17	10	15	15	0
50000	Direct cause	800	39	38	28	10	31	28	3
50000	Direct cause	1000	51	57	43	14	46	42	4
50000	Ancestor	1000	51	59	46	13	48	45	3
30000	Direct cause	1000	51	54	36	18	33	33	0
50000	Direct cause	2000	114	121	87	34	90	82	8
50000	Ancestor	2000	114	110	87	23	99	86	13
30000	Direct cause	2000	114	110	77	33	64	64	0
50000	Direct cause	3000	163	215	122	93	111	108	3
50000	Ancestor	3000	163	216	121	95	128	110	18
30000	Direct cause	3000	163	168	114	54	115	97	18

Table 3 Comparison of different evaluation metrics between our proposed I-GCM and the PC-simple algorithm

Overall sample size	**Environment variable type**	**No. of input variables**	**No. of overall true causal**	**PC-simple**						**I-GCM**					
				PPV	Sensitivity	NPV	Specificity	F1-score	F0.5-score	PPV	Sensitivity	NPV	Specificity	F1-score	F0.5-score
50000	Direct cause	200	11	0.692	0.818	0.989	0.979	0.750	0.714	1.000	0.636	0.979	1.000	0.778	0.897
50000	Direct cause	400	21	0.630	0.810	0.989	0.974	0.709	0.659	1.000	0.714	0.985	1.000	0.833	0.926

50000	Direct cause	800	39	0.737	0.718	0.986	0.987	0.727	0.733	0.903	0.718	0.986	0.996	0.800	0.859
50000	Direct cause	1000	51	0.754	0.843	0.990	0.987	0.796	0.770	0.913	0.824	0.990	0.998	0.866	0.894
50000	Ancestor	1000	51	0.780	0.902	0.995	0.986	0.837	0.802	0.938	0.882	0.994	0.997	0.909	0.926
30000	Direct cause	1000	51	0.667	0.706	0.984	0.981	0.686	0.674	1.000	0.647	0.981	1.000	0.786	0.902
50000	Direct cause	2000	114	0.719	0.763	0.988	0.980	0.740	0.727	0.911	0.719	0.983	0.999	0.804	0.865
50000	Ancestor	2000	114	0.791	0.763	0.986	0.988	0.777	0.785	0.869	0.754	0.985	0.993	0.807	0.843
30000	Direct cause	2000	114	0.700	0.675	0.980	0.983	0.687	0.695	1.000	0.561	0.974	1.000	0.719	0.865
50000	Direct cause	3000	163	0.567	0.748	0.985	0.967	0.645	0.596	0.973	0.663	0.981	0.999	0.789	0.890
50000	Ancestor	3000	163	0.560	0.742	0.985	0.967	0.638	0.589	0.859	0.675	0.982	0.994	0.756	0.815
30000	Direct cause	3000	163	0.679	0.699	0.983	0.981	0.689	0.683	0.843	0.595	0.977	0.994	0.698	0.778

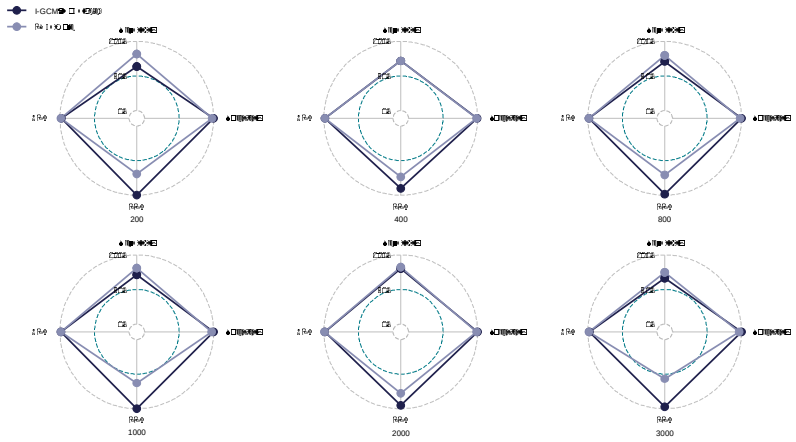

Fig.1 Comparison of performance in causal variables discovery between our proposed I-GCM and the PC-simple algorithm

Both direct cause and ancestor variables can be selected as environment variables to identify direct causal variables. They demonstrated comparable performance in terms of the prediction metrics (refer to Table 3). It's noteworthy that the direct cause seemed to be a more potent candidate for environment variable in eliminating false positives. We hypothesize that interventions on direct cause probably induced more discernable and detectable heterogeneity in the corresponding subsets than ancestors, aiding in distinguishing true causal variables from false ones. This observation aligns with current studies on environment variable identification for causal inference[20-22].

We also compared our method with the commonly used PC-simple algorithm in high-dimensional settings. Our method outperformed PC-simple algorithm in identifying direct causal variables, yielding high PPV and specificity while maintaining comparable power in identifying actual direct causal variables. Most importantly, we observed that I-GCM substantially improves the PPV, with the difference in the PPV between the I-GCM and the PC-simple reaching 40% in some scenarios (refer to Fig. 2). The improvement was particularly pronounced in high-dimensional settings. Even though the sensitivity for our proposed method was sometimes slightly lower than that for the PC-simple algorithm, it achieved a substantially higher PPV, and that I-GCM consistently outperformed PC-simple in the F1 and F0.5 scores (which considers both the PPV and sensitivity).

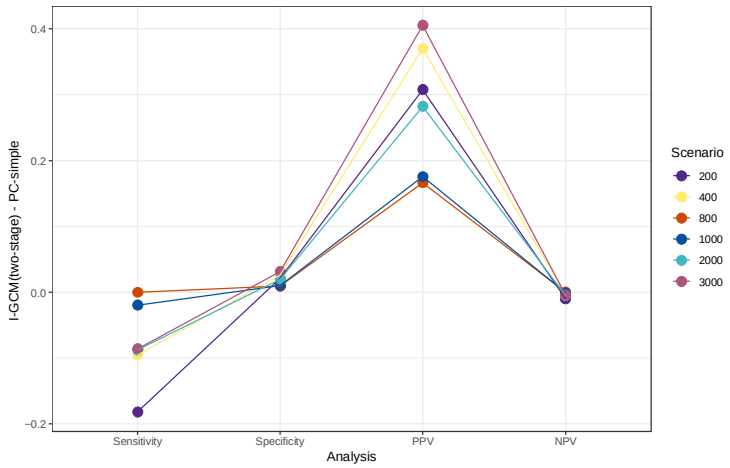

Fig.2 The differences regarding 4 evaluation metrics between the proposed I-GCM and the PC-simple algorithm

We also evaluated the performance of the PC-simple between a stringent (i.e., $\alpha = 0.01$) and the default alpha cutoff (i.e., $\alpha = 0.05$). As anticipated, a decrease in alpha modestly improved the PPV, albeit at the expense of reduced power (refer to Fig. 3, Table 4). Most importantly, we discovered that the I-GCM could still significantly improve the PPV even with a stringent cutoff, particularly in high-dimensional settings (refer to Fig. 4). This further proved the reliability and superiority of our proposed method. It's also noteworthy that the proposed I-GCM constantly delivers satisfactory performance even in the presence of non-linear relations (Table S1).

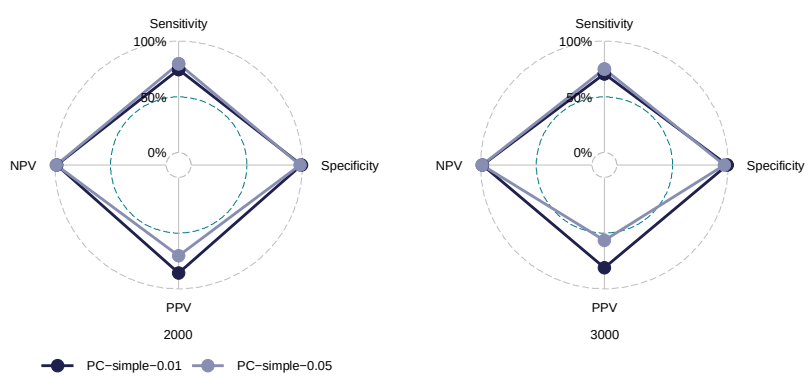

Fig. 3 Comparison of performance in causal variables discovery by the PC-simple algorithm with different cutoffs

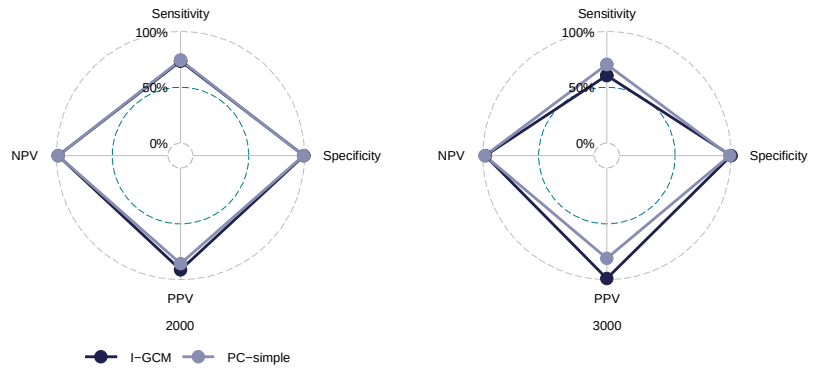

Fig. 4 Comparison of performance in causal variables discovery between our proposed I-GCM and the PC-simple algorithm with a more stringent alpha cutoff (i.e., 0.01)

Table 4 Comparison of the performance between our proposed I-GCM and the PC-simple algorithm under different alpha cutoffs

Alpha for PC-simple	Environment variable type	No. of input variables	No. of overall true causal	PC-simple						I-GCM					
				PPV	Sensitivity	NPV	Specificity	F1-score	F0.5-score	PPV	Sensitivity	NPV	Specificity	F1-score	F0.5-score
0.01	Direct cause	2000	114	0.746	0.859	0.993	0.985	0.797	0.831	0.913	0.737	0.984	0.995	0.815	0.891
0.05	Direct cause	2000	114	0.719	0.763	0.988	0.98	0.74	0.727	0.911	0.719	0.983	0.999	0.804	0.865
0.01	Direct cause	3000	163	0.81	0.705	0.983	0.99	0.754	0.861	0.99	0.607	0.978	0.999	0.746	0.961
0.05	Direct cause	3000	163	0.567	0.748	0.985	0.967	0.645	0.596	0.973	0.663	0.981	0.999	0.789	0.89

As described above, we also performed a pilot study with stability selection, with $p = 1000\ and\ n = 50000$. Under a π_{thr} threshold of 0.9, on average 48 variables were selected in each subsampling run. The expected number of falsely selected variables was estimated to be 3, under our recommended settings of α = 5e-03 and T (change of GCM statistics)= 0.2. Our model achieved superior performance (PPV=0.963, F1 score=0.932, F0.5 score=0.949) compared with the PC-simple method (PPV=0.706, F1 score=0.807, F0.5 score=0.843). Although this testing scenario involved a moderate feature size, we expect stability selection to demonstrate similar performance in higher-dimensional settings.

Results for real data application

We applied our method to several clinical traits to identify their corresponding causal variables. Table 5 summarizes the number of identified causal variables using "predicted" tissue-specific gene expression levels and clinical variables. The included clinical variables were the same for both COVID infection and severe COVID. However, different clinical variables were included for HDL and triglycerides. Only target-relevant clinical variables were included for the causal variables set identification. All the included clinical variables were extracted from the UK-Biobank dataset with application no. 37268. As illustrated in Table 5, the number of direct causal variables identified for the same target varied when gene expression profiles from different tissues were included. Despite this variation, the identified causal clinical variables remained relatively stable across tissues.

Table 5 Identified causal variables for different traits

Phenotype	No. of total causal variables	No. of clinical variables	No. of genes	Tissue
COVID infection	36	16	20	Whole blood
Severe COVID	90	31	59	Whole blood

COVID infection	70	17	53	Lung
Severe COVID	156	33	123	Lung
HDL	19	8	11	Whole blood
Triglycerides (TG)	58	16	42	Whole blood

Causal variable identification results for COVID infection

For the blood-specific analysis, 36 (direct) causal variables were identified, while 70 were identified for the lung-specific analysis. Even though the overall causal variables varied between tissues, the (direct) causal clinical variables were relatively stable across tissues for COVID infection. The identified clinical variables include age, doses of vaccine, Townsend index, BMI, LDL, ethnic group, coronary artery disease (CAD), dementia, atrial fibrillation (AF), and so on (for full details, please refer to Table S2). Encouragingly, all these clinical variables have been reported to be associated with COVID infection based on current studies. For example, age has been identified as the most important causal variable for COVID infection both for the analyses with the whole blood and the lung. Existing studies have demonstrated that older people, especially those over 70, are more susceptible to severe COVID infection than young adults[23, 24].

The identified direct causal genes for COVID infection differed across tissues. We separately identified 20 and 53 direct causal genes in the whole blood and the lung. More causal genes were identified based on imputed expression from the lung. Interestingly, we found that the ranking of identified causal genes varied between tissues. This aligned with our expectation, as COVID is a lung-related disease, and many disease-related genes are tissue specific. We compared the performance of our method with the PC-simple algorithm in identifying potential targets using the Fisher's exact test. Our proposed method performed comparably with the PC-simple, although it tended to identify fewer variables. Many of the identified genes have been reported to be COVID-associated by existing studies, with some even identified as potential drug targets for COVID (Table 6, full results please refer to Table S2), e.g., *CCR5*, *FYCO1*, *KCNN4*, etc. Pattersson et al.[25] reported that inhabiting *CCR5* in COVID patients could lead to decreased inflammatory cytokines and COVID RNA in the plasma. *FYCO1* has been identified COVID-associated gene from several GWAS[26, 27]. More specifically, abnormal expression of this gene may lead to respiratory failure for COVID patients[28].

As expected, many of the enriched pathways have been reported to be relevant to COVID or related pathophysiology (Table S3). Here we highlighted a few interesting pathways. As one of the top enriched pathways for COVID infection, the N-acetylglucosamine (NAC) degradation pathway has been reported to be involved in chronic anti-inflammatory reactions, and a clinical study reported that NAC may lead to reduced hospital stay and ICU admission for COVID patients[29]. Acute viral myocarditis was another top-enriched pathway. Notably, acute myocarditis has been reported as a complication of COVID infection [30]. The mitochondrial transcription initiation pathway, which was significantly enriched for COVID infection, has been reported to be associated with the mediation effects of COVID on innate immunity[31].

Table S4 demonstrates the drug enrichment results for COVID infection. Encouragingly, some of the enriched drugs have been tested/used to treat COVID patients in clinical practice. For example, ethinylestradiol was significantly enriched for COVID infection based on the gene sets identified from the whole blood. There is an ongoing clinical trial to evaluate whether estrogen therapy could lead to reduced hospitalization stays in non-severe COVID patients[29]. Another study showed that recent hormone replacement therapy (HRT) was associated with reduced all-cause mortality in patients with COVID-19[32]

Causal variable identification results for severe COVID

For severe COVID, we identified 90 and 156 (direct) causal variables for the whole blood-specific and lung-specific analyses, respectively. Again, the identified clinical causal variables remained stable across tissues (i.e., whole blood and lung). The identified clinical variables include doses of vaccine, stroke, Townsend index, non-COVID pneumonia, type 2 diabetes mellitus (T2DM), renal failure, chronic obstructive pulmonary disease (COPD), renal failure, heart failure, venous thrombus embolism (VTE), waist circumference, HDL etc. (full results please refer to Table S2). Again, all identified clinical variables have been reported as risk factors for COVID severity in existing studies. For example, two doses of COVID vaccine was identified as the most important direct causal variable from our study. This finding aligns with most existing studies on COVID

severity, which acknowledged that two doses of mRNA were highly effective in preventing hospital admission/death from COVID. A meta-analysis on 7,267,055 COVID patients revealed that stroke could lead to increased COVID mortality by an effect size of 1.3[33]. T2DM was also one of the earliest identified risk factors for COVID severity[34]. COVID patients with T2DM have been reported to demonstrate poor therapeutic effects[35]. Other identified causal variables like stroke, non-COVID pneumonia, townsend index, renal failure, and COPD were also well-acknowledged risk factors for COVID severity.

We respectively identified 59 and 123 causal genes for the whole blood and the lung (Table S2). Similar to COVID infection, the rankings of the identified direct causal genes were tissue specific. We compared our identified causal genes with those from the PC-simple algorithm to examine their potential to be listed as targets for COVID. We conducted fisher's exact test to examine whether there exist significant detection power differences between our method and the PC-simple algorithm. As expected, comparable power was observed even though fewer causal genes were identified by our method. In other words, the sensitivity for our method is not significantly different from that for PC-simple algorithm. Encouragingly, many of the identified genes have been reported to be associated with severe COVID in existing studies, e.g., *DNZL, JAK1, CCR5, PDE8B, POM121*, etc. (Table 6, for more details, please refer to Table S4). Anderini et al.[36] reported that DNLZ could affect the binding of zinc(II), which may further lead to an inefficient immune response to COVID infection. Chen et al.[37] revealed that inhibition of JAK1 could lead to reduced cytokine release syndrome. It is also worth noting that baricitinib, a treatment that targets JAK1, has been used to treat COVID patients in clinical practice.

We also performed pathway enrichment analysis on the identified causal gene set (Table S3). Encouragingly, many of the enriched pathways were significantly associated with the pathophysiology of COVID or related complications. For example, other interleukin signaling was one of the top-enriched pathways for severe COVID. Interleukin signaling pathways were intensively involved in anti-inflammatory activity [38], and thus may be used for early identification of severe COVID patients. Membrane Trafficking was also identified as a top-enriched pathway. As suggested by Banerjee et al.[39], interrupted membrane protein trafficking could lead to the disruption of signal recognition particles, which in turn could promote the propagation of COVID. Vesicle-mediated transport was another significantly enriched pathway. In a previous study by Hassanpour et al.[40], exosome (an extracellular vesicle) may play a role in COVID-19 virus infection.

Drug enrichment analysis was also performed for severe COVID (Table S4). Again, we found some of the enriched drugs have been tested/used for treating COVID patients. Here we highlight a few examples. Doxorubicin was one of the top-enriched drugs for severe COVID. Several existing studies have implicated its clinical usefulness in treating COVID patients[41-43]

Table 6 Examples of identified direct causal genes for studied clinical traits

Phenotype	Direct causal genes	Tissue
COVID infection	*CCR5, HLA-DPB2, JAK1, SPR14, RNF138*	Whole blood
Severe COVID	*DNLZ, JAK1, PUS10, GPM6A, ORAI1*	Whole blood
COVID infection	*FYCO1, KCNN4, VPS16, TULP2, RPUSD4*	Lung
Severe COVID	*POM121, PRSA,CCR5, PDE8B, FYC01*	Lung
HDL	*LPL, CCDC92, ACP2, FAM76B, GPR180*	Whole blood
Triglycerides (TG)	*BACE1, NRBP1, LPL, TRIM74, SLC56A*	Whole blood

Causal variable identification results for HDL

In total, we identified 19 direct causal variables for HDL, comprising 8 clinical variables and 11 genes (Table S2). The identified clinical variables included apolipoprotein A, TG, waist-hip ratio (WHR), cholesterol, hemoglobin (HB), low-density lipoprotein (LDL), BMI, and hip circumference. These clinical variables have been reported to be important indicators for HDL levels. For example, apolipoprotein A is the major apolipoprotein for HDL, and its concentration level in the plasma could directly reflect that of HDL. TG and HDL are well-known risk factors for coronary artery disease (CAD). Jeppesen et al.[44] demonstrated that the risk for CAD could be substantially reduced with low TG and high HDL levels. In a previous study, Hamalainen et al.[45] revealed that HB concentration in the plasma was closely associated with HDL particle size. More specifically, a high HB level could lead to large HDL particles, which is associated with an increased risk for various cardiovascular diseases like diabetes and metabolic syndromes. Furthermore, many of the identified causal genes have been demonstrated to be closely related to HDL (Table 6, full results please refer to Table

S2c). For example, a previous study by Xiao et al.[46] reported that *CCDC92* has effect on the size and concentration of HDL particles in plasma. As suggested by Tsutsumi et al.[47], a decrease in LPL activity was associated with unfavorable HDL levels in the plasma.

Pathway enrichment analysis was also performed for HDL (Table S3). Some top pathways include 'composition of lipid particles', 'metabolic pathway of LDL, HDL and TG, including diseases', and 'role of ppar-gamma coactivators in obesity and thermogenesis were found to be significantly enriched for HDL. Kersten[48] reported that the activation of PPAR receptors could lead to an increased HDL level in plasma. For more details about the enriched pathways, please refer to Table S3.

Causal variable identification results for TG

For TG, we identified a total of 58 direct causal variables, with 16 clinical variables and 42 genes (Table S2). The identified clinical variables included HDL, apolipoprotein A, glucose, glycated hemoglobin (HbA1c), ethnic group (black or not), lipoprotein A, cholesterol, T2DM, hypertension (HTN), smoking status, bipolar, etc (full results please refer to Table S2d). As expected, these clinical variables have been shown to be associated with TG. For instance, Srinivasan et al.[49] reported that poor glucose metabolism is associated with high TG levels. Also, TG and glucose can serve as a cost-effective marker for insulin resistance. A study by Naqvi et al.[50] showed that HbA1c can act as an indicator of TG level in the plasma. High HbA1c concentration was shown to reflect unfavorable TG levels in the plasma.

In addition, many identified genes were found to be closely associated with TG (Table 6, full results please refer to Table S2d). For example, Meankin et al.[51] suggested that the loss of BACE1 could lead to unfavorable lipid levels. *NRBP1* has been identified as a susceptibility gene for TG based on an independent GWAS study by Read et al.[52].

Table S3 demonstrates the pathway enrichment analysis result for TG. Here we will highlight a few top pathways. Integrin signaling pathway was one of the top-enriched pathways based on the identified causal gene set. In a related study, Xiao et al.[53] demonstrated that integrin $\beta 3$ deficiency was associated with elevated triglyceride levels. Cholesterol metabolism was another significantly enriched pathway for TG. According to Feingold[54], the removal of triglycerides in very low-density lipoprotein was associated with cholesterol levels. NRF2 pathway was also found to be significantly enriched for TG. In a study by Tanaka et al.[55], the NRF2 pathway was implicated to inhibit the accumulation of triglycerides in the blood.

Discussion

Overview

This study introduced a novel method, I-GCM, to identify direct causal variables for the target of interest. The method combines the GCM (developed for causal relations detection) with the invariance property of causal relationships for effective causal variables discovery. Simulation results validated the efficacy of the proposed method in detecting actual causal variables. Notably, our method consistently outperformed the PC-simple algorithm in uncovering true direct causal variables, especially in terms of higher PPV, while also maintaining comparable sensitivity.

Also, we applied our proposed approach to 2 binary and 2 continuous traits extracted from the UK Biobank (i.e., COVID infection, severe COVID, HDL-C, and TG) to uncover the corresponding causal clinical and genetic variable sets in the whole blood and lung tissues. Encouragingly, most of the identified clinical causal variables are known risk factors for the studied traits. Additionally, we found that the gene sets identified by our method were significantly enriched in pathways involved in the pathophysiology of the studied traits. Given the satisfactory performance of our proposed method in identifying true causal variables, it proves particularly useful for selecting genes for follow-up studies. Furthermore, the identified causal genes may serve as potential targets for novel treatments and drugs.

Strengths

The proposed framework I-GCM has several strengths. A key advantage is its superior performance in uncovering direct causal variables (especially in terms of PPV or precision) compared to the PC-simple algorithm alone, while maintaining comparable sensitivity, especially in high-dimensional settings. Importantly, we showed that integrating structural causal discovery methods (e.g. PC algorithm-based methods) with the more recently proposed invariance-based methods (ICP) may improve both causal discovery performance and computational speed. This advancement opens up possibilities for applying ICP methods to high-dimensional data.

Another strength of our proposed method is its ability to capture both linear and non-linear relationships, eliminating the need for prior knowledge about the underlying generative model. We employed the GCM as a measure of dependence, which allows non-linear relationship between the covariates and the outcome or environment. Many existing or common causal inference approaches, such as PC/PC-simple and the original ICP approach, assume linear relationships between the variables. Although we utilized gradient boosted trees (XGBoost) in our study, any machine learning methods, including deep neural networks, could be employed to calculate GCM to detect the causal relations between variables and the target. It is also worth noting that our method combines datasets collected from different experimental settings (e.g., observational/interventional datasets) and leverage them for causal discovery. This makes our method a natural fit for datasets from different sources, data obtained under different interventions (e.g. knockout or perturbation of different genes, different medication treatment, different experimental conditions etc.), or data obtained under both interventional and observational settings. However, it may be challenging to use the PC-simple algorithm per se in such situations as it is not straightforward how the different kinds of data can be combined together. (In our proposed framework, PC-simple is just used for pre-screening but not as the final step for causal discovery; one may perform pre-screening separately in different environments and take the union of results).

Furthermore, performing variable selection before causal variable discovery substantially reduces computation complexity. Since we ranked the preselected variables and used a backward feature selection method to enumerate the candidate causal variable sets, we dramatically reduce the number of candidate sets ($q\ vs\ 2^q$), leading to improved computational speed. To the best of our knowledge, this study is the first to exploit invariant causal prediction in genetic epidemiology studies, and the first to apply ICP principles in biobank-scale datasets with high-dimensional genomic data. Given the increasing availability of large-scale datasets, our method shows great potential for identifying reliable causal variables for various diseases. The identified causal variables could provide insights into underlying disease mechanisms and inform more effective treatment and prevention strategies.

Limitations

There are a few limitations in this study. The PC-simple algorithm, used to preselect a subset of features for further analysis, was designed to handle linear relations, and may not adequately capture all non-linear associations. On the other hand, pre-selection of variables based on more complex algorithms to account for non-linearity is likely computationally too demanding. In addition, in practice many strong causal factors may also have a linear component. Our simulation results demonstrated that our proposed I-GCM performed reasonably well in the presence of non-linear relations. Also, when identifying the causal variable set for the target, we only consider single-directional relations, thereby ignoring reverse causality. This could potentially lead to the discovery of false positives. However, since the expression profiles were "predicted" from genotypes, reverse causality between genes and the target was highly unlikely. Incorporating timestamps could also address this issue, as effects cannot precede causes. In this study, when extracting exposures and covariates for the outcome, we only took the covariates measured before the outcome occurred. Furthermore, domain knowledge could be exploited to mitigate this problem. This could serve as a promising direction for future work.

In summary, we have proposed a novel framework for causal variable discovery with high precision. We consider our method a useful tool to prioritize variables, especially genes, for follow-up studies. Our proposed framework is flexible and may be extended to other omics studies. For example, given the substantial increase of single-cell RNA-sequencing (scRNA-seq) datasets in recent years, our proposed I-GCM may represent a new avenue for causal analysis based on multiple scRNA-seq datasets. Furthermore, the proposed I-GCM is a useful extension to existing causal inference methods.

Data availability: UK biobank data is available to any researchers who formally apply for the data. However, the data is not publicly available due to privacy concerns.

Code availability: The source codes and R package to reproduce our experiments for this work will be available upon publication of this study.

Conflicts of interest

The authors declare no relevant conflicts of interest.

Acknowledgements

This work was supported partially by a Young Collaborative research grant (C4003-23Y), a National Natural Science Foundation China grant (81971706), the Lo Kwee Seong Biomedical Research Fund from The Chinese University of Hong Kong and the KIZ-CUHK Joint Laboratory of Bioresources and Molecular Research of Common Diseases, Kunming Institute of Zoology and The Chinese University of Hong Kong, China.

References

1. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*. 2018;113(523):1228–1242.

2. Athey S, Tibshirani J, Wager S. Generalized random forests. . 2019.

3. Imbens GW. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*. 2020;58(4):1129–1179.

4. Yin L, Feng Y, Lau A, Qiu J, Sham P, So H. A bayesian network-based framework to uncover the causal effects of genes on complex traits based on GWAS data. *medRxiv*. 2022:2022.12. 25.22283943.

5. Sobel ME. Causal inference in the social sciences. *Journal of the American Statistical Association*. 2000;95(450):647–651.

6. Bühlmann P, Kalisch M, Maathuis MH. Variable selection in high-dimensional linear models: Partially faithful distributions and the PC-simple algorithm. *Biometrika*. 2010;97(2):261–278.

7. Peters J, Bühlmann P, Meinshausen N. Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society.Series B (Statistical Methodology)*. 2016:947–1012.

8. Heinze-Deml C, Peters J, Meinshausen N. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*. 2018;6(2).

9. Shah RD, Peters J. The hardness of conditional independence testing and the generalised covariance measure. . 2020.

10. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. . 2016:785–794.

11. Haas M. A note on the absolute moments of the bivariate normal distribution. *Economics Bulletin*. 2018;38(1):650–656.

12. Shojaie SRH, Aminghafari M, Mohammadpour A. On the expected absolute value of a bivariate normal distribution. *Journal of Statistics Theory and Applications*. 2012;11(4):371–377.

13. Kalisch M, Mächler M, Colombo D, Maathuis MH, Bühlmann P. Causal inference using graphical models with the R package pcalg. *Journal of statistical software*. 2012;47:1–26.

14. Gamazon ER, Wheeler HE, Shah KP, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet*. 2015;47(9):1091–1098.

15. Kamburov A, Stelzl U, Lehrach H, Herwig R. The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res*. 2013;41(D1):D793–D800.

16. Herwig R, Hardt C, Lienhard M, Kamburov A. Analyzing and interpreting genome data at the network level with ConsensusPathDB. *Nature protocols*. 2016;11(10):1889–1907.

17. Creager E, Jacobsen J, Zemel R. Environment inference for invariant learning. . 2021:2189–2200.

18. Heinze-Deml C, Meinshausen N. Conditional variance penalties and domain shift robustness. *arXiv preprint arXiv:1710.11469*. 2017.

19. Arjovsky M, Bottou L, Gulrajani I, Lopez-Paz D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*. 2019.

20. Goldstein E, Lipsitch M, Cevik M. On the effect of age on the transmission of SARS-CoV-2 in households, schools, and the community. *J Infect Dis*. 2021;223(3):362–369.

21. Davies NG, Klepac P, Liu Y, Prem K, Jit M, Eggo RM. Age-dependent effects in the transmission and control of COVID-19 epidemics. *Nat Med*. 2020;26(8):1205–1211.

22. Patterson BK, Seethamraju H, Dhody K, et al. CCR5 inhibition in critical COVID-19 patients decreases inflammatory cytokines, increases CD8 T-cells, and decreases SARS-CoV2 RNA in plasma by day 14. *International Journal of Infectious Diseases*. 2021;103:25–32.

23. Shelton JF, Shastri AJ, Ye C, et al. Trans-ancestry analysis reveals genetic and nongenetic associations with COVID-19 susceptibility and severity. *Nat Genet*. 2021;53(6):801–808.

24. Pairo-Castineira E, Clohisey S, Klaric L, et al. Genetic mechanisms of critical illness in COVID-19. *Nature*. 2021;591(7848):92–98.

25. Severe Covid-19 GWAS Group. Genomewide association study of severe covid-19 with respiratory failure. *N Engl J Med*. 2020;383(16):1522–1534.

26. Hassan AE. An observational cohort study to assess N-acetylglucosamine for COVID-19 treatment in the inpatient setting. *Annals of Medicine and Surgery*. 2021;68:102574.

27. Beşler MS, Arslan H. Acute myocarditis associated with COVID-19 infection. *Am J Emerg Med*. 2020;38(11):2489. e1–2489. e2.

28. Miller B, Silverstein A, Flores M, et al. Host mitochondrial transcriptome response to SARS-CoV-2 in multiple cell models and clinical samples. *Scientific reports*. 2021;11(1):3.

29. Wang X, Liu Z, Li J, et al. Impacts of type 2 diabetes on disease severity, therapeutic effect, and mortality of patients with COVID-19. *The Journal of Clinical Endocrinology & Metabolism*. 2020;105(12):dgaa535.

30. Dambha-Miller H, Hinton W, Wilcox CR, Joy M, Feher M, De Lusignan S. Mortality in COVID-19 among women on hormone replacement therapy: A retrospective cohort study. *Fam Pract*. 2022;39(6):1049–1055.

31. Li S, Ren J, Hou H, et al. The association between stroke and COVID-19-related mortality: A systematic review and meta-analysis based on adjusted effect estimates. *Neurological Sciences*. 2022;43(7):4049–4059.

32. Norouzi M, Norouzi S, Ruggiero A, et al. Type-2 diabetes as a risk factor for severe COVID-19 infection. *Microorganisms*. 2021;9(6):1211.

33. Andreini C, Arnesano F, Rosato A. The zinc proteome of SARS-CoV-2. *Metallomics*. 2022;14(7):mfac047.

34. Chen C, Wang J, Li H, Yuan L, Gale RP, Liang Y. JAK-inhibitors for coronavirus disease-2019 (COVID-19): A meta-analysis. *Leukemia*. 2021;35(9):2616–2620.

35. Rodríguez-Hernández MÁ, Carneros D, Núñez-Núñez M, et al. Identification of IL-6 signalling components as predictors of severity and outcome in COVID-19. *Frontiers in Immunology*. 2022;13.

36. Banerjee AK, Blanco MR, Bruce EA, et al. SARS-CoV-2 disrupts splicing, translation, and protein trafficking to suppress host defenses. *Cell*. 2020;183(5):1325–1339. e21.

37. Hassanpour M, Rezaie J, Nouri M, Panahi Y. The role of extracellular vesicles in COVID-19 virus infection. *Infection, Genetics and Evolution*. 2020;85:104422.

38. Al-Motawa MS, Abbas H, Wijten P, et al. Vulnerabilities of the SARS-CoV-2 virus to proteotoxicity—opportunity for repurposed chemotherapy of COVID-19 infection. *Frontiers in pharmacology*. 2020;11:585408.

39. Sajid Jamal QM, Alharbi AH, Ahmad V. Identification of doxorubicin as a potential therapeutic against SARS-CoV-2 (COVID-19) protease: A molecular docking and dynamics simulation studies. *Journal of Biomolecular Structure and Dynamics*. 2022;40(17):7960–7974.

40. Singh MB, Vishvakarma VK, Lal AA, Chandra R, Jain P, Singh P. A comparative study of 5-fluorouracil, doxorubicin, methotrexate, paclitaxel for their inhibition ability for mpro of nCoV: Molecular docking and molecular dynamics simulations. *Journal of the Indian Chemical Society*. 2022;99(12):100790.

41. Jeppesen J, Hein HO, Suadicani P, Gyntelberg F. Low triglycerides–high high-density lipoprotein cholesterol and risk of ischemic heart disease. *Arch Intern Med*. 2001;161(3):361–366.

42. Hämäläinen P, Saltevo J, Kautiainen H, Mäntyselkä P, Vanhala M. Hemoglobin level and lipoprotein particle size. *Lipids in Health and Disease*. 2018;17(1):1–6.

43. Xiao L, Shi D, Zhang H, et al. Association between single nucleotide polymorphism rs11057401 of CCDC92 gene and the risk of coronary heart disease (CHD). *Lipids in health and disease*. 2018;17:1–5.

44. Tsutsumi K. Lipoprotein lipase and atherosclerosis. *Current vascular pharmacology*. 2003;1(1):11–17.

45. Kersten S. Peroxisome proliferator activated receptors and lipoprotein metabolism. *PPAR research*. 2008;2008.

46. Srinivasan K, Viswanad B, Asrat L, Kaul CL, Ramarao P. Combination of high-fat diet-fed and low-dose streptozotocin-treated rat: A model for type 2 diabetes and pharmacological screening. *Pharmacological research*. 2005;52(4):313–320.

47. Naqvi S, Naveed S, Ali Z, et al. Correlation between glycated hemoglobin and triglyceride level in type 2 diabetes mellitus. *Cureus*. 2017;9(6).

48. Meakin PJ, Harper AJ, Hamilton DL, et al. Reduction in BACE1 decreases body weight, protects against diet-induced obesity and enhances insulin sensitivity in mice. *Biochem J*. 2012;441(1):285–296.

49. Read RW, Schlauch KA, Lombardi VC, et al. Genome-wide identification of rare and common variants driving triglyceride levels in a nevada population. *Frontiers in Genetics*. 2021;12:639418.

50. Xiao B, Mao J, Sun B, et al. Integrin β3 deficiency results in hypertriglyceridemia via disrupting LPL (lipoprotein lipase) secretion. *Arterioscler Thromb Vasc Biol*. 2020;40(5):1296–1310.

51. Feingold KR. Introduction to lipids and lipoproteins. *endotext [internet]*. 2021.

52. Tanaka Y, Aleksunes LM, Yeager RL, et al. NF-E2-related factor 2 inhibits lipid accumulation and oxidative stress in mice fed a high-fat diet. *J Pharmacol Exp Ther*. 2008;325(2):655–664.