JOURNAL OF DATA SCIENCE | Vol.2026, Issue 3, No.8
eISSN:2805-5160

Causal Discovery-Based Data Mining for Explainable Risk Factor Identification in Public Health Data

Rinci Kembang Hapsari[1]*, Kusmaya[2], John Edmon A. Alcomendas[3]

[1]Informatics Engineering Department, Institut Teknologi Adhi Tama Surabaya, East Java, Indonesia
[2]Universitas Langlangbuana, Indonesia
[3]Department of Mathematics and Statistics, Polytechnic University of the Philippines, Manila, Philippines

Email: rincikembang@itats.ac.id[1]*, kusmaya7165@gmail.com[2], john_edmon56@yahoo.com[3]

Abstract

Conventional machine learning methods are widely used in public health analytics, but many models mainly capture statistical correlations without clearly distinguishing causal relationships. This limitation reduces interpretability and weakens the reliability of data-driven recommendations for health intervention planning. This study proposes a causal discovery-based data mining framework for explainable risk factor identification in public health data. A controlled synthetic public health dataset was constructed to represent demographic, lifestyle, clinical, and socioeconomic factors associated with chronic disease risk. The proposed framework integrates data preprocessing, Bayesian Network modelling, PC algorithm-based causal structure learning, causal graph interpretation, and comparative evaluation against correlation-based analysis. The framework was evaluated using structural Hamming distance, prediction accuracy, causal consistency, and interpretability assessment. The results show that the proposed causal discovery framework achieved 86.40% prediction accuracy, 0.82 causal consistency score, and lower structural error than individual causal discovery models. The discovered graph identified plausible dependency pathways involving body mass index, systolic blood pressure, glucose level, smoking status, physical activity, and socioeconomic status. Compared with correlation-based analysis, the proposed framework provided more structured explanations by distinguishing direct risk-related variables, indirect pathways, and potential upstream intervention targets. This study contributes to explainable data mining by demonstrating how causal discovery can support risk factor interpretation and evidence-oriented public health decision-making.

Keywords

Causal discovery; Bayesian networks; Explainable data mining; Risk factor identification; Health risk

Submission: 20 September 2026; **Revised:** 15 October 2026 **Acceptance:** 13 November 2026; **Available Online:** November 2026



Copyright: © 2026. All the authors listed in this paper. The distribution, reproduction, and any other usage of the content of this paper is permitted, with credit given to all the author(s) and copyright owner(s) in accordance to common academic practice. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license, as stated in the website: https://creativecommons.org/licenses/by/4.0/

Introduction

It is for this reason that data mining techniques have come to be relied upon for predicting future scenarios from current information. Data-driven solutions provide public health systems with the means to analyze disease risk factors, design preventive efforts, and devise population-targeted interventions. The rapid proliferation of multiple sources of health data, including records of electronic health data systems (EHR), public health and community health screening surveys, community-based screening initiatives, and digitized epidemiological data, offers unprecedented opportunities for large-scale analyses of health data (Chakarverti et al., 2019).

Most conventional machine learning models possess the ability to yield high levels of prediction. However, many are inherently geared to identify statistical associations. This is problematic for health care analytics for the simple reason that explainable AI (XAI) systems are needed to enhance the interpretability and understandability of model outputs to better assist the health care domain and public health systems (Singh et al., 2026). Variables may exhibit strong statistical associations with disease risk without necessarily representing direct causal relationships. In predictive models, many variables may be considered insignificant, but can still affect disease risk through indirect pathways. This limitation can constrain the use of predictive models for public health planning and clinical decision support. It is essential to identify which risk factors provide direct pathways, which offer indirect pathways, and which provide opportunities for risk factors that ensure pathways are more effectively targeted.

Causal discovery methods estimate plausible dependency structures and directional pathways among variables under defined assumptions. Causal discovery measures dependency structures among variables and represents such structures in the form of causal graphs. In a causal graph, variables are represented as nodes, and possible causal relationships are represented as directed edges. With such structures, it becomes possible to express whether lifestyle behaviours affect clinical indicators, whether clinical indicators affect disease risk, and whether socioeconomic indicators affect disease outcomes through behavioural pathways.

The distinction between statistical association and causal interpretation is particularly important in public health research. (Indrayan, 2025) This is evidenced by the fact that public health prevention campaigns targeting modifiable risk factors are recommended. Intervention prioritization is often the problem in public health. (Teh & Woon, 2021) For instance, the intervention prioritization problem may manifest itself in the decision to target a variable that influences risk indirectly, such as a socioeconomic variable, through the loss of access to healthy behaviours. (Cilloni & Dowdy, 2024) Causal discovery may support the identification of plausible intervention pathways by distinguishing upstream, intermediate, and more proximal risk-related variables.

Utilizing causal discovery techniques in public health data mining poses a number of difficulties. There are mixed variable types, missing values, and confounding and interdependent relationships among many risk factors in the public health data. Furthermore, a lot of data mining research focuses on model accuracy as opposed to the causal interpretability of the model. (Zhang et al., 2025) Therefore, a functional framework is needed to incorporate causal discovery algorithms and address explainable risk factor identification with evaluation criteria specific to public health data.

The framework builds on the integration of several techniques and tools, namely, data mining and risk factor identification, Bayesian Network, and evaluation in terms of prediction accuracy and interpretability. This study advances the idea of explainable risk factor identification with causal discovery in public health data. (You et al., 2025) The study examines how causal

discovery can support interpretable risk-factor analysis and identify plausible pathways relevant to public health decision-making.

Methodology

Research Design

This study adopted a quantitative computational research design based on the integration of data mining and causal discovery. The study extended beyond health-risk prediction by examining risk-factor relationships through an explainable causal-discovery framework. The proposed framework encompassed seven core functions: dataset construction, data preprocessing, variable selection, causal discovery modelling, causal graph construction, risk-factor interpretation, and model evaluation.

The workflow of the proposed framework is illustrated in Figure 1. Initially, public health data were preprocessed through cleaning, encoding, and normalization. After that, some of the data were used to learn the causal structures through Bayesian Network and the PC algorithm. Direct, indirect, and upstream risk factors were interpreted from the causal graph. Finally, the framework was tested through structural, predictive, and interpretability assessments.

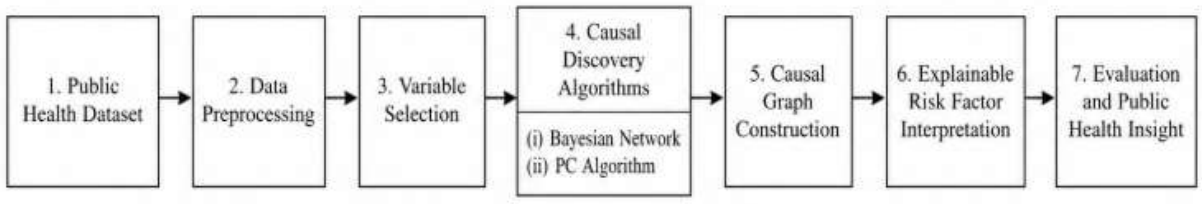


Figure 1. Proposed causal discovery-based data mining framework

Figure 1 depicts the process of the proposed framework. The framework starts from the preparation of public health data and ends at the causal interpretation, which provides the explanation of risk factors and guides the planning of public health interventions.

Dataset and Variables

A controlled synthetic public health dataset was constructed to enable evaluation against a predefined causal reference structure. The dataset was designed to represent demographic, socioeconomic, lifestyle, and clinical variables commonly considered in chronic-disease risk analysis. The dataset included 1,200 records and 10 variables.

The variables included in the dataset were age, gender, socioeconomic status, smoking status, physical activity and diet, body mass index, blood pressure, blood glucose, and disease risk status. Disease risk status was defined as a binary outcome, with 1 representing high risk and 0 representing low risk. Dataset construction followed a predefined dependency structure in which socioeconomic and lifestyle variables influenced selected clinical indicators, which in turn contributed to disease-risk status. This design enabled evaluation of whether the proposed framework could recover meaningful causal relationships from the dataset.

Dataset construction used a fixed random seed of 42 to support reproducibility of the computational experiment. Age was generated within the range of 18 to 80 years, while gender was represented as a binary categorical variable. Socioeconomic status was generated as an ordinal variable with three levels, namely low, medium, and high. Lifestyle variables were configured to represent plausible patterns relevant to public health risk analysis. Smoking status was generated as a binary behavioural variable, while physical activity was represented as weekly activity frequency and diet quality was represented using a numerical score ranging from low to high dietary quality. Clinical indicators were generated using dependency assumptions in which body mass index was influenced by physical activity and diet quality, while systolic blood pressure and glucose level were influenced by age and body mass index. Disease risk status was then generated as a binary outcome using a weighted risk score derived from body mass index, systolic blood pressure, glucose level, smoking status, age, and lifestyle-related variables. Synthetic records with risk scores above the predefined threshold were labelled as high risk, while the remaining records were labelled as low risk. This approach provided a known reference structure for evaluating causal recovery while maintaining plausible relationships among variables of differing public health relevance.

To clarify the structure of the synthetic public health dataset, Table 1 describes the variable groups, individual variables, data types, and roles in analysis for the proposed causal discovery framework.

Table 1. Dataset characteristics and variable description

Variable Group	Variable	Data Type	Role in Analysis
Demographic	Age	Numerical	Background risk factor
Demographic	Gender	Categorical	Background characteristic
Socioeconomic	Socioeconomic status	Ordinal categorical	Upstream determinant
Lifestyle	Smoking status	Binary categorical	Behavioural risk factor
Lifestyle	Physical activity	Numerical	Protective lifestyle factor
Lifestyle	Diet quality	Numerical	Behavioural health factor
Clinical	Body mass index	Numerical	Mediating clinical indicator
Clinical	Systolic blood pressure	Numerical	Direct clinical risk factor
Clinical	Glucose level	Numerical	Direct clinical risk factor
Outcome	Disease risk status	Binary categorical	Target outcome

Table 1 presents details about the variables that comprise the synthetic public health dataset. The variables were chosen to depict the relationships frequently described in public health analytics, where demographic and socioeconomic conditions are likely to affect behavioral patterns, behavioral patterns are likely to affect clinical measurements, and clinical measurements are likely to affect the risk of developing diseases.

The controlled dataset provided a known reference structure against which the recovery of dependency relationships could be evaluated. In public health data, the real causal relationships are typically unknown; consequently, it is not clear whether the discovered causal relationships are valid. Because the dataset was generated based on a hypothesized causal structure, the discovered relationships could be compared with the predefined reference structure to quantify structural error. Although this controlled design does not substitute for validation using operational public health datasets, it provides an appropriate initial setting for evaluating methodological performance.

Data Preprocessing

Prior to conducting the causal discovery modeling, the data were preprocessed. Missing values were assessed and imputed using the median for continuous variables and the mode for categorical variables. Categorical variables were encoded into numerical representations suitable for computational analysis. Continuous data were scaled to the same range to eliminate the effects of different scales among age, body mass index, blood pressure, glucose level, and the quality of physical activity and dietary intake.

An 80:20 configuration was applied to the dataset to create training and testing subsets. The training subset was used for the learning of causal structures, and the testing subset was used for evaluating predictive utility. To eliminate unstable estimations of causal structures, variables with extremely low variance were considered for exclusion. Each of the ten variables was retained in the final preprocessed dataset as each variable was considered to inform the constructed public health risk framework.

Preprocessing was also used to reduce scale-related effects that could distort model estimation. If variables are allowed to coexist in the model despite being measured on disparate numerical scales, they will most likely exert undue influence on model estimation. (Pasinringi et al., 2025) For example, age, glucose level, systolic blood pressure, physical activity, and diet quality are measured using different value ranges. Normalization was therefore applied to ensure that differences in scale did not dominate the causal discovery process. Similarly, categorical encoding was required so that variables such as gender, smoking status, and socioeconomic status could be included in the computational analysis. These steps ensured that the causal discovery algorithms received a stable and consistent input representation.

Causal Discovery Procedure

The proposed framework used Bayesian Network modelling and the PC algorithm as the main causal discovery methods (Wilcox & Meredith, 2026). A Bayesian Network represents the joint probability distribution of variables using a directed acyclic graph. Each variable is conditionally dependent on its parent variables in the graph (Pérez-López et al., 2024). The joint probability distribution is expressed in Equation 1.

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^{n} P\left(X_i \mid Pa(X_i)\right)$$

where X_i represents a variable in the dataset and $Pa(X_i)$ represents the set of parent variables of X_i in the causal graph.

For Bayesian Network modelling, a score-based structure learning strategy was applied. The graph structure was estimated by searching for a directed acyclic graph that maximized the goodness-of-fit score while preserving model parsimony. (Suzuki, 2026) This approach was selected because it is suitable for learning probabilistic dependency structures from tabular public health data. The learned Bayesian Network was then used to represent possible probabilistic dependencies among demographic, lifestyle, clinical, and disease risk variables.

For the PC algorithm, a constraint-based causal discovery procedure was applied using conditional independence testing. The algorithm began with a fully connected undirected graph and removed edges when conditional independence was detected between variables. A significance level of 0.05 was used as the threshold for conditional independence testing. After the skeleton structure was obtained, edge orientation rules were applied to form a directed graph where possible. The resulting graph represented plausible directional dependency pathways among the public health variables. Though mixed public health variables were present in the dataset with some variables generated under controlled synthetic assumptions, the relationships identified from the dataset were interpreted as plausible causal pathways under the assumptions of the controlled experimental design. These relationships should not be interpreted as confirmed causal effects and require validation using real-world data, stronger identification strategies, and domain expertise.

The results from the causal discovery were incorporated into a proposed framework. The final framework focused on the stability of the causal structure, the assessed contribution of the predictors, and the discovered relationships. Based on the developed causal graph, risk factors were classified as direct, indirect, and upstream risk factors.

Evaluation Metrics

The proposed framework was assessed by structural Hamming distance, prediction accuracy, causal consistency, and interpretability assessment. Structural Hamming distance was used to determine the gap between the discovered graph and the predefined reference graph used during construction of the controlled dataset; it is shown in Equation 2.

$$SHD(G, G') = A + D + R$$

where $SHD(G, G')$ represents the structural Hamming distance between the expected graph G and the discovered graph G', A represents the number of edges that need to be added, D represents the number of edges that need to be deleted, and R represents the number of edges with reversed direction.

To determine prediction accuracy, it was evaluated whether the discovered causal structure allowed for classification of disease risk. Causal consistency measured how stable the discovered causal relationships were across repeated samplings. For the interpretability assessment, the discovered relationships were evaluated to determine if they were represented as direct risk factors, indirect risk factors, and pathways along which public health interventions may be relevant.

Causal consistency was defined as the number of stable directed relationships in repeated sampling runs of the dataset. The dataset was resampled multiple times, and causal discovery was employed in each case. A directed relationship was considered consistent when most sampling runs arrived at that relationship. A causal consistency score was computed as the number of stable causal edges over the number of relevant edges in the reference/discovered structure. A causal consistency score of 0.82 indicates that 82% of the relevant directed relationships were repeatedly recovered across resampling runs.

For the interpretability assessment, a qualitative rubric was employed that included three evaluation criteria: the clarity of the relationship direction; the ability to differentiate between direct and indirect risk factors; and the utility of the identified relationship for public health. A result was considered limited when it was able to identify statistical associations, but was not able to define the role of the associations in the causal chain. It was considered high when the method was able to delineate the major risk factors, but still produced a number of obfuscated relationships. It was classified as very high when the method was able to clearly distinguish direct risk factors and indirect pathways and separated upstream factors that could be employed for public health reasoning. Following this rubric, the proposed framework was classified as having very high interpretability as it described pathways of disease risk in a graph-based framework. For the proposed framework, the metrics employed and their interpretations as relevant to structural, predictive, stability, and interpretability perspectives were summarized in Table 2.

Table 2. Evaluation metrics used in the proposed framework

Metric	**Purpose**	**Interpretation**
Structural Hamming Distance	Measures causal graph structure error	Lower value indicates better structural recovery (Kelley, 2022)
Prediction Accuracy	Measures predictive utility of the discovered structure	Higher value indicates stronger disease risk classification support (Nimsarkar, 2024)
Causal Consistency	Measures stability of discovered causal relationships across repeated sampling	Higher value indicates more reliable causal structure (Ali & Suneeta, 2025)
Interpretability Assessment	Evaluates whether discovered relationships can be explained as direct, indirect, or upstream risk factors (Paulisova et al., 2021)	Higher clarity supports public health decision-making

The evaluation metrics used to assess our proposed framework are presented in Table 2. Since data mining based on causal discovery should be evaluated in terms of interpretability and structural correctness in addition to the performance of predictions, these metrics were chosen.

Results and Discussion

Dataset Overview

The final dataset contained 1,200 records and 10 variables. No variables were eliminated, as all of them were pertinent to the public health causal framework constructed. As for the outcome variable, 39.50% of the population was classified as high risk, while 60.50% were classified as low risk. Missing data were addressed through median and mode imputation, as appropriate. Data were partitioned into training and test sets in an 80:20 ratio. The training set was employed for causal structure learning, and the test set for predictive assessment.

The processed dataset was appropriate for causal pathway analysis as it incorporated the relevant variables at the behavioral and clinical levels as well as the outcome variables. Background determinants of the dataset were socioeconomic and demographic variables, while variables at the behavioral level were lifestyle variables. Clinical variables were intermediate health variables, and risk status represented the final health outcome.

Discovered Causal Structure

The proposed framework recovered several meaningful and plausible dependency relationships within the predefined public health structure. From the controlled datasets, socioeconomic status was positioned as an upstream factor associated with physical activity and diet quality, which were subsequently linked to body mass index. Body mass index was connected directionally to systolic blood pressure and glucose level within the recovered graph. The recovered structure contained direct graph connections from systolic blood pressure, smoking status, and body mass index to disease-risk status.

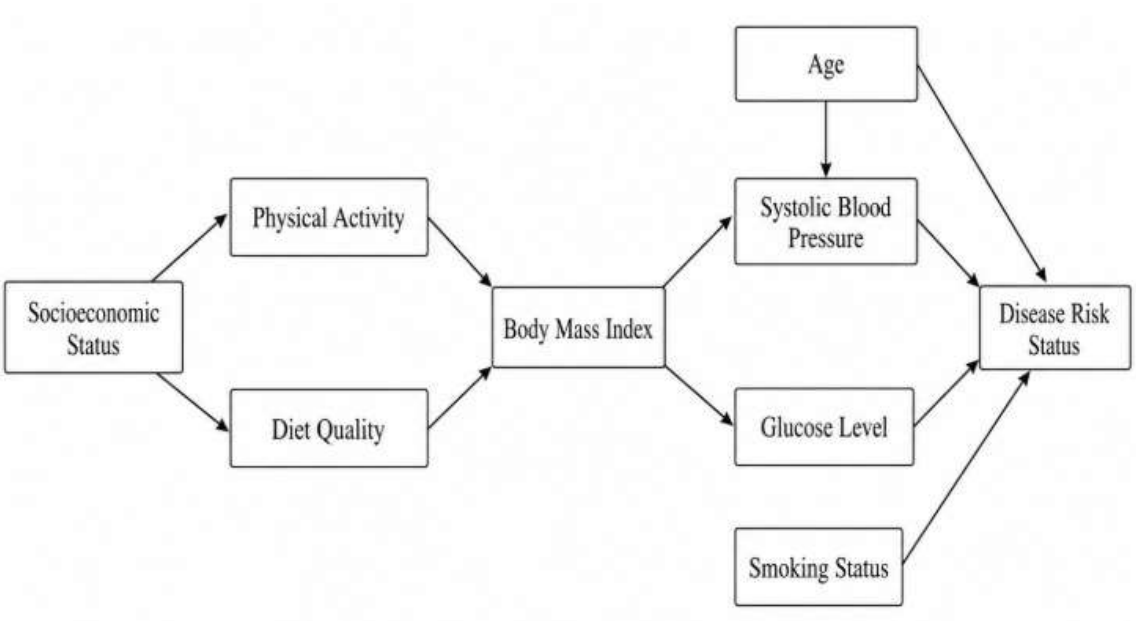


Figure 2. Discovered causal graph of public health risk factors

Figure 2 shows the discovered causal graph in a simplified format with example public health risk factors. This graph shows that disease risk is impacted by clinical risk factors but also lifestyle and socioeconomic factors through indirect pathways.

The discovered graph reflects a clearer model than a correlation-based analysis because it can separate direct and indirect relationships (Chen et al., 2023). For instance, in the model, socioeconomic status may not directly result in disease risk, but can influence lifestyle factors that then affect clinical risk factors (Loughman et al., 2023). Likewise, the model can show that physical activity can indirectly impact disease risk through body mass index and blood pressure. This representation is relevant to public health interpretation because it distinguishes more proximal clinical factors from upstream and lifestyle-related pathways.

From the discovered graph, the risk of adolescent obesity can be seen as the outcome of many factors across individual, family, and school domains, rather than the result of a single, most obvious factor (Huang & Chen, 2026). This is aligned with the nature of many public health concerns, in which risk is typically the result of an interaction between many of the behavioral, biological, and social elements (Wen & Koo, 2025). In this model, risk factors of a more clinical nature, such as blood pressure and glucose levels, are positioned closer to the outcome of disease risk, whereas lifestyle and social factors are more distant in the causal structure. This helps to explain, from a public health perspective, how risk can be addressed at many levels of intervention, from upstream to lifestyle changes to the management of clinical risks.

Explainable Risk Factor Identification

The suggested model placed the recognized variables into direct risk factors, indirect risk factors, and upstream determinants. Variables directly connected to disease-risk status were interpreted as direct risk-related factors, such as systolic blood pressure, glucose level, smoking, and body mass index. Of these variables, body mass index was interpreted as a clinical and metabolic risk-related variable, whereas systolic blood pressure and glucose level were considered clinical risk variables.

Physical activity and dietary quality were considered indirect risk factors, as their impacts were mainly associated with body mass index and other clinical risk variables (Mao & Li, 2025). Insufficient physical activity may elevate body mass index and subsequently raise blood pressure and the risk of disease (Cicek et al., 2025). Similarly, dietary quality may affect body mass index and glucose level, thus creating an indirect pathway to the risk of disease (Nimsarkar, 2024). Of all the variables, socioeconomic status was considered an upstream determinant as it affected the lifestyle-related variables and not the disease outcome. Age was also an important background variable as it affected clinical risk variables like blood pressure.

This interpretation demonstrates how the framework distinguishes the roles of variables across direct, indirect, and upstream risk pathways. This is important to public health as it adds value to the planning of interventions to address clinical risk factors, lifestyle risk factors, and upstream determinants.

Model Evaluation Results

The proposed framework, Bayesian Network modelling, PC algorithm structure learning, and correlation-based analysis were evaluated regarding structural accuracy, predictive utility, causal consistency, and interpretability (Jiet et al., 2024). To show how the proposed framework

stands against the baseline approaches, Table 3 summarizes the results pertaining to each method's structures, predictive accuracy, consistency, and interpretability.

Table 3. Performance comparison of causal discovery and correlation-based analysis

Method	Structural Hamming Distance	Prediction Accuracy	Causal Consistency	Interpretability
Correlation-based analysis	Not applicable	82.10%	0.41	Limited
Bayesian Network	8	84.70%	0.74	High
PC Algorithm	7	83.90%	0.77	High
Proposed framework	5	86.40%	0.82	Very high

The results indicate that the proposed framework achieved the strongest overall performance across the structural, predictive, consistency, and interpretability criteria evaluated. The framework obtained the lowest structural Hamming distance and thus the highest structural accuracy. The proposed framework also attained the highest scores in predictive accuracy and causal consistency. It is important to note that there was a small margin in the score comparison for predictive accuracy. However, the proposed framework showed a clearer interpretability advantage because it differentiated direct, indirect, and upstream pathways, unlike the other approaches that were able to identify statistical associations.

The correlation-based analysis could identify statistical associations and thus gained a moderate score in predictive accuracy. However, it could not provide an explanation of the causal relationships. It identified risk-related variables, but it was not able to classify those variables as being direct risk factors, indirect risk factors, or upstream determinants. The proposed framework, on the other hand, was able to provide a graph-based interpretation of how different risk-related variables are interconnected.

Discussion

The findings suggest that integrating causal discovery with public health data mining can improve the structural interpretability of risk-factor analysis (Tian et al., 2025). Most conventional data mining methods focus almost exclusively on the prediction aspect. The emphasis of the proposed method on explanation, in addition to the prediction aspect, can provide a significant advantage. This is especially the case in public health because the justification of the prediction is of utmost importance. For example, a model that predicts disease risk without providing interpretable reasoning may have limited utility for intervention planning.

The causal graph indicated that the clinical indicators of systolic blood pressure, glucose levels, and body mass index had direct relationships with the disease risk. The smoking status, physical activity, and diet quality had more of a behavioral risk classification, and their causal roles were more indirect. Smoking status was a direct behavioral risk factor, while physical activity and diet quality mainly influenced disease risk through body mass index and the clinical indicators.

It was observed that the risk of chronic kidney disease in adolescents was affected not only by clinical indicators such as undiagnosed chronic inflammatory foci, anemia, and blood pressure changes but also by lifestyle factors such as smoking, excessive consumption of energy drinks,

and overweight. Following the proposed method, the focus on the first stage of the causal chain was of particular importance (Olimova & Axmedova, 2025). Within the recovered graph, physical activity and diet quality were linked indirectly to disease risk through body mass index and subsequent clinical indicators.

In the case of indirect risk factors, the inclusion of health education and community-level behavioral interventions can be considered, as evidence from a cluster-randomized controlled trial among healthcare assistants shows that a behaviour change intervention improved health literacy and reduced behavioural risk factors of non-communicable diseases. The results indicate that causal discovery allows for the progressive differentiation of response strategies along the risk continuum (Gowrishankar & Stacy-Boddapati, 2025). At the clinical end of direct risk, factors can be managed through a combination of risk screening, monitoring, and medical treatment (Mashood et al., 2026). The proposed framework helps to identify relevant risk factors, but its greatest value is related to the placement of risk factors along the continuum of health intervention, and the potential impact of risk factor modification on health outcomes.

The integration of causal discovery and explainable risk factor identification in public health data mining reflects the unique contribution of this study. The proposed framework is not a substitute for epidemiological inference or domain-expert judgement; rather, it provides a computational approach for generating interpretable hypotheses for subsequent empirical evaluation.

This study has some limitations. First, the study used a controlled synthetic dataset constructed under predefined dependency assumptions. Although this design enables direct evaluation of structural recovery against a known reference, operational public health datasets may contain greater noise, measurement error, unobserved confounding, and missing information (Angeruzzi & Albertini, 2025). Second, the formulation and interpretation of causal relations are based on the assumption of certain sufficiency and evaluation of conditional independence. The discovered causal relations may be incomplete and/or erroneous if some of the controlling variables are omitted. Future research should evaluate the framework using independently sourced public health datasets and incorporate domain-expert validation before policy-oriented interpretation.

The proposed framework has direct implications for explainable health data mining. In several health care and public health instances, decision-makers will be reluctant to trust computational models if the results remain opaque (Olesen & Trabert, 2022). Causal graphs provide relatively easy-to-understand model results (Dayal & Murugesan, 2023). In these graphs, variables are portrayed with their relationships and the direction of their interrelationships. This can help present the results of advanced model outputs in a manner constructively interpretable in the given domain. For researchers, these graphs can be used to generate new hypotheses. For health care practitioners, the graphs can provide the means to communicate the risk pathways with greater clarity. For policymakers, the graphs can provide means to focus on the risk factors with direct interrelationships as well as those that may require intervention to address more indirectly.

Conclusion

This study developed and evaluated a framework combining data mining and causal discovery for explainable risk-factor identification in public health data. The framework includes data preparation, Bayesian Network, PC algorithm, causal structure learning, causal graph interpretation, and evaluation based on structural Hamming distance, prediction accuracy, causal

consistency, and interpretability. The proposed framework was evaluated using a controlled synthetic public health dataset with a predefined reference structure.

The proposed framework was also able to provide a prediction accuracy of 86.40% and a causal consistency score of 0.82, as well as the best performance overall, in the case of the structural Hamming distance. Based on the framework, the most relevant elements in the risk of disease pathways were determined to be: systolic blood pressure, blood glucose level, body mass index, smoking, physical activity, diet, socioeconomic factors, and age. The framework also provided interpretable risk-factor pathways by distinguishing direct risk-related variables, indirect pathways, and upstream determinants.

In the field of explainable data mining, this framework illustrates the use of causal discovery for the better interpretation of risk factors in public health data. Future research should validate the framework using independently sourced public health datasets, investigate potential unmeasured confounding, and assess its generalizability across different disease outcomes and population contexts.

Acknowledgements

The researcher did not receive any funding for this study, and the results have not been published in any other sources. Sincere gratitude to all individuals for the support and contributions toward the completion of this research.

References

Ali, M., Suneeta, V. (2025). Causal structure of higher curvature gravity. Physical Review D. https://doi.org/10.1103/pzdl-crjb

Angeruzzi, A. S., Albertini, M. K. (2025). Longitudinal Synthetic Data Generation from Causal Structures. Anais do XIII Symposium on Knowledge Discovery, Mining and Learning (KDMiLe 2025). https://doi.org/10.5753/kdmile.2025.247519

Chakarverti, M., Sharma, N., Divivedi, R. R. (2019). Prediction Analysis Techniques of Data Mining: A Review. https://doi.org/10.2139/ssrn.3350303

Chen, L., Zhang, Y., Xu, D., Wang, Y., Gao, H. (2023). Causal relationships of neonatal jaundice, direct bilirubin and indirect bilirubin with autism spectrum disorder: A two-sample Mendelian randomization analysis. Frontiers in Public Health. https://doi.org/10.3389/fpubh.2023.1137383

Cicek, G., Isik, O., Ayhan, C., Turgut, A., Deryahanoglu, G. (2025). Effects of physical activity and body mass index on sleep quality and depression among Turkish adults. BMC Public Health. https://doi.org/10.1186/s12889-025-23912-9

Cilloni, L., Dowdy, D. W. (2024). Tuberculosis in US Indigenous Communities: A Need for Public Health Prioritization. American Journal of Public Health. https://doi.org/10.2105/ajph.2023.307544

Dayal, V., Murugesan, A. (2023). Causal Graphs. Demystifying Causal Inference. https://doi.org/10.1007/978-981-99-3905-3_5

Gowrishankar, D., Stacy-Boddapati, A. (2025). The Role of Cultural Risk Factors in Clinical Decision-Making for Pediatric Lead Screening. Screening, prevention, and health promotion. https://doi.org/10.1370/afm.23.s1.8092

Huang, L., Chen, J. (2026). Interpretable machine learning for identifying adolescent obesity risk and identifying key determinants. Frontiers in Public Health. https://doi.org/10.3389/fpubh.2026.1657467

Indrayan, A. (2025). Research Methodology and Biostatistics Series VIII – Statistical Relationship Versus Cause-Effect Relationship Between Medical Factors. Research Methodology and Biostatistics Series VIII – Statistical Relationship Versus Cause-Effect Relationship Between Medical Factors. https://doi.org/10.62830/mmj2-04-32c

Jiet, M. M., Verma, P., Kamble, A., Puri, C. (2024). A Review on Bayesian Methods for Uncertainty Quantification in Machine Learning Models Enhancing Predictive Accuracy and Model Interpretability. 2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI). https://doi.org/10.1109/icoici62503.2024.10696308

Kelley, A. (2022). The System that Supports Recovery. Treatment Program Evaluation. https://doi.org/10.4324/9781003290728-4

Loughman, A., Adler, C., Macpherson, H. (2023). Unlocking Modifiable Risk Factors for Alzheimer's Disease: Does the Oral Microbiome Hold Some of the Keys?. Journal of Alzheimer s Disease. https://doi.org/10.3233/jad-220760

Mao, D., Li, B. (2025). Evaluating the Role of School-Based Physical Activity in Mitigating Cardiometabolic Risk Factors in Children and Adolescents with Overweight or Obesity: A Systematic Review and Meta-Analysis. Children. https://doi.org/10.3390/children12040439

Mashood, I., Samaranayake, D., Kumarapeli, V. (2026). Effectiveness of a behaviour change intervention on health literacy for behavioural risk factors of non-communicable diseases among health care assistants of government hospitals in Colombo District: a cluster-randomized controlled trial. BMC Public Health. https://doi.org/10.1186/s12889-025-26110-9

Nimsarkar, D. A. D. (2024). The Relationship Between Body Mass Index (BMI) and Cardiovascular Disease Risk: A Meta-Analysis. African Journal of Biomedical Research. https://doi.org/10.53555//ajbr.v27i4s.8264

Olesen, S. W., Trabert, E. (2022). Infectious Disease Modeling: Recommendations for Public Health Decision-Makers. Disaster Medicine and Public Health Preparedness. https://doi.org/10.1017/dmp.2022.99

Olimova, G., Axmedova, N. (2025). Clinical assessment of risk factors for the development of chronic kidney disease in adolescents. Scientific and Innovative Therapy. https://doi.org/10.64582/ivit.uz.306

Pasinringi, S. A., Rivai, F., Ismi, A. (2025). Internal Marketing Analysis in the Healthcare Field: Employee Satisfaction and Well-Being as Mediation Variables. South Eastern European Journal of Public Health. https://doi.org/10.70135/seejph.vi.5527

Paulisova, E., Orosova, O., Barbierik, L. (2021). Direct and indirect relationships between disordered eating, sociocultural factors, satisfaction of basic psychological needs and controlled regulation of exercise in university students. Polish Psychological Bulletin. https://doi.org/10.24425/ppb.2021.137258

Pérez-López, F. R., Rodríguez, I., García-Alfaro, P. (2024). The use of the directed acyclic graph to disentangle confounding variables from causal factors in observational studies. Maturitas. https://doi.org/10.1016/j.maturitas.2024.107936

Singh, A., Bhardwaj, P., Garg, P., & Singh, N. (2026). Introduction to explainable artificial intelligence in healthcare. In *Explainable AI in clinical practice* (pp. 23–44). Elsevier. https://doi.org/10.1016/B978-0-443-44111-0.00003-2

Suzuki, J. (2026). Score-Based Structure Learning. Graphical Models and Causal Discovery with Python. https://doi.org/10.1007/978-981-95-5308-2_8

Teh, H. S., Woon, Y. L. (2021). Burden of cancers attributable to modifiable risk factors in Malaysia. BMC Public Health. https://doi.org/10.1186/s12889-021-10412-9

Tian, X., Bai, J., Gong, X., Lei, R., Yan, Y. (2025). Leveraging causal discovery and causal inference to improve interpretability of Charpy impact toughness. Computational Materials Science. https://doi.org/10.1016/j.commatsci.2025.114164

Wen, L., Koo, K. (2025). Public health factors and social determinants of health in urological research. Translational Urology. https://doi.org/10.1016/b978-0-323-90186-4.00076-6

Wilcox, Meredith (2026). Causal Discovery in Big Data Using an Adapted Bayesian Local Causal Discovery Algorithm Based on the Identification of Y-Structures. https://doi.org/10.25148/fiuir.31651

You, M., Pan, X., Zhu, C. (2025). Risk identification and assessment of Internet public opinion on public emergencies based on Bayesian network and association rule mining. Frontiers in Public Health. https://doi.org/10.3389/fpubh.2025.1642960

Zhang, W., Li, Y., Wang, Y. (2025). Data debiasing via causal diffusion model. Data Mining and Knowledge Discovery. https://doi.org/10.1007/s10618-025-01111-9