Tsinghua Science and Technology
Doi: 10.26599/TST.2026.90100015

Research Article

CDAF: A Causality Driven Analysis Framework for Disease Risk Factor Identification

Wuyan Zhao, Peng Li, Canqing Yu, Wei Wei* (✉), and Xu Yang* (✉)

Abstract Exploring potential disease risk factors is critical for precision medicine, enabling the identification of biomarkers and supporting accurate diagnostic tools and interventions. However, traditional statistical and machine learning methods often fail to reduce disease risk effectively, as they focus on correlations while ignoring essential causal relationships. Causality, on the other hand, could identify causal relationships and predict effects from large-scale observational data, discovering underlying data generation mechanisms. Therefore, we propose a novel Causality Driven Analysis Framework (CDAF) to identify causal risk factors from observational data. CDAF employs a converging causality growth algorithm to discover essential causal relationships through Bayesian scoring criterion, modeled as a causal graph. It estimates the average treatment effects (ATE) of direct causal edges using causal inference, ranking the importance of risk factors. CDAF demonstrated excellent performance on both benchmark datasets and other real clinical datasets. We empirically demonstrate that our framework offers a powerful tool for uncovering complex disease mechanisms, advancing precision medicine, and enabling personalized treatment strategies.

Keywords causality; precision medicine; disease risk factor

1 Introduction

Precision medicine is a novel approach to understanding health and disease based on patient-individual data [1]. The primary aim is to accurately identify the causes and therapeutic targets of diseases, classify them, and achieve personalized treatments [2]. The exploration of potential disease risk factors in the context of precision medicine aligns with the current development needs of public health. It facilitates early detection and treatment of diseases, which are crucial for decision-making in preventive medicine [3]. Meanwhile, the identification of disease risk factors allows the identification of disease-related biomarkers, facilitating early detection of diseases and the development of more accurate diagnostic tools, interventions and personalized treatments. In general, exploring potential disease risk factors covers a wide range of areas, from basic research to clinical applications, involving various disciplines. It is not only a crucial direction for medical research but also serves as a foundation for advancing precision medicine.

The development of precision medicine has been heavily influenced by the exponential increase in the amount of data that represents information from multiple domains. Identifying the causes of diseases in precision medicine is enabled by several data collection and analytics technologies [1], its success relies on the interpretation of complex, multi-dimensional datasets, where statistical analysis plays a key role [4]. Advanced statistical techniques enable the extraction of meaningful insights from large-scale genomics, proteomic, and metabolomic data, which are the cornerstones of the exploration [5].

Traditional statistical methods in medicine focused primarily on hypothesis testing and linear models [6]. Common methods include t-test [7], ANOVA (analysis of variance) [8], linear regression [9], logistic regression [10], etc., which can check the differences between data, identify patterns in the data, realize disease prediction, and analyze disease risk factors [11–13]. For example, logistic regression can find correlations between a patient's insulin levels, diet, body mass index, and the occurrence of diabetes [14].

In precision medicine, the complexity and quantity of data are on the rise. As a result, there is an urgent need for more efficient inference methods, which prompts the use of machine learning (ML) algorithms [15, 16]. ML methods can efficiently identify complex patterns in data with higher flexibility and scalability. Popular and most commonly used ML methods for medical data analysis include support vector machines [17], random forests [18], k-nearest neighbors [19]

• Wuyan Zhao, Peng Li and Xu Yang are with School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China. E-mail:zhaowuyan@bit.edu.cn;lipeng360@bit.edu.cn;pyro_yangxu@bit.edu.cn.
• Canqing Yu is with the Department of Epidemiology and Biostatistics, School of Public Health, Peking University Health Science Center, Beijing, 100191, China and the Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education, Beijing, 100191, China. E-mail: yucanqing@pku.edu.cn.
• Wei Wei is with Beijing Friendship Hospital, Capital Medical University, Beijing 100050, China. E-mail: weiweibfh@ccmu.edu.cn.
* To whom correspondence should be addressed.
Manuscript received: 2025-03-03; revised: 2025-10-17; accepted: 2026-01-28

and deep neural networks [20], etc. ML methods have already been reported on correlatively predicting mortality of acute coronary syndrome [21], complications of bariatric surgery [22], related risk of metabolic syndrome [23] and so on.

However, statistical methods and ML methods are based on correlation and cannot intuitively reflect the causes of disease. In addition, understanding causal effects and ranking the importance of risk factors are essential for personalized and precise treatments. Under this background, causality provides strong support by learning a model that works under different distributions and contains causal mechanisms, and making interventions or counterfactual inferences [24, 25]. Going beyond simple correlation analysis, causality elevates learning models towards more powerful causal interpretations [26]. It directly fits the observed samples from the perspective of the data generation mechanism, and does not need complex data labels. Causality can more comprehensively identify the characteristics of the disease, mitigate confounding bias, so as to discover the most essential causes of the disease.

In the context of precision medicine, we can obtain a large amount of observable medical data from numerous patients, based on which we can discover the causal relationships among related factors. For each patient, we record the observed data of their disease symptoms and outcomes, forming a matrix-shaped dataset. Each row represents the observation results of a patient sample, and each column corresponds to an indicator. Starting from this data matrix, causal learning calculates the marginal likelihood probabilities between variables. Under certain graph-based constraints, it discovers the causal graph structure that best matches the generation mechanism of this data matrix, intuitively reflecting the causal relationships between variables.

While to the best of our knowledge, research using the whole process of causality to explore the causes of diseases has not been proposed. Most existing studies employ statistical and ML methods to identify correlations in the data, often treating these correlations as causal factors of disease and cannot accurately quantifying the average treatment effects (ATE) between causal relationships. In contrast, conventional causal discovery methods primarily focus on recovering the underlying graph structure from observational data, without providing a complete analytical pipeline. In our proposed framework, causal discovery and causal inference are explicitly integrated within a unified process that enables both structure identification and effect quantification. This end-to-end workflow allows us to generate clinically interpretable insights that go beyond correlation-based analyses or purely structural discovery, thereby bridging methodological development with clinical applicability to support timely disease risk prevention.

Therefore, we propose a framework that incorporates causality, aiming to find out the risk factors causally related to the disease for preventing the risk of disease in time. In this work, we firstly propose a causal disentanglement strategy for observed medical data to separate causal and non-causal factors in disease risk factors. Then, we design a converging causality growth algorithm to hierarchically describe causality in the form of a causal graph. Finally, a causal intervention modeling method is proposed to quantify the causal effect between causal relationships.

Overall, we innovatively propose a causality driven analysis framework (CDAF), as illustrated in Fig. 1. CDAF can discover and quantify the causes of diseases from a causal perspective, which provides an efficient and low-cost scheme for the application of precision medicine. The key point of our work is to form a complete causal learning process framework, which directly discovers causal relationships from observable medical data, eliminates non-causal confounding factors and quantifies the importance of these factors. Specifically, our main contributions are as follows:

- We propose a novel strategy for causal disentangling of medical data, which can effectively separate causal and confounding factors in disease influencing factors and improve the accuracy of disease causes' mining.
- We design a converging causality growth algorithm to verify whether there are causal relationships between disease risk factors and the disease outcomes, then describe the relationships of risk factors hierarchically in the form of causal graph, so as to find the cause most related to the disease outcome.
- We propose a causal intervention modeling method to fit causal effect values according to three rules, so as to quantify the influence of causal relationships of risk factors to disease outcomes.

2 Related Work

In the field of medical research, both traditional medical statistics and ML methods based on correlation analysis have been extensively applied, while causal learning has emerged as a crucial approach. We review their progress, applications, and limitations in this section.

2.1 Medical Statistics and ML Based on Correlation

Correlation analysis, represented by statistical models and ML methods, has made significant progress in accurately

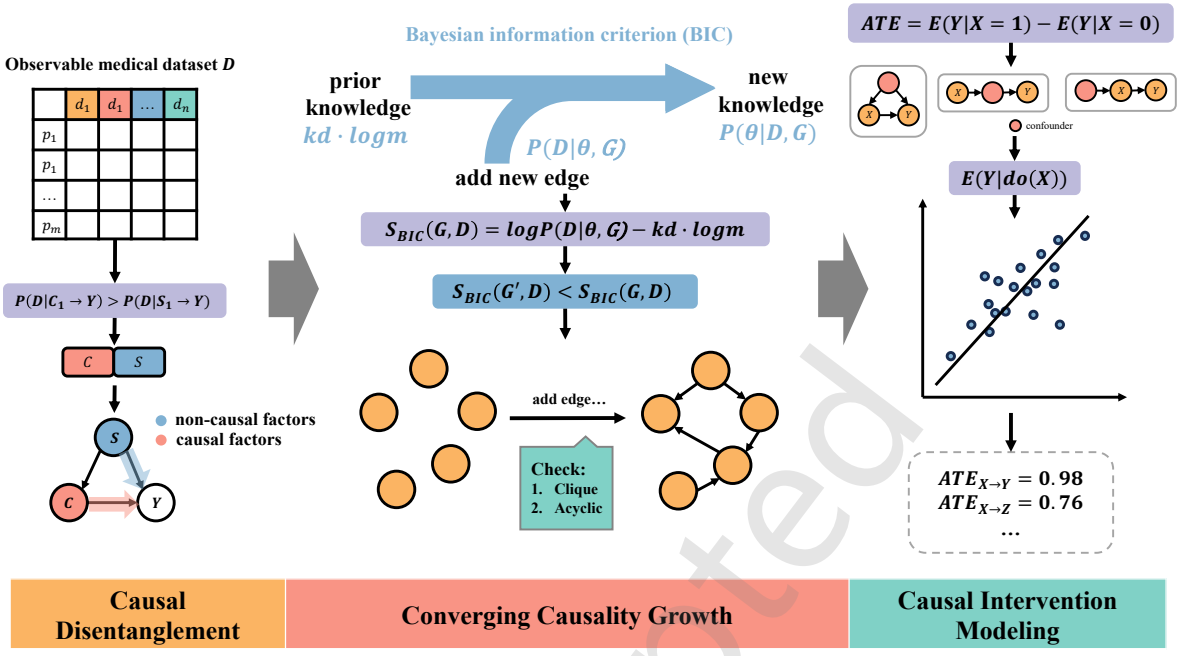


Fig. 1 Total structure of the CDAF framework. The framework consists of three components: a causal disentanglement strategy, a converging causality growth algorithm, and a causal intervention modeling method.

analyzing the causes of diseases and effectively identifying risk factors.

Linear regression models the linear relationship between the dependent variable and one or more independent variables by minimizing the sum of squared errors to estimate the regression coefficients [27]. Logistic regression is a generalized linear model that is often used in classification problems. It estimates the probability that the dependent variable is a binary or multivariate classification. In medical research, logistic regression has a wide range of applications [5, 28]. [29] demonstrates the application of logistic regression in estimating the related factors of cancer incidence. Besides, many statistical methods have been used in discovering risk factors of diseases, such as the identification of predictive biomarkers [30] and the development of genetic risk factors for cardiovascular diseases [31].

Generally, ML can be classified as supervised, unsupervised and reinforced [32, 33]. Supervised learning aims to find a pattern mapping the input data to the output data with labeled data, while unsupervised learning discovers unique structures in the input data without labels [34]. In the medical field, SVM [17] and random forest [18] belonging to supervised learning and unsupervised learning methods such as principal component analysis (PCA), k-means [35] are frequently used methods [36]. In practice, above methods have been used in many health-related applications [37, 38], such as cancers [39], cardiovascular disease [40], etc.

However, both statistical and ML methods only find correlations between variables, which measures how the variables change together. In simple terms, if two variables are strongly correlated, it means that they may exhibit similar trends in certain situations, indicating the strength of the relationship between them. This trend does not imply that a change in one variable directly leads to a change in the other, so their causal relationship cannot be determined directly. But in our CDAF, we go beyond superficial correlation analysis and use causality methods to explore the relationship between potential risk factors and disease outcomes.

2.2 Causal Learning

Causal learning focuses on identifying and inferring the causal relationships between events, has become increasingly important in various disciplines [41, 42]. Currently, causal learning involves two important issues: causal discovery and causal inference [43]. The former starts with a set of observable data and learns a causal graph to represent the causal relationships between all variables. The latter focuses on testing whether two variables are related and quantifying the influence of one variable on another.

Causal learning has become increasingly important as it allows for the extraction of explanatory structures from observable data [44]. It is capable of encoding invariant mechanisms for handling various types of causal relationships and data distributions by means of modeling conditional

independences and interventions [45, 46].

For example, [47] applied Kendall rank correlation to Bayesian structure learning, effectively mining the causal relationship among lung cancer data. [48] discovered the causal graph based on the structured table, and combined the existing tabular learning method with causal graphs to preserve the correlation between features. Furthermore, causal learning has been used to model the relationship between diseases and imaging features based on data such as X-ray [49]. Beyond these, [50] leveraged causal Bayesian networks on large-scale breast cancer cohorts to refine predictive modeling and risk assessment. [51] introduced Missdag, a causal discovery method that explicitly addresses incomplete medical records, enhancing robustness under missing data. Similarly, [52] advanced causal inference methods for epidemiological datasets, improving the reliability of risk factor identification in complex biomedical contexts. And CurvatureNet [53] leveraged causal graph refinement for analyzing complex neurological disorders.

Nevertheless, these works often primarily focused on structure recovery or disease prediction/classification tasks. On the one hand, the extracted causal relationships tend to stay at a surface level and are rarely applied to deeply explore potential risk factors, which limits research into the biological mechanisms of diseases. On the other hand, they often lack a systematic follow-up process to quantify causal effects and prioritize risk factors. In contrast, our CDAF directly mines observable medical data to identify causal risk factors for disease outcomes, eliminates spurious associations, and quantifies causal effects through ATE-based ranking. This enables the separation of confounding factors and the construction of clinically interpretable causal graphs, thereby improving both the reliability and the depth of etiology exploration.

3 Method

3.1 Causal disentanglement of observable medical data

From the causal perspective, we formulate the disentangled representation learning of observable medical data with a structural causal model (SCM) [41] as shown in Fig. 2. Let C and S denote the invariant and variant factors formed by causal and spurious patterns, representing causal factors and non-causal factors related to the disease, respectively. Y denotes the disease outcome. Intuitively, the causal relationships can be expressed as $C \rightarrow Y \leftarrow S$ and $S \rightarrow C$. The former means that the disease outcome is affected by both causal and non-causal factors simultaneously, and the latter means that causal factors also affect non-causal factors.

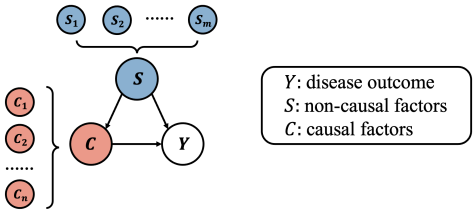


Fig. 2 The illustration of the structural causal model (SCM). The arrow $\rightarrow$ indicates causal relationship from one variable to another, i.e., cause $\rightarrow$ effect. The path C $\rightarrow$ Y means direct causal relationship, the path S $\rightarrow$ Y and S $\rightarrow$ C lead to spurious relationship between C and Y.

Common causes between variables, termed confounders, may induce backdoor paths [41] (e.g. $C \leftarrow S \rightarrow Y$) that lead to spurious correlations. Models relying on such spurious relations are unstable under perturbations of non-causal factors. Therefore, disentangling causal from spurious components provides a reliable basis for identifying disease causes.

In the background of Bayesian statistics, we proceed our causal disentanglement strategy from the perspective of describing the generating mechanisms of medical data. We assume that medical datasets contain a large number of indicators related to diseases, forming an i.i.d. set of variables $\mathcal{D} = \{d_1, d_2..., d_n\}$. Specifically, for each candidate variable, we define causal model $\mathcal{M}_C : Y \sim (C, \theta_C, U_Y)($ indicating $C \rightarrow Y)$ and non-causal model $\mathcal{M}_S : Y \sim (S, \theta_S, U_Y)$(indicating $S \nrightarrow Y$).

Theorem 1 *Given an observed dataset $\mathcal{D}$, the marginal likelihood under the causal hypothesis is strictly greater than that under the non-causal hypothesis, i.e.,*

$$Score(D \mid \mathcal{M}_C) > Score(D \mid \mathcal{M}_S) \tag{1}$$

where $Score(D \mid \mathcal{M}) = \int P(D \mid \theta, \mathcal{M})\, P(\theta \mid \mathcal{M})\, \mathrm{d}\theta$, θ *denotes model parameters.*

The proof is provided in Appendix A.1. We use marginal likelihood as the discriminative criterion since it balances data fit and model complexity, providing a principled basis for distinguishing causal from non-causal hypotheses. Consequently, the theorem establishes that causal factors are statistically more consistent with the data-generating process than spurious factors, thus supporting the disentanglement of causal and non-causal components in medical datasets.

For optimization, we use do-calculus $do(S)$ to intervene spurious patterns and therefore cut the links from causal patterns to spurious patterns [41]. Therefore, for a set of variables (C, S) that are expected to be disentangled, we calculate the $Score(Y|do(C))$ and $Score(Y|do(S))$ respectively:

$$Score(P|do(C)) = \sum_{s} P(Y|C, S = s)P(S = s) \quad (2)$$

$$Score(P|do(S)) = \sum_{c} P(Y|S, C = c)P(C = c) \quad (3)$$

where $Score()$ is determined according to Theorem (1).

Theorem 2 *If* $Score(Y|do(S)) \rightarrow 0$ *while* $Score(Y|do(C))$ *remains unchanged, then* S *is identified as a spurious factor and* C *as a causal factor.*

The proof is provided in Appendix A.2. In summary, this disentanglement strategy establishes a solid theoretical foundation from a mathematical perspective. We will detail the specific scoring function and algorithmic implementation used to operationalize this strategy in the next section.

3.2 Converging Causality Growth Algorithm

By disentangling causal and non-causal factors, we identify the components of risk factors that causally contribute to a disease outcome and visualize the relationships between these risk factors. Therefore, we propose a converging causality growth algorithm, which is a score-driven causal edge addition method, as shown in Fig. 3. Through the fitting of data and causal relationships, the causal edges are iteratively discovered, then a hierarchical causal graph is generated to describe the causal relationships between risk factors and disease outcomes. Specifically, we introduce the Bayesian information criterion (BIC) based causality iterative fitting form, and give causal edge generation checking conditions.

3.2.1 *BIC-based Iterative Fit Form*

We believe that Meek's conjecture holds [54, 55], then we propose the converging causality growth algorithm based on it. Through our method, a finite sequence of edge additions will be generated. Applying this sequence to an initial graph G produces a new graph H, which still preserves the independence relations of G throughout the process. Finally, the procedure converges when $G = H$. Specifically, we iteratively validate causal relationships between risk factors and disease outcomes by using the BIC to judge the fit of the data and causality. We select BIC because it offers a computationally efficient and theoretically consistent criterion that balances model fit, complexity, and interpretability better than alternative scoring functions.

BIC evaluates the degree of fit between the directed acyclic graph (DAG) and the observable medical data, and tests the causality in terms of independence. It uses the marginal likelihood to measure the degree of fit, and uses another prior

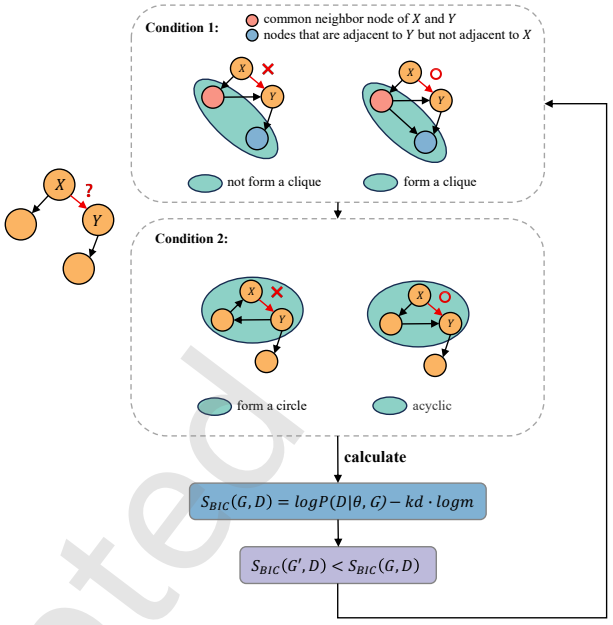


Fig. 3 The converging causality growth algorithm. X and Y are two risk factor variables, we suppose the edge to be inserted is X → Y.

knowledge about the graph to ensure the current structure. Accordingly, given the observable medical dataset D and the number of data m, we iteratively generate a causal graph G of dimension d with the following fit score:

$$S_{BIC}(G, D) = logP(D|\theta, G) - kd \cdot logm \quad (4)$$

where θ is the parameter of the maximum likelihood function, k is a hyperparameter, d is the dimension of G.

Starting from an initial empty graph G, we iteratively add edges with causal relationships to G, and calculate the BIC fitting score. After a finite number of adding edges, the final causal graph will converge to an invariant state. And between the new graph G' obtained at each iteration and the graph G of the previous iteration, we have the following relation:

$$S_{BIC}(G', D) < S_{BIC}(G, D) \quad (5)$$

Based on Eq. 5, our target is to find the graph structure G with the minimum BIC score, to achieve convergence.

3.2.2 *Add Edge Condition Check*

When iteratively adding edges to the graph, we propose two check conditions. They ensure causality and acyclicity, respectively, maintaining the properties of the causal graph. We traverse all pairs of nodes (X, Y) and consider the edge to be inserted as $X \rightarrow Y$. Define T as the set of nodes that are adjacent to Y but not adjacent to X, and S as the common neighbor nodes of X and Y.

Condition 1: Ensure that the validity of the causal graph structure is maintained after new edges are inserted. We

Algorithm 1 Check condition 1: clique

Data: Current DAG G, candidate node set $U = common\_neighbors(X, Y) \cup (neighbors(Y) \setminus neighbors(X))$
Result: Boolean value indicating whether U forms a clique in G
Initialize flag *isClique* $\leftarrow$ True
for each node $u \in U$ **do**
  for each node $v \in U$ with $v \neq u$ **do**
    if edge (u, v) does not exist in G **then**
      isClique $\leftarrow$ False **break** inner loop
    end if
  end for
  if *isClique* = False **then**
    break outer loop
  end if
end for
return isClique

Algorithm 2 Check condition 2: acyclicity

Data: Current DAG G, candidate edge $X \rightarrow Y$
Result: Boolean value indicating whether adding $X \rightarrow Y$ preserves acyclicity
Initialize an empty queue Q and visited set V
Enqueue Y into Q, add Y to V
while Q is not empty **do**
  $current \leftarrow Dequeue(Q)$
  if $current = X$ **then**
    return False
  end if
  for each child c of $current$ in G **do**
    if $c \notin V$ **then**
      Enqueue c into Q;
      add c to V
    end if
  end for
end while
return True

first introduce the V-structure, which is a specific causal structure, denoted as: $A \rightarrow C \leftarrow B$, where A and B are independent and act on C together. Using this structure, we can determine the direction of edges in causal graph, so as to identify the causal relationships intuitively and quickly. When we add an edge $X \rightarrow Y$ to the graph, we may add or destroy the V-structure, leading to a change in causality. For example, a new V-structure: $X \rightarrow Y \leftarrow Z$ is added due to the new edge $X \rightarrow Y$, where Z belongs to $S \cup T$.

In order to ensure this change will not destroy the original causal relationships, we introduce **clique** to realize the check condition. The detailed checking steps of this condition is shown in Algorithm 1.

If $S \cup T$ forms a clique, then according to its characters, all nodes in it share the same causal source , that is, the same set of parent nodes. In this way, any added or lost causal relationships are naturally constrained by the closure property of cliques, which ensures the rationality of causal relationships.

Condition 2: Ensure that causal graph remains acyclic by confirming no cycles are created after new edges are inserted. Causality is unidirectional, a factor cannot be both a cause and an outcome, which is consistent with the properties of DAG. Therefore, we need to check whether a circle is formed after inserting edges to ensure the rationality of the causal relationships. Based on the depth-first search algorithm, we start from node X and check whether the path from Y to X contains nodes in $S \cup T$. If it is not included, it ensures the acyclic property will not be destroyed after inserting edge $X \rightarrow Y$. The detailed checking steps of this condition is shown in Algorithm 2

Therefore, as long as the edge $X \rightarrow Y$ to be inserted passes the above two conditions, BIC is used to calculate the change of score to realize the causality generation. The final causal graph contains the disease outcome, risk factors and their causal relationships. We define this causal graph as three levels: direct, indirect, and other.

Direct factors refer to factors that are directly linked to the disease outcome, with edges pointing toward the outcome. Indirect factors are directly linked to direct factors, and the direction of the arrow points to direct factors. Other factors refer to the ones left unmentioned. In this hierarchical relationship, direct and indirect causal factors can be extracted from the graph clearly, hence guiding the precise implementation of medical interventions.

The pseudocode of converging causality growth algorithm is shown in Algorithm 3.

3.3 Causal Intervention Modeling Method

In order to quantify the causal effect between causal relationships and determine the importance of disease risk factors, we propose an intervention modeling method. We replace the traditional controlled experimental intervention method and only intervene from the observed data and causal graph to realize the estimation of causal effect value. Specifically, we further discuss the method in two parts: causal effect modeling and regression-based medical data estimation.

Algorithm 3 Converging Causality Growth Algorithm

Data: Data D
Result: Final Causal Graph G
Initialize BIC score class and graph G
Compute initial score of graph G
while True **do**
  Set $score_{new} = score$
  for each pair of nodes (X, Y) where $X \neq Y$ **do**
    calculate T: nodes adjacent to Y but not adjacent to X,
    calculate S: the common neighbor nodes of X and Y.
    if no edge exists between X and Y **then**
      for each subset $S \cup T$ **do**
        Test condition1 & condition2
        if both conditions are satisfied **then**
          Compute score $change$ by inserting $X \rightarrow Y$
          if score improves **then**
            Update graph G with $X \rightarrow Y$
            Set $score_{new} = score + change$
          end if
        end if
      end for
    end if
  end for
  if $score_{new} = score$ **then**
    break
  end if
end while

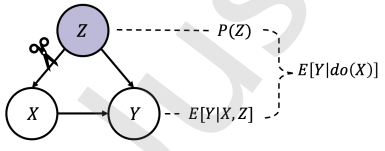


Fig. 4 Shared confounding correction rule. X is intervention variable, Y is outcome variable, Z is shared confounding variable.

3.3.1 *Causal effect modeling*

Causal effect estimates the causal relationship between a given treatment or intervention and the target outcome variable. In the potential outcomes framework [41], units in the intervention and control groups will show different potential outcomes, and the causal intervention effect is the difference between these two outcomes. However, it is costly to carry out clinical randomized controlled trials to identify the outcome differences between the intervention and control groups. Therefore, we used data simulation to construct the causal effect expression. Based on the observable medical data and the causal relationships found in causal graph, we calculate the average treatment effect as causal effect:

$$ATE = E(Y|X = 1) - E(Y|X = 0) \tag{6}$$

where it is assumed that all subjects receive the intervention representation as $X = 1$ and all subjects receive the control representation as $X = 0$.

Nevertheless, due to the complex connectivity of causal relationships contained in the causal graph, a variable may be affected by multiple variables at the same time, and the causal chain is not a simple linear influence relationship. Therefore, it is necessary to identify the relationship according to special rules, and to model the specific conditional expectation expression based on the complex interaction between variables. We further discuss three forms of conditional expectation modeling:

Shared confounding correction rule: In the medical field, intervention variable X (such as medication or treatment) and outcome variable Y (such as disease outcome or recovery) are often influenced by common cause Z (such as age, sex, past medical history, etc.). According to the backdoor criterion [56], we suppose that the intervention variable X and the outcome variable Y are affected by the common cause Z, as shown in Fig. 4. The variable Z is observable, then the causal effect value of $X \rightarrow Y$ can be expressed by blocking the path $X \leftarrow Y \rightarrow Z$, and the conditional expectation expression is as follows:

$$E\left[Y|do\left(X\right)\right] = \sum_{Z} P\left(Z\right) E\left[Y|X, Z\right] \tag{7}$$

Therefore, we can quickly identify which variable (Z) in the model is used to estimate the causal relationship between X and Y, and thus estimate the conditional expectation of $X \rightarrow Y$ conveniently.

Mediated confounding control rule: If Z is unobservable, the causal effect value of $X \rightarrow Y$ cannot be estimated by blocking the backdoor path. In this case, we look for a front door path of $X \rightarrow Z \rightarrow Y$ [41], as shown in Fig. 5. The front door criterion requires that (1) Z cuts off all directed paths from X to Y, (2) there is no backdoor path from X to Z, (3) all backdoor paths from Z to Y are blocked by X. We then estimate the conditional expectation as follows:

$$E\left[Y|do\left(X\right)\right] = \sum_{Z} P\left(Z|X\right) P\left(Y|Z\right) \tag{8}$$

Therefore, the conditional expectation of $X \rightarrow Y$ can also be obtained after Z is identified as the mediating variable of $X \rightarrow Y$ in the causal graph.

External influence variable adjustment rule: If a valid confounding variable cannot be extracted from the causal

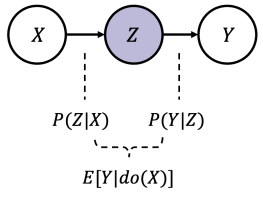


Fig. 5 Mediated confounding control rule. X is intervention variable, Y is outcome variable, Z is mediated confounding variable.

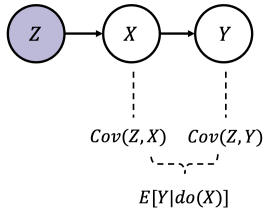


Fig. 6 External influence variable adjustment rule. X is intervention variable, Y is outcome variable, Z is external influence variable.

graph based on the above rules, we then use an external adjustment to estimate the conditional expectation expression [57]. We set a dummy variable Z without arrows, which is required to satisfy two properties: (1) it is independent of the outcome variable Y, (2) it is directly related to the intervention variable X, as shown in Fig. 6. With the aid of Z , we calculate covariance to get the following conditional expectation expression:

$$E\left[Y|do\left(X\right)\right] = \frac{Cov\left(Z,Y\right)}{Cov\left(Z,X\right)} \tag{9}$$

Consequently, after identifying the external influence variable Z from the causal graph, the conditional expectation of $X \rightarrow Y$ can be calculated.

3.3.2 Regression-based medical data estimation

Due to the limited availability of observable medical data and sample size, and considering that the relationships between disease and risk factors can be approximately linear, we use a linear regression-based method for estimation. This approach allows effective parameter estimation without requiring a large number of samples and avoids the high data demands of complex models.

We assume that the outcome variable Y is linearly related to the intervention variable X and the confounding variable Z. Their relationship is denoted as:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon \tag{10}$$

where β_1 is the regression coefficient for the intervention variable X, representing the linear effect of X on Y after controlling for the confounding variable Z. ϵ is the error term, assuming $E[\epsilon] = 0$.

After determining the causal effect model $E[Y|do(X)]$, the form for controlling the confounding variable Z is established and passed into Eq. 10. The linear regression model can calculate the conditional expectation, so as to fit the causal effect of X on Y, which is an intuitive linear approximation. Specifically, $E[Y|X = 0] = \beta_0$, $E[Y|X = 1] = \beta_0 + \beta_1$, and according to Eq. 6, β_1 is the ATE we need to estimate.

4 Experiments

4.1 Data and Preprocess

For real-world clinical datasets, we selected the China Kadoorie Biobank (CKB) dataset and the Parkinson's disease (PD) dataset for the experiment. The CKB dataset contains detailed health information from approximately 510,000 adults aged 30 to 79 across 10 regions in China, collected between 2004 and 2008 [58, 59]. It includes key indicators that may determine common chronic diseases, such as gender, age, diet, smoking, drinking, exercise, medical history, medication use, weight, blood pressure, and blood sugar levels. The PD dataset was sourced from Beijing Friendship Hospital, Capital Medical University. It includes data from 25 Parkinson's disease patients and 43 healthy controls. Multimodal magnetic resonance imaging data were collected for these samples, reflecting brain iron deposition in Parkinson's patients. Both datasets were ethically approved and obtained with participant consent.

We conducted the following steps for data preprocessing.
Feature Selection: For the CKB dataset, feature selection was guided by a targeted review of epidemiological literature and clinical guidelines concerning three representative diseases—stomach cancer, liver cancer, and acute myocardial infarction (AMI)—and further refined through expert consultation based on a predefined set of criteria to ensure objectivity and clinical relevance. For the PD dataset, we reconstructed the collected quantitative susceptibility mapping images and extracted the susceptibility values of 12 regions of interest (ROIs), including the substantia nigra (SN), red nucleus (RN), caudate nucleus (CN), putamen (PUT), globus pallidus (GP) and thalamus (TH) in the left and right brain regions. In addition, we selected scale information reflecting the participants' mental state, anxiety, depression, etc.
Data Cleaning: For the CKB dataset, to mitigate outcome imbalance and enhance the stability of likelihood-based causal discovery, we adjusted the sample to maintain an approximately 1:1 ratio between diseased and non-diseased records. For the PD dataset, due to unavoidable errors in the image

acquisition and reconstruction stages, we directly eliminated samples with a significant amount of missing values and filled data with fewer missing values with the mean.

Data Normalization: For the CKB and PD datasets, data normalization was performed to ensure the stability of the algorithm model during the subsequent phases. All feature values were scaled to the range $[0, 1]$ to eliminate dimensional effects.

4.2 Experimental Results

To verify the effectiveness of our method, we applied CDAF to the CKB dataset and PD dataset, focusing on stomach cancer, liver cancer, acute myocardial infarction (AMI) and temporal pattern of brain iron deposition in Parkinson's patients. We compared the discovered causal relationships with findings from existing medical studies.

4.2.1 The CKB Dataset

We searched the literature to find known high-risk factors and pathogenic mechanisms of the diseases, and compared them with the nodes and causal relationships in the causal graph results. Unlike benchmark datasets with established ground-truth structures, the CKB clinical dataset lacks a universally recognized causal graph, making direct computation of quantitative metrics infeasible. Therefore, we focused on examining whether the causal effect between direct risk factors and disease development is consistent with clinical data and studies, then to further verify the rationality of the results.

Stomach Cancer. Fig. 7 illustrates the causal relationships among risk factors for stomach cancer. Direct factors include frequency of smoking and spicy food consumption, with ATE of 0.09 and -0.05, respectively. When compared to medical research findings, these two causal relationships corroborate that behavioral factors, such as smoking and alcohol consumption, contribute to the onset of stomach cancer [60–63]. Meanwhile, stomach cancer is closely related to modifiable factors such as diet. Studies indicate that higher intake of fruits and vegetables may reduce stomach cancer risk [64–67]. This conclusion is reflected in our causal graph result, where stomach cancer is linked to the intake frequency of fresh vegetables and fruits as other factors. Additionally, the ATE from stomach cancer to the frequency of meat consumption is 0.22, further emphasizing the potential impact of dietary habits on stomach cancer risk.

Therefore, it is believed that smoking and the consumption of spicy food are direct causal risk factors contributing to stomach cancer. Additionally, the intake of fresh vegetables and fruits, along with other dietary habits, may serve as potential risk factors which need further investigation.

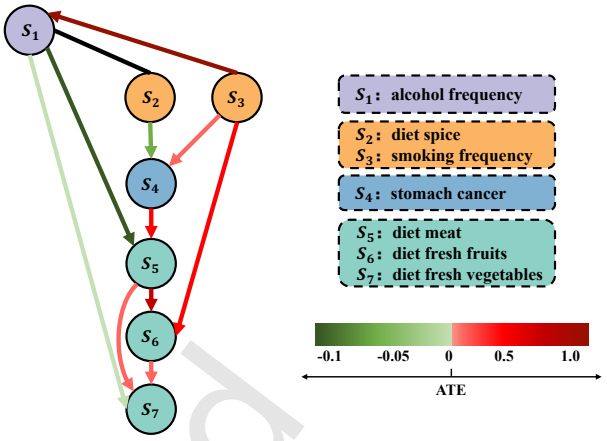


Fig. 7 Causal graph of stomach cancer. The edges represent causal relationships, with colors indicating the ATE value: red edges denote positive effects, and green edges denote negative effects. The intensity of the color reflects the magnitude of the effect, with darker shades indicating stronger causal effects. This graph highlights the degree of influence of risk factors contributing to stomach cancer.

Liver Cancer. Fig. 8 illustrates the causal relationships of risk factors for liver cancer. The results indicate three direct factors: BMI, drinking frequency, and recent spicy food consumption frequency. According to the relevant medical research on liver cancer, alcoholism, alcohol-related cirrhosis, metabolic syndrome, obesity and various dietary exposures are all at high risk for the disease [68–73]. Studies further indicate that individuals with type-2 diabetes or BMI above 30 kg/m^2 face a higher risk of liver cancer [74–76]. This aligns with the indirect factors shown in our result, where diabetes influences BMI, suggesting that diabetes may indirectly increase liver cancer risk through obesity. Moreover, among the three identified direct factors, the ATE was 0.01 for BMI, 0.04 for drinking frequency, and 0.05 for the frequency of spicy food consumption. In contrast, the ATE for indirect factors diabetes and weight loss, which indirectly affect liver cancer risk through BMI, was significantly higher, at 2.09 and 2.25, respectively. These findings emphasize the importance of focusing on diabetes and weight management as critical risk factors influencing liver cancer, reminding researchers to pay more attention to them.

Acute Myocardial Infarction (AMI). Fig. 9 illustrates the causal relationships of risk factors for AMI. It was found that a history of coronary heart disease and hypertension, as well as the symptom of oxygen deficiency when walking on flat ground, are directly related to AMI. These factors are regarded as direct causal factors contributing to AMI. The ATE for hypertension and coronary heart disease is 0.25 and 0.26 respectively, which means that there is approximately a 25% probability of causing AMI and it's a relatively high risk.

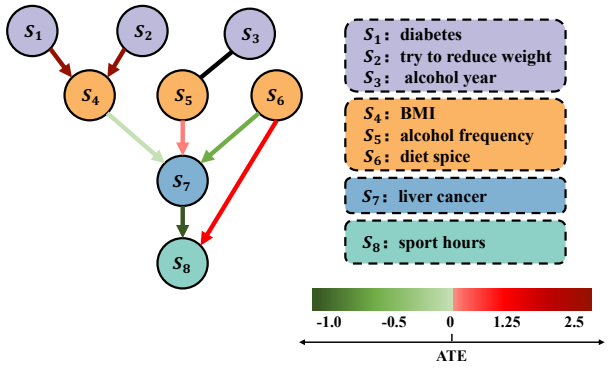


Fig. 8 Causal graph of liver cancer. This graph highlights the degree of influence of risk factors contributing to liver cancer.

Additionally, AMI is related to smoking frequency, exercise duration, and the regular consumption of spicy food, so these factors are regarded as other factors. The ATE for smoking frequency and exercise duration is approximately 0.65 and 0.56, respectively, indicating about a 50% probability of mutual influence. Above findings align with existing medical studies [77–84], so organizing these factors will aid in treating the AMI.

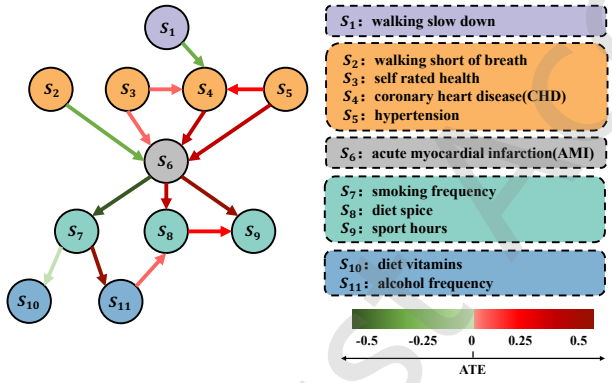


Fig. 9 Causal graph of acute myocardial infarction (AMI). This graph highlights the degree of influence of risk factors contributing to AMI.

4.2.2 The PD dataset

Fig. 10 illustrates that brain iron deposition in Parkinson's patients can be categorized into three distinct stages. The initial stage shows increased iron deposition in the thalamus and putamen, occurring simultaneously. The second stage involves the caudate nucleus and red nucleus. The final stage, characterized by increased iron deposition in the substantia nigra and globus pallidus, occurs at a deeper level. Meanwhile, it is found that brain iron deposition in the left and right brain regions basically occurs in the same period or adjacent periods. Additionally, causal relationships also exist between factors such as anxiety, depression, and cognitive levels, represented by the red nodes in the Fig. 10. These psychological symptoms are often linked to iron deposition in specific brain areas.

Medical researches indicate that brain iron deposition in Parkinson's patients initially occurs in the thalamus [85]. In the early stages of PD, patients show morphological changes in the right putamen [86] and early dopaminergic neuron dysfunction in the caudate nucleus [87]. Increased iron levels in the substantia nigra serve as a biomarker for disease status in advanced PD [88]. All of the research findings above accurately and effectively support our causal graph obtained. Analyzing it aids in evaluating the risk factors of PD and offers a new direction for exploring the potential mechanisms of brain disorders in Parkinson's patients.

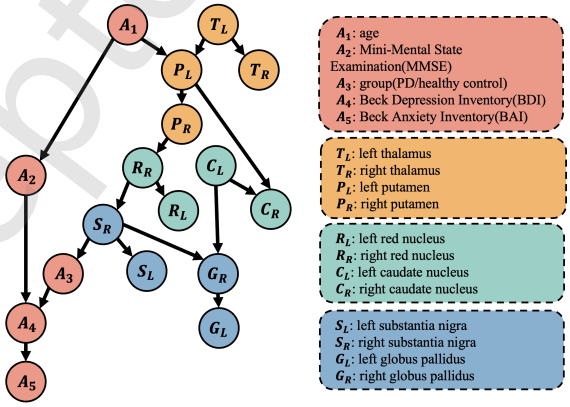


Fig. 10 Causal graph of Parkinson's disease (PD). The edges represent causal relationships over time, illustrating the dynamic interactions between different brain regions involved in iron deposition. This graph provides insights into the temporal patterns of iron deposition and their potential role in the pathogenesis of PD.

4.3 Effectiveness and Robustness

To confirm the effectiveness and robustness of our CDAF, we experimentally verify the converging causality growth algorithm and causal intervention modeling, respectively.

4.3.1 Effectiveness of converging causality growth algorithm

On CKB dataset and PD dataset, we use the proposed converging causality growth algorithm to discover hierarchical causal graph. In the converging process, the most reliable edge was inserted into the causal graph based on the smallest BIC score in each iteration. The trend of the scores for the stomach cancer, liver cancer, AMI and brain iron deposition in Parkinson's patients are presented in Fig. 11.

Fig. 11 shows that the score exhibits a clear declining trend. All four curves in the later iterations smoothly converge to a lower BIC score within 20 iterations. This indicates that the

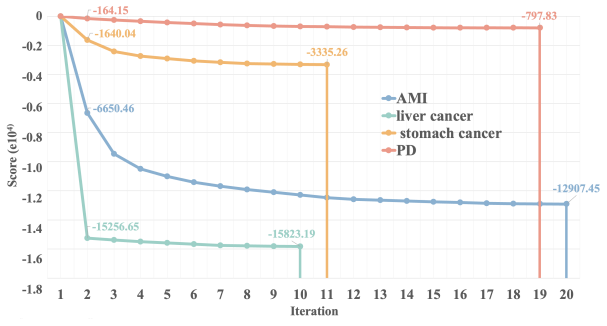


Fig. 11 BIC score iterations of four diseases. Figure illustrates the iterative optimization process of the converging causality growth algorithm, which proves the stable convergence for four diseases. Each curve represents the BIC score trend for a particular disease, indicated by a different color. The consistent decrease in BIC scores over multiple iterations confirms the effectiveness of the algorithm in discovering causal relationships.

proposed algorithm has significant convergence properties and efficient data processing capability, ensuring a relatively balanced state is reached in the limited iteration steps. It also reflects the effectiveness of the internal mechanism of the algorithm, which overcomes the complexity and uncertainty caused by the different distribution and scale of different observable medical data.

Meanwhile, in the initial iterations, it can be seen from Fig. 11 that the score decreases rapidly. The score of the curve representing liver cancer decreases by more than 5 orders of magnitude after the first iteration of edge insertion. This indicates that in the early stage of convergence, the edges inserted to the causal graph have more important significance and are causal relationships with a higher degree of fitting to the data. The stomach cancer and PD curves in Fig. 11 show smoother declines, indicating that the added causal relationships contribute more uniformly across iterations.

Specifically, the scores initially start near 0. For stomach cancer, the score stabilizes at approximately -3335; for liver cancer, it converges to around -24551; and for AMI, it reaches -12907. In the analysis of brain iron deposition in Parkinson's disease patients, the score converges to -797 after 19 iterations. Notably, the score for liver cancer converges to the lowest value (approximately -24551), indicating that the learned causal graph structure for liver cancer exhibits the highest model fit, suggesting the causal relationships it discovers are more accurate.

Therefore, the converging causality growth algorithm could effectively discover convergent causal graphs across various types of datasets. Each iteration successfully inserts edges that represent causal relationships to the causal graph.

4.3.2 Robustness of importance ranking

We have ranked the degree of influence of causal relationships based on the calculated ATE, then we screen out factors with higher risks and stronger causality, focusing on in subsequent treatment and research. We use four robustness checks to refute the obtained estimates. Specifically, we implement four different treatments on the original data to re-estimate the ATE value: (1) adding noise to the dataset, (2) adding random confounding effect, (3) extracting subsets, and (4) placebo test.

Refutation result for stomach cancer is shown in Table 1. Among them, the two edges recomputed by the adding random confounding effect method obtained the highest p-value of 1. On average, the extracting subsets method achieves a mean p-value of 0.94, which verifies the effectiveness of the refutation test. Refutation results for liver cancer and AMI are shown in Table 2 and Table 3, through which we can see that the refutation test basically achieved a p-value bigger than 0.9 on average, indicating the consistency of the recalculated degree of influence before and after data treatments. The refutation test results verify the robustness of causal intervention modeling, which is crucial to the effect of our proposed causal effect calculation method. Moreover, it can be proved that our ranking is reliable and the determined high-risk factors are valuable.

4.4 Comparison Experiment

4.4.1 Comparison with baselines

To further strengthen the evaluation, we expanded the comparison to include 10 baseline causal discovery methods covering three major families: (1) neural and deep learning–based approaches (ICD [89], DAG-GNN [90], RL [91], CORL [92]), (2) continuous optimization–based approaches (NOTEARS [93], GOLEM [94], DAGMA [95], DiffAN [96]), and (3) constraint/search–based approaches (PC [97], FCI [98]). We conducted benchmarking experiments on two widely used public datasets (Sachs [99] and ASIA [100]), reporting precision, recall, F1-score, and runtime for each method. The results (as shown in Table 4) show that CDAF achieves competitive or superior accuracy compared with both classical and recent baselines, while maintaining a substantially lower runtime than many optimization- or learning-based algorithms. These results highlight that CDAF balances accuracy and efficiency across heterogeneous algorithmic paradigms, underscoring its robustness and practical suitability for biomedical causal discovery tasks.

Table 1 Refutation Result for stomach cancer. P-values for null hypothesis validity, large p-value (close to 1) indicates that the causal effect estimate is robust.

Causal edge	add noise	p value	add random confounder	p value	extracting subsets	p value	placebo test	p value
$S_3 \to S_1$	0.6638	0.92	0.6651	0.92	0.6654	0.98	0.0015	0.94
$S_5 \to S_6$	0.3587	0.98	0.3604	0.96	0.3607	0.94	0.0007	0.98
$S_4 \to S_5$	0.2280	0.86	0.2206	0.9	0.2217	0.98	0.0049	0.8
$S_3 \to S_6$	0.1178	0.94	0.1182	1	0.1183	0.92	0.0010	0.9
$S_3 \to S_4$	0.0892	0.98	0.0892	0.88	0.0887	0.86	0.0008	0.76
$S_5 \to S_7$	0.0350	0.74	0.0357	0.96	0.0354	0.96	-0.0008	0.74
$S_6 \to S_7$	0.0226	0.94	0.0225	0.96	0.0225	0.9	0.0003	0.98
$S_1 \to S_7$	-0.0208	0.94	-0.0205	1	-0.0205	0.9	0.0000	0.92
$S_2 \to S_4$	-0.0505	0.98	-0.0505	0.84	-0.0506	0.98	0.0004	0.88
$S_1 \to S_5$	-0.0924	0.9	-0.0938	0.92	-0.0939	0.98	-0.0005	0.88
Avg		**0.918**		**0.934**		**0.94**		**0.878**

Table 2 Refutation Result for liver cancer

Causal edge	add noise	p value	add random confounder	p value	extracting subsets	p value	placebo test	p value
$S_2 \to S_4$	2.2572	1	2.2541	0.98	2.2538	0.92	-0.0687	0.92
$S_1 \to S_4$	2.0819	0.94	2.0908	0.98	2.0837	0.92	-0.0062	1
$S_5 \to S_7$	0.0355	0.88	0.0352	0.98	0.0352	0.94	0.0000	0.9
$S_6 \to S_8$	0.1984	0.84	0.1943	0.9	0.1954	0.96	0.0020	0.98
$S_4 \to S_7$	-0.0118	0.84	-0.0115	0.9	-0.0116	0.96	0.0000	1
$S_6 \to S_7$	-0.0517	0.88	-0.0513	0.84	-0.0515	0.94	-0.0003	0.88
$S_7 \to S_8$	-0.7903	0.82	-0.7854	0.8	-0.7871	0.96	0.0073	0.9
Avg		**0.886**		**0.911**		**0.943**		**0.94**

Table 3 Refutation Result for AMI

Causal edge	add noise	p value	add random confounder	p value	extracting subsets	p value	placebo test	p value
$S_7 \to S_{11}$	0.6510	1	0.6504	0.8	0.6500	0.94	-0.0006	0.84
$S_6 \to S_9$	0.5656	0.96	0.5627	0.98	0.5630	0.98	0.0085	0.84
$S_6 \to S_8$	0.2743	0.98	0.2780	0.84	0.2785	0.9	-0.0026	0.88
$S_4 \to S_6$	0.2595	0.82	0.2620	0.96	0.2633	0.86	-0.0007	0.94
$S_5 \to S_6$	0.2515	0.96	0.2521	0.84	0.2519	0.96	0.0031	0.88
$S_5 \to S_4$	0.1241	0.92	0.1245	1	0.1246	0.98	0.0008	0.9
$S_8 \to S_9$	0.1298	1	0.1291	0.92	0.1281	0.92	0.0047	0.88
$S_{11} \to S_8$	0.0745	0.9	0.0750	0.84	0.0760	0.82	0.0001	1
$S_3 \to S_6$	0.0578	0.98	0.0578	0.92	0.0581	0.94	-0.0003	0.96
$S_3 \to S_4$	0.0282	0.82	0.0284	0.94	0.0287	0.72	-0.0003	0.88
$S_7 \to S_{10}$	-0.0063	0.82	-0.0064	1	-0.0064	0.96	0.0000	0.92
$S_1 \to S_4$	-0.1663	0.96	-0.1663	0.86	-0.1669	0.82	-0.0010	0.94
$S_2 \to S_6$	-0.2347	0.8	-0.2314	0.9	-0.2300	0.74	-0.0004	0.98
$S_6 \to S_7$	-0.3725	0.9	-0.3764	0.7	-0.3767	0.98	0.0012	0.89
Avg		**0.916**		**0.893**		**0.894**		**0.909**

4.4.2 *Comparative Case Study on the CKB Dataset*

We contrast our CDAF with the results of four existing causal discovery algorithms: GOLEM, RL, CORL, and DAG-GNN, aiming to demonstrate the advantages of our framework on observable medical data. When applied to the CKB dataset, these algorithms identified causal relationships among the same variables, resulting in different causal graphs. Fig. 12 illustrates the comparison results. It is confirmed that our framework has better applicability for the CKB dataset.

Among these causal graphs, few causal relationships are marked in red, as the four comparative algorithms (RL, DAG-GNN, CORL, and GOLEM) identified limited causal relationships. But these red edges largely overlap with those discovered by our framework, indirectly validating the effectiveness of our method. The blue edges, representing causal relationships jointly identified by CORL and GOLEM, also overlap with our framework and further show the partial consistency between these algorithms. In contrast, the green

Table 4 Performance comparison of causal discovery methods on Sachs and ASIA datasets

Methods	Sachs Dataset				ASIA Dataset			
	Precision	Recall	F1	Runtime	Precision	Recall	F1	Runtime
CDAF	**0.407**	**0.550**	**0.468**	**10.05 s**	**0.857**	**0.75**	**0.799**	**1.57 s**
GOLEM	0.318	0.350	0.333	2m33 s	0.154	0.250	0.191	16.6 s
RL	0.333	0.350	0.341	12m2 s	0.333	0.375	0.353	5m34 s
CORL	0.444	0.400	0.421	22m22 s	0.429	0.375	0.400	50m46 s
DAG-GNN	0.200	0.150	0.171	56m4 s	0.455	0.625	0.529	8m37 s
PC (FisherZ)	0.111	0.150	0.128	1.2 s	0.333	0.429	0.375	0.2 s
NOTEARS	0.263	0.500	0.342	4.51 s	–	–	–	–
FCI	0.300	0.200	0.240	0.8 s	1.000	0.375	0.545	0.11 s
ICD	0.231	0.450	0.305	0.26 s	0.556	0.625	0.588	0.01 s
DAGMA	0.571	0.400	0.471	17m47 s	0.500	0.500	0.500	11m58 s
DiffAN	0.0909	0.278	0.137	11m53.8s	0.185	0.625	0.286	7m59.5s

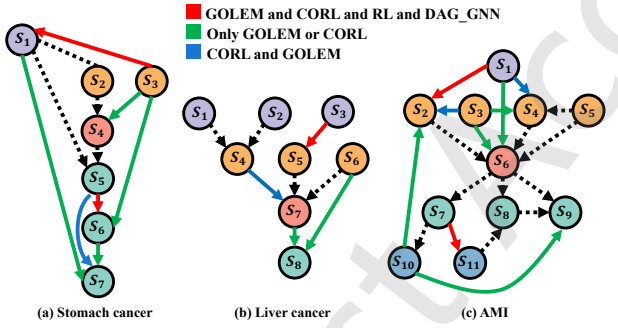


Fig. 12 Comparison of causal graphs generated by different algorithms for (a) stomach cancer, (b) liver cancer, and (c) acute myocardial infarction (AMI). Four algorithms are used: GOLEM, CORL, RL and DAG-GNN. The red edges represent causal relationships identified by four algorithms. The blue edges jointly identified by CORL and GOLEM, demonstrating partial consistency between these methods. The green edges uniquely identified by either CORL or GOLEM (but not both), highlighting additional discoveries made by these algorithms. The black dashed edges identified exclusively by our framework, which were not discovered by any of the four comparative algorithms.

edges, corresponding to causal relationships uniquely identified by either CORL or GOLEM, demonstrate additional discoveries made by these algorithms, some of which were not discovered by us. Notably, the black dashed edges represent causal relationships exclusively identified by our framework, which were not discovered by any of the comparative algorithms. This unique discovery of approximately half of the causal relationships emphasizes the superior capability of our method in identifying key causal relationships directly related to disease outcomes. While the comparative algorithms provide valuable supplementary insights and help validate the effectiveness of our framework, they fail to fully discover key risk factors. The overlapped relationships provide an indirect but credible validation signal, since edges repeatedly discovered across diverse algorithms are less likely to be artifacts of any single method and more likely to reflect stable causal patterns in the data. This consistency shows that CDAF preserves the core structural relationships recognized by other approaches, while further extending them with effect quantification and clinical interpretability.

5 Discussion

In medical research, exploring risk factors of diseases can reveal pathogenesis and support the development of precision medicine. Traditional statistical methods and machine learning methods only identify correlations and do not adequately discover the essential causal relationships, which limits their application. Therefore, we propose the CDAF to bridge the gap between causal learning and traditional medical methods. We design a causal disentanglement method for observable medical data to reduce the influence of confounding factors, then discover and quantify the causal relationships of disease risk factors through the converging causality growth algorithm and causal intervention modeling method proposed.

Experiments demonstrate that CDAF effectively identifies

high risk factors for stomach cancer, liver cancer, and AMI, which are highly consistent with existing research in the medical field. Fig. 7-9 show the direct causal, indirect causal and other risk factors of these three diseases in the form of causal graph clearly. From Fig. 10, we identify patterns of brain iron deposition in Parkinson's patients, offering new insights into the pathogenesis of neurodegenerative diseases. Besides, we verify the effectiveness and robustness through the declining trend of the BIC score and refutation tests, as shown in Table 1-3. As we show in Fig. 12, when comparing to other causal discovery methods, our result identifies more accurate causal relationships.

During the transition to precision medicine practice, CDAF guides clinical decision making based on this two-step workflow: (i) stratifying patients according to direct versus indirect causal factors identified in the causal graph, and (ii) prioritizing interventions within each stratum by ranking factors with the largest average treatment effects (ATE). For illustration, a patient with hypertension and self-rated poor health would be placed in a high-priority stratum due to the presence of direct causal factors. Among these, hypertension exhibits a higher ATE than self-rated health, indicating that intensive blood pressure management should be prioritized, while self-rated health can inform complementary monitoring strategies. This workflow connects our methodology with translational precision medicine. And we emphasize that this workflow represents a methodological reference rather than a ready-to-use clinical protocol; its translation into practice must be guided by clinicians, established medical interventions, and accumulated clinical expertise.

This innovative CDAF not only provides a new path to explore causality in medical data, but also combines causal learning with the experience and knowledge of traditional medicine to form a powerful tool. Our CDAF goes beyond the general correlation and learns the most essential causal relationship behind the data, so as to provide a more accurate basis for the diagnosis and treatment of diseases, and has a broad application prospect.

However, it is worth noting that this study has certain limitations. The manual feature selection process may have excluded potentially relevant factors, which we aim to address in future work by incorporating disease-specific prior knowledge. In addition, while linear regression was employed as an interpretable baseline for effect estimation, CDAF is inherently estimator-agnostic and can be extended to more advanced nonlinear methods. Besides, challenges such as high-dimensional variables, temporal dependencies, and cross-population validation impose demanding requirements on datasets and were beyond the scope of this study. Future research will pursue these directions with more targeted improvements, ultimately strengthening CDAF's role as a systematic and generalizable framework for uncovering disease etiology and advancing precision medicine. Finally, while residual bias from unmeasured confounding is inevitable in observational studies, CDAF mitigates spurious associations and provides methodological causal evidence that requires further validation and fairness considerations in future applications.

6 Conclusions

The CDAF we proposed provides an efficient tool for exploring the causal risk factors of diseases. The framework first disentangles causal and non-causal factors, then uses the converging causality growth algorithm to identify the hierarchical causal graph of causal relationships, and finally quantifies the degree of influence through causal intervention modeling to rank their importance. Experimental results show that CDAF can accurately reveal the causal risk factors of stomach cancer, liver cancer, AMI and Parkinson's disease, which are highly consistent with existing medical research. Overall, CDAF, as a comprehensive causal learning framework, bridges the gap between causality and clinical application, providing a powerful tool for discovering the causes of diseases, which demonstrates significant potential in advancing precision medicine. In the future, we plan to enhance CDAF by integrating multi-omics data and exploring the integration of domain-specific prior knowledge, in order to capture a more comprehensive picture of disease mechanisms and improve the generalizability of the framework.

Appendix

A.1. Proof of Theorem 1

According to mutual information theory, for causal factors C, confounding factors S and outcome Y, we have:

$$I(C;S) \leqslant H(C) \tag{A.1-1}$$

where H(·) is the entropy. Since the causal factor C completely determines the outcome Y (i.e., C is in one-to-one correspondence with Y without noise interference), so we have:

$$I(C;Y) = H(C) = H(Y) \tag{A.1-2}$$

Furthermore, we have:

$$I(S;Y) \leqslant H(Y) \tag{A.1-3}$$

Therefore, the following inequality holds:

$$I(C;Y) = H(Y) \geqslant I(S;Y) \tag{A.1-4}$$

And the equality only holds when the distributions of C and S are exactly the same. Based on the causal model $\mathcal{M}_C : Y \sim (C, \theta_C, U_Y)$ and non-causal model $\mathcal{M}_S : Y \sim (S, \theta_S, U_Y)$, thus, using marginal likelihood to fit I, we obtain the following:

$$Score(D \mid \mathcal{M}_C) > Score(D \mid \mathcal{M}_S) \tag{A.1-5}$$

A.2. Proof of Theorem 2

Assume $Y \perp S \mid C$, then

$$P(Y \mid C, S = s) = P(Y|C), \forall s \tag{A.2-1}$$

Substituting into the definition of $Score(Y|do(C))$:

$$\begin{aligned} Score(Y|do(C)) &= \sum_s P(Y \mid C, S = s)P(S = s) \\ &= \sum_s P(Y|C)P(S = s) = P(Y \mid C) \end{aligned} \tag{A.2-2}$$

Thus the interventional score remains invariant with respect to $do(C)$. And if S is not the direct causal parent of Y, then

$$P(Y \mid S, C = c) = P(Y|C = c) \tag{A.2-3}$$

Substituting into $Score(Y \mid do(S))$:

$$\begin{aligned} Score(Y \mid do(S)) &= \sum_c P(Y|S, C = c)P(C = c) \\ &= \sum_c P(Y|C = c)P(C = c) \end{aligned} \tag{A.2-4}$$

Since S has no direct causal effect on Y, intervening on S does not provide additional information to Y. In the limiting case where the correlation between S and C is fully broken by the intervention, the resulting distribution of C is independent of S, leading to

$$\begin{cases} Score(Y \mid do(S)) \to 0 \\ Score(Y \mid do(C)) \; remains \; unchanged \end{cases} \tag{A.2-5}$$

Acknowledgment

This research was funded by the National Key Research and Development Plan "Inter-governmental International Science and Technology Innovation Cooperation" Program, Minis of Science and Technology (Grant No. 2025YFE0112100), the Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education (Grant NO.2025102), and the Fundamental Research Funds for the Central Universities 2025XC11020.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Reference

[1] K. B. Johnson, W.-Q. Wei, D. Weeraratne, M. E. Frisse, K. Misulis, K. Rhee, J. Zhao, and J. L. Snowdon, "Precision medicine, ai, and the future of personalized health care," *Clinical and translational science*, vol. 14, no. 1, pp. 86–93, 2021.

[2] M. R. Subbiah, "Nutrigenetics and nutraceuticals: the next wave riding on personalized medicine," *Translational Research*, vol. 149, no. 2, pp. 55–61, 2007.

[3] Z. Ahmed, K. Mohamed, S. Zeeshan, and X. Dong, "Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine," *Database*, vol. 2020, p. baaa010, 2020.

[4] A. Alyass, M. Turcotte, and D. Meyre, "From big data analysis to personalized medicine for all: challenges and opportunities," *BMC medical genomics*, vol. 8, pp. 1–12, 2015.

[5] X. Chen, "Analyses and concerns in precision medicine: A statistical perspective," *arXiv preprint arXiv:2401.06899*, 2024.

[6] B. Myors, K. R. Murphy, and A. Wolach, *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests*. Routledge, 2010.

[7] T. K. Kim, "T test as a parametric statistic," *Korean journal of anesthesiology*, vol. 68, no. 6, pp. 540–546, 2015.

[8] L. St, S. Wold *et al.*, "Analysis of variance (anova)," *Chemometrics and intelligent laboratory systems*, vol. 6, no. 4, pp. 259–272, 1989.

[9] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[10] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013.

[11] J. C. De Winter, "Using the student's t-test with extremely small sample sizes," *Practical Assessment, Research, and Evaluation*, vol. 18, no. 1, p. 10, 2019.

[12] A. Yuan, X. Chen, Y. Zhou, and M. T. Tan, "Subgroup analysis with semiparametric models toward precision medicine," *Statistics in Medicine*, vol. 37, no. 11, pp. 1830–1845, 2018.

[13] P. Xia, Z. Bai, Y. Yao, L. Xu, H. Zhang, L. Du, X. Chen, Q. Ye, Y. Zhu, P. Wang, X. Li, G. Wang, and Z. Fang, "Advanced deep neural network with unified feature-aware and label embedding for multi-label arrhythmias classification," *Tsinghua Science and Technology*, vol. 30, no. 3, pp. 1251–1269, 2025.

[14] P. W. Wilson, J. B. Meigs, L. Sullivan, C. S. Fox, D. M. Nathan, and R. B. D'Agostino, "Prediction of incident diabetes mellitus in middle-aged adults: the framingham offspring study," *Archives of internal medicine*, vol. 167, no. 10, pp. 1068–1074, 2007.

[15] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN computer science*, vol. 2, no. 3, p. 160, 2021.

[16] H. S. R. Rajula, G. Verlato, M. Manchia, N. Antonucci, and V. Fanos, "Comparison of conventional statistical methods with machine learning in medicine: Diagnosis, drug development, and treatment," *Medicina*, vol. 56, no. 9, p. 455, Sep. 2020.

[17] C. Cortes, "Support-vector networks," *Machine Learning*, 1995.

[18] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[19] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.

[20] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[21] J. A. Hernesniemi, S. Mahdiani, J. A. Tynkkynen, L.-P. Lyytikäinen, P. P. Mishra, T. Lehtimäki, M. Eskola, K. Nikus, K. Antila, and N. Oksala, "Extensive phenotype data and machine learning in prediction of mortality in acute coronary syndrome–the maddec study," *Annals of medicine*, vol. 51, no. 2, pp. 156–163, 2019.

[22] J. Nudel, A. M. Bishara, S. W. de Geus, P. Patil, J. Srinivasan, D. T. Hess, and J. Woodson, "Development and validation of machine learning models to predict gastrointestinal leak and venous thromboembolism after weight loss surgery: an analysis of the mbsaqip database," *Surgical endoscopy*, vol. 35, pp. 182–191, 2021.

[23] A. Shimoda, D. Ichikawa, and H. Oyama, "Prediction models to identify individuals at risk of metabolic syndrome who are unlikely to participate in a health intervention program," *International journal of medical informatics*, vol. 111, pp. 90–99, 2018.

[24] D. C. Castro, I. Walker, and B. Glocker, "Causality matters in medical imaging," *Nature Communications*, vol. 11, no. 1, p. 3673, Jul. 2020.

[25] J. G. Richens, C. M. Lee, and S. Johri, "Improving the accuracy of medical diagnosis with causal machine learning," *Nature Communications*, vol. 11, no. 1, p. 3923, Aug. 2020.

[26] N. Tagasovska, V. Chavez-Demoulin, and T. Vatter, "Distinguishing cause from effect using quantiles: Bivariate quantile causal discovery," in *International Conference on Machine Learning*. PMLR, 2020, pp. 9311–9323.

[27] N. Draper, *Applied regression analysis*. McGraw-Hill. Inc, 1998.

[28] M. Jike, O. Itani, N. Watanabe, D. J. Buysse, and Y. Kaneita, "Long sleep duration and health outcomes: a systematic review, meta-analysis and meta-regression," *Sleep medicine reviews*, vol. 39, pp. 25–36, 2018.

[29] M. Meysami, V. Kumar, M. Pugh, S. T. Lowery, S. Sur, S. Mondal, and J. M. Greene, "Utilizing logistic regression to compare risk factors in disease modeling with imbalanced data: a case study in vitamin d and cancer incidence," *Frontiers in Oncology*, vol. 13, p. 1227842, 2023.

[30] D.-Y. Oh and Y.-J. Bang, "Her2-targeted therapies—a role beyond breast cancer," *Nature reviews Clinical oncology*, vol. 17, no. 1, pp. 33–48, 2020.

[31] P. W. Franks, W. T. Cefalu, J. Dennis, J. C. Florez, C. Mathieu, R. W. Morton, M. Ridderstråle, H. H. Sillesen, and C. D. Stehouwer, "Precision medicine for cardiometabolic disease: a framework for clinical translation," *The Lancet Diabetes & Endocrinology*, vol. 11, no. 11, pp. 822–835, 2023.

[32] T. M. Mitchell and T. M. Mitchell, *Machine learning*. McGraw-hill New York, 1997, vol. 1, no. 9.

[33] G. Hinton and T. J. Sejnowski, *Unsupervised learning: foundations of neural computation*. MIT press, 1999.

[34] Z. Li, X. Jiang, Y. Wang, and Y. Kim, "Applied machine learning in alzheimer's disease research: omics, imaging, and clinical data," *Emerging topics in life sciences*, vol. 5, no. 6, pp. 765–777, 2021.

[35] E. A. Patrick and F. P. Fischer III, "A generalized k-nearest neighbor rule," *Information and control*, vol. 16, no. 2, pp. 128–152, 1970.

[36] H. Husi, *Computational biology*. Codon Publications, 2019.

[37] Q. Zhao, H. Xu, J. Li, F. A. Rajput, and L. Qiao, "The application of artificial intelligence in alzheimer's research," *Tsinghua Science and Technology*, vol. 29, no. 1, pp. 13–33, 2024.

[38] L. Lan, G. Hu, R. Li, T. Wang, L. Jiang, J. Luo, Z. Ji, and Y. Wang, "Machine learning for selecting important clinical markers of imaging subgroups of cerebral small vessel disease based on a common data model," *Tsinghua Science and Technology*, vol. 29, no. 5, pp. 1495–1508, 2024.

[39] J. Cruz, "Ds wishart applications of machine learning in cancer prediction and prognosis., 2007, 2," *DOI: https://doi.org/10.1177/117693510600200030*, pp. 59–77.

[40] S. J. Al'Aref, K. Anchouche, G. Singh, P. J. Slomka, K. K. Kolli, A. Kumar, M. Pandey, G. Maliakal, A. R. Van Rosendael, A. N. Beecy *et al.*, "Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging," *European heart journal*, vol. 40, no. 24, pp. 1975–1986, 2019.

[41] J. Pearl, *Causality*. Cambridge university press, 2009.

[42] H. Gao, C. Yao, J. Li, L. Si, Y. Jin, F. Wu, C. Zheng, and H. Liu, "Rethinking causal relationships learning in graph neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 11, 2024, pp. 12 145–12 154.

[43] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1228–1242, 2018.

[44] D. Lopez-Paz, R. Nishihara, S. Chintala, B. Scholkopf, and L. Bottou, "Discovering causal signals in images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6979–6987.

[45] J. Pearl, "Causal inference in statistics: An overview," 2009.

[46] X. Wang, Q. Zhao, L. Wang, and W. Liu, "Causality-based contrastive incremental learning framework for domain generalization," *Tsinghua Science and Technology*, vol. 30, no. 4, pp. 1636–1647, 2025.

[47] J. Yang, L. Jiang, K. Xie, Q. Chen, and A. Wang, "Lung nodule detection algorithm based on rank correlation causal structure learning," *Expert Systems with Applications*, vol. 216, p. 119381, 2023.

[48] D. Chen, H. Zhao, J. He, Q. Pan, and W. Zhao, "An causal xai diagnostic model for breast cancer based on mammography reports," in *2021 IEEE international conference on bioinformatics and biomedicine (BIBM)*. IEEE, 2021, pp. 3341–3349.

[49] X. Li, R. Guo, J. Lu, T. Chen, and X. Qian, "Causality-driven graph neural network for early diagnosis of pancreatic cancer in non-contrast computerized tomography," *IEEE Transactions on Medical Imaging*, vol. 42, no. 6, pp. 1656–1667, Jun. 2023.

[50] M. da Câmara Ribeiro-Dantas, H. Li, V. Cabeli, L. Dupuis, F. Simon, L. Hettal, A.-S. Hamy, and H. Isambert, "Learning interpretable causal networks from very large datasets, application to 400,000 medical records of breast cancer patients," *Iscience*, vol. 27, no. 5, 2024.

[51] E. Gao, I. Ng, M. Gong, L. Shen, W. Huang, T. Liu, K. Zhang, and H. Bondell, "Missdag: Causal discovery in the presence of missing data with continuous additive noise models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5024–5038, 2022.

[52] A. Zanga, A. Bernasconi, P. J. Lucas, H. Pijnenborg, C. Reijnen, M. Scutari, and F. Stella, "The impact of missing data on causal discovery: A multicentric clinical study," *arXiv preprint arXiv:2305.10050*, 2023.

[53] F. G. Febrinanto, A. Simango, C. Xu, J. Zhou, J. Ma, S. Tyagi, and F. Xia, "Refined causal graph structure learning via curvature for brain disease classification," *Artificial Intelligence Review*, vol. 58, no. 8, p. 222, 2025.

[54] C. Meek, "Causal inference and causal explanation with background knowledge," *arXiv preprint arXiv:1302.4972*, 2013.

[55] D. M. Chickering, "Optimal structure identification with greedy search," *Journal of machine learning research*, vol. 3, no. Nov, pp. 507–554, 2002.

[56] J. Pearl, *Causal inference in statistics: a primer*. John Wiley & Sons, 2016.

[57] M. Baiocchi, J. Cheng, and D. S. Small, "Instrumental variable methods for causal inference," *Statistics in medicine*, vol. 33, no. 13, pp. 2297–2340, 2014.

[58] Z. Chen, L. Lee, J. Chen, R. Collins, F. Wu, Y. Guo, P. Linksted, and R. Peto, "Cohort profile: the kadoorie study of chronic disease in china (kscdc)," *International journal of epidemiology*, vol. 34, no. 6, pp. 1243–1249, 2005.

[59] Z. Chen, J. Chen, R. Collins, Y. Guo, R. Peto, F. Wu, and L. Li, "China kadoorie biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up," *International journal of epidemiology*, vol. 40, no. 6, pp. 1652–1666, 2011.

[60] J. Poorolajal, L. Moradi, Y. Mohammadi, Z. Cheraghi, and F. Gohari-Ensaf, "Risk factors for stomach cancer: a systematic review and meta-analysis," *Epidemiology and Health*, vol. 42, p. e2020004, Feb. 2020.

[61] Y. Minami, S. Kanemura, T. Oikawa, S. Suzuki, Y. Hasegawa, K. Miura, Y. Nishino, Y. Kakugawa, and T. Fujiya, "Associations of cigarette smoking and alcohol drinking with stomach cancer survival: A prospective patient cohort study in japan," *International Journal of Cancer*, vol. 143, no. 5, pp. 1072–1085, Sep. 2018.

[62] K. D. Crew and A. I. Neugut, "Epidemiology of gastric cancer," *World journal of gastroenterology: WJG*, vol. 12, no. 3, p. 354, 2006.

[63] G. Luo, Y. Zhang, P. Guo, L. Wang, Y. Huang, and K. Li, "Global patterns and trends in stomach cancer incidence: Age, period and birth cohort analysis," *International Journal of Cancer*, vol. 141, no. 7, pp. 1333–1344, 2017.

[64] S. Tsugane and S. Sasazuki, "Diet and the risk of gastric cancer: review of epidemiological evidence," *Gastric Cancer*, vol. 10, no. 2, pp. 75–83, Jun. 2007.

[65] W.-C. You, W. Blot, Y.-S. Chang, A. Ershow, Z.-T. Yang, Q. An, B. Henderson, G.-W. Xu, J. Fraumeni, and T.-G. Wang, "Diet and high risk of stomach cancer in shandong, china," *Health Policy*, vol. 14, no. 2, pp. 156–157, Mar. 1990.

[66] Richa, N. Sharma, and G. Sageena, "Dietary factors associated with gastric cancer-a review," *Translational Medicine Communications*, vol. 7, no. 1, p. 7, 2022.

[67] H. Bahrami and M. Tafrihi, "Global trends of cancer: The role of diet, lifestyle, and environmental factors," *Cancer innovation*, vol. 2, no. 4, pp. 290–301, 2023.

[68] A. Marengo, C. Rosso, and E. Bugianesi, "Liver cancer: Connections with obesity, fatty liver, and cirrhosis," *Annual Review of Medicine*, vol. 67, no. 1, pp. 103–117, Jan. 2016.

[69] E. E. Calle, C. Rodriguez, K. Walker-Thurmond, and M. J. Thun, "Overweight, obesity, and mortality from cancer in a prospectively studied cohort of u.s. adults," *New England Journal of Medicine*, vol. 348, no. 17, pp. 1625–1638, Apr. 2003.

[70] Y. Fu, L. Maccioni, X. W. Wang, T. F. Greten, and B. Gao, "Alcohol-associated liver cancer," *Hepatology*, vol. 80, no. 6, pp. 1462–1479, 2024.

[71] S.-C. Chuang, Y.-C. A. Lee, G.-J. Wu, K. Straif, and M. Hashibe, "Alcohol consumption and liver cancer risk: a meta-analysis," *Cancer Causes & Control*, vol. 26, no. 9, pp. 1205–1231, 2015.

[72] P. T. Campbell, C. C. Newton, N. D. Freedman, J. Koshiol, M. C. Alavanja, L. E. Beane Freeman, J. E. Buring, A. T. Chan, D. Q. Chong, M. Datta *et al.*, "Body mass index, waist circumference, diabetes, and risk of liver cancer for us adults," *Cancer research*, vol. 76, no. 20, pp. 6076–6083, 2016.

[73] Y. Wang, B. Wang, F. Shen, J. Fan, and H. Cao, "Body mass index and risk of primary liver cancer: a meta-analysis of prospective studies," *The oncologist*, vol. 17, no. 11, pp. 1461–1468, 2012.

[74] Z.-J. Zhang, Z.-J. Zheng, R. Shi, Q. Su, Q. Jiang, and K. E. Kip, "Metformin for liver cancer prevention in patients with type 2 diabetes: a systematic review and meta-analysis," *The Journal of Clinical Endocrinology & Metabolism*, vol. 97, no. 7, pp. 2347–2353, 2012.

[75] Y. Pang, C. Kartsonaki, I. Turnbull, Y. Guo, R. Clarke, Y. Chen,

F. Bragg, L. Yang, Z. Bian, I. Y. Millwood *et al.*, "Diabetes, plasma glucose, and incidence of fatty liver, cirrhosis, and liver cancer: a prospective study of 0.5 million people," *Hepatology*, vol. 68, no. 4, pp. 1308–1318, 2018.

[76] H. B. El-Serag, T. Tran, and J. E. Everhart, "Diabetes increases the risk of chronic liver disease and hepatocellular carcinoma," *Gastroenterology*, vol. 126, no. 2, pp. 460–468, 2004.

[77] M. T. Roe, A. R. Halabi, R. H. Mehta, A. Y. Chen, L. K. Newby, R. A. Harrington, S. C. Smith Jr, E. M. Ohman, W. B. Gibler, and E. D. Peterson, "Documented traditional cardiovascular risk factors and mortality in non–st-segment elevation myocardial infarction," *American heart journal*, vol. 153, no. 4, pp. 507–514, 2007.

[78] S. Yusuf, S. Hawken, S. Ôunpuu, T. Dans, A. Avezum, F. Lanas, M. McQueen, A. Budaj, P. Pais, J. Varigos *et al.*, "Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the interheart study): case-control study," *The lancet*, vol. 364, no. 9438, pp. 937–952, 2004.

[79] G. Waller, U. Janlert, M. Norberg, R. Lundqvist, and A. Forssén, "Self-rated health and standard risk factors for myocardial infarction: a cohort study," *BMJ open*, vol. 5, no. 2, p. e006589, 2015.

[80] F. Sanchis-Gomar, C. Perez-Quilis, R. Leischik, and A. Lucia, "Epidemiology of coronary heart disease and acute coronary syndrome," *Annals of translational medicine*, vol. 4, no. 13, p. 256, 2016.

[81] R. Pedrinelli, P. Ballo, C. Fiorentini, S. Denti, M. Galderisi, A. Ganau, G. Germanò, P. Innelli, A. Paini, S. Perlini *et al.*, "Hypertension and acute myocardial infarction: an overview," *Journal of cardiovascular medicine*, vol. 13, no. 3, pp. 194–202, 2012.

[82] D. G. Kang, M. H. Jeong, Y. Ahn, S. C. Chae, S. H. Hur, T. J. Hong, Y. J. Kim, I. W. Seong, J. K. Chae, J. Y. Rhew *et al.*, "Clinical effects of hypertension on the mortality of patients with acute myocardial infarction," *Journal of Korean medical science*, vol. 24, no. 5, pp. 800–806, 2009.

[83] M. S. Lockheart, L. M. Steffen, H. M. Rebnord, R. L. Fimreite, J. Ringstad, D. S. Thelle, J. I. Pedersen, and D. R. Jacobs Jr, "Dietary patterns, food groups and myocardial infarction: a case–control study," *British Journal of Nutrition*, vol. 98, no. 2, pp. 380–387, 2007.

[84] I. Biyik and O. Ergene, "Alcohol and acute myocardial infarction," *Journal of international medical research*, vol. 35, no. 1, pp. 46–51, 2007.

[85] G. M. Halliday, "Thalamic changes in parkinson's disease," *Parkinsonism & related disorders*, vol. 15, pp. S152–S155, 2009.

[86] D. Sigirli, S. T. Ozdemir, S. Erer, I. Sahin, I. Ercan, R. Ozpar, M. O. Orun, and B. Hakyemez, "Statistical shape analysis of putamen in early-onset parkinson's disease," *Clinical Neurology and Neurosurgery*, vol. 209, p. 106936, 2021.

[87] J. Pasquini, R. Durcan, L. Wiblin, M. G. Stokholm, L. Rochester, D. J. Brooks, D. Burn, and N. Pavese, "Clinical implications of early caudate dysfunction in parkinson's disease," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 90, no. 10, pp. 1098–1104, 2019.

[88] X. Guan, M. Xuan, Q. Gu, P. Huang, C. Liu, N. Wang, X. Xu, W. Luo, and M. Zhang, "Regionally progressive accumulation of iron in parkinson's disease as measured by quantitative susceptibility mapping," *NMR in Biomedicine*, vol. 30, no. 4, p. e3489, 2017.

[89] R. Y. Rohekar, S. Nisimov, Y. Gurwicz, and G. Novik, "Iterative causal discovery in the possible presence of latent confounders and selection bias," *Advances in Neural Information Processing Systems*, vol. 34, pp. 2454–2465, 2021.

[90] S. Lachapelle, P. Brouillard, T. Deleu, and S. Lacoste-Julien, "Gradient-based neural dag learning," Feb. 2020.

[91] S. Zhu, I. Ng, and Z. Chen, "Causal discovery with reinforcement learning," *arXiv preprint arXiv:1906.04477*, 2019.

[92] X. Wang, Y. Du, S. Zhu, L. Ke, Z. Chen, J. Hao, and J. Wang, "Ordering-based causal discovery with reinforcement learning," *arXiv preprint arXiv:2105.06631*, 2021.

[93] X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing, "Dags with no tears: Continuous optimization for structure learning," *Advances in neural information processing systems*, vol. 31, 2018.

[94] I. Ng, A. Ghassami, and K. Zhang, "On the role of sparsity and dag constraints for learning linear dags," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 943–17 954, 2020.

[95] K. Bello, B. Aragam, and P. Ravikumar, "Dagma: Learning dags via m-matrices and a log-determinant acyclicity characterization," *Advances in Neural Information Processing Systems*, vol. 35, pp. 8226–8239, 2022.

[96] P. Sanchez, X. Liu, A. Q. O'Neil, and S. A. Tsaftaris, "Diffusion models for causal discovery via topological ordering," in *The Eleventh International Conference on Learning Representations*.

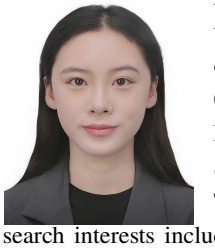
[97] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, prediction, and search*. MIT press, 2000.

[98] P. Spirtes, C. Meek, and T. Richardson, "Causal inference in the presence of latent variables and selection bias," in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 1995, pp. 499–506.

[99] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan, "Causal protein-signaling networks derived from multiparameter single-cell data," *Science*, vol. 308, no. 5721, pp. 523–529, 2005.

[100] S. L. Lauritzen and D. J. Spiegelhalter, "Local computations with probabilities on graphical structures and their application to expert systems," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 50, no. 2, pp. 157–194, 1988.

Author biography

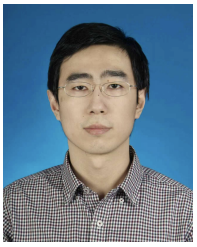


Zhao Wuyan received the B.S. degree from the School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China, in 2024. She is currently pursuing the M.Sc. degree with the School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China. Her current research interests include causal learning, data mining, and deep learning.

Peng Li works in the School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China. His research interests include complexity research and artificial intelligence.

Canqing Yu received the B.S. degree and Ph.D. degrees from Peking University, Beijing, China. He is currently a research fellow with the School of Public Health, Peking University, Beijing, China. He serves as an editorial board member for several journals, including Health Data Science and China CDC Weekly, and holds leadership positions in professional associations such as the Chinese Association for Health Promotion and Education. His research interests focus on the epidemiology of major chronic diseases and the application of data science in large-scale cohort studies.

Wei Wei is currently an associate research fellow and the deputy director of the Office of Medical Digital Intelligence Innovation Center at Beijing Friendship Hospital. Her research interests include the application of artificial intelligence algorithms in clinical medicine.



Xu Yang is currently an associate professor with the School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China. He received the B.S. and Ph.D. degrees from Tsinghua University, Beijing, China. His research interests include brain-inspired AI and causal-based AI.