

A Causally Informed Bayesian Network Framework for Physical Fitness Risk Assessment Using Propensity Score Matching–Guided Structural Priors

Yuanhang Zhou¹, Lulu Geng¹, Jinjie Chen¹, Jiaoyang Zhang¹, Sheng Guo¹, Fei Wang²,
X. San Liang^{2,3}, Tao Zhang^{1,2*}

¹Mental Health Education Center and School of Science, Xihua University, Chengdu, China

²School of Artificial Intelligence, and the Artificial Intelligence Department, Division of Frontier Research, Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Sun Yat-sen University, Zhuhai 519000, China

³Department of Atmospheric and Oceanic Sciences, Fudan University, Shanghai, China

*Corresponding authors: Tao Zhang zhangtao1698@xhu.edu.cn

Abstract

Background Bayesian networks (BNs) are useful for probabilistic prediction, but conventional structure-learning approaches are mainly association-driven and may overlook causal information. This limitation can reduce model interpretability and compromise the clinical or practical relevance of inferred dependency structures. Here, we proposed a causally informed Bayesian network framework that incorporates propensity score matching (PSM)-guided causal effects as structural priors for health risk assessment.

Methods This study proposed a PSM-guided Bayesian network framework (PSM-BN) that incorporates causal effect estimates as structural priors for BN construction. We analyzed physical fitness data using categorical predictors (e.g., Sit and reach, Situps, Grip force, Age) matched through PSM to estimate average treatment effects on the treated. These causal paths were then utilized to structure a BN. Model performance was assessed via five-fold cross-validation, comparing PSM-BN with traditional structure-learning algorithms.

Results The PSM-BN consistently identified sit-ups, sit and reach, and age as the dominant predictors of physical fitness risk. The proposed model achieved an AUC of 0.888 and an accuracy of 82.3%, exceeding the performance of conventional BN models, whose AUCs ranged from 0.845 to 0.861. Explainability analyses showed that the identified predictors contributed robustly and consistently to model output, supporting the transparency of the proposed framework.

Conclusion By integrating PSM-derived causal priors into BN structure learning, the proposed framework enhances both predictive accuracy and model interpretability. This causally informed modeling strategy may serve as a useful computational approach for biomedical risk assessment and other health-related decision-support applications.

Keywords Causal inference, Risk assessment, Bayesian network, Propensity score matching, Physical fitness

Introduction

Effective risk assessment requires computational models that not only achieve adequate predictive performance but also provide interpretable and practically meaningful reasoning. In health-related decision support, transparency is especially important because model outputs may inform screening, risk stratification, and individualized intervention planning. Although recent machine learning and deep learning approaches have demonstrated strong predictive capability, many operate as black-box systems and offer limited insight into the underlying dependency structure of risk factors [1]. This limitation reduces their practical value in public health and biomedical settings where explainability and causal plausibility are essential [2]. Therefore, there is a need for computational frameworks that combine robust prediction with transparent and structurally interpretable modeling.

Bayesian networks (BNs) provide a principled and interpretable framework by encoding conditional dependencies in a directed acyclic graph (DAG) [3]. Yet, learning BN structures from observational health data is still challenged by confounding, spurious associations, and ambiguity in causal direction. For example, Ordovas, et al. [4] applied BN to cardiovascular risk prediction, but the resulting network required expert review to remove implausible connections. Zhao, et al. [5] used hierarchical BNs to enhance laboratory safety assessments, achieving better interpretability through domain knowledge but reducing objectivity. Moreover, many health systems exhibit nonlinear interactions and dynamic feedback—such as threshold effects or behavioral adaptation—that cannot be captured by purely correlation-based methods [6]. This highlights the need for a more robust understanding of complex causal relationships when constructing BNs. Yet commonly used causal inference approaches—including instrumental variables, difference-in-differences, and structural equation modeling—often face limitations such as high computational complexity, unstable estimates, and heavy reliance on correlation structures [7-9]. These issues restrict their applicability and generalizability in real-world causal analysis. As a result, practitioners may either accept questionable causal directions or impose ad hoc expert constraints, sacrificing validity or objectivity. Consequently, there is a pressing need in public health and healthcare for analytical models that are more generalizable, interpretable, and capable of delivering reliable performance.

To address these challenges, this study proposes a framework that integrates Propensity Score Matching (PSM) into BN structure learning to enhance causal interpretability and mitigate confounding bias. PSM is a widely used causal inference method for observational studies and works by estimating the probability of receiving a treatment based on observed covariates. Through matching individuals with similar propensity scores, PSM balances confounders and improves the credibility of causal effect estimation [10, 11]. In health research, PSM is commonly applied to evaluate intervention impacts and uncover causal relationships between lifestyle factors and health outcomes.

We apply this framework to a large-scale physical fitness dataset from the Korean Sports Promotion Foundation (KSPF), which includes 13,393 participants, 11 physical indicators, and health status

graded from A to D. By integrating PSM into BN structure learning, we guide the discovery process toward causally consistent structures. Balancing confounders prior to structure estimation provides credible causal priors, improving identification of true causal pathways among fitness indicators and health outcomes. Compared with conventional data-driven BN learning, our approach reduces reliance on expert-specified structures while enhancing interpretability and validity. Additionally, we incorporate explainable machine learning techniques such as SHAP (SHapley Additive exPlanations) values and feature importance [12] to support and communicate identified risk factors. While these techniques traditionally emphasize statistical associations, integrating PSM-based causal estimates strengthens their interpretability and practical relevance, enabling more reliable identification of key drivers and improving trust in model predictions among stakeholders.

Materials and methods

Propensity score matching (PSM)

PSM is a causal inference method widely used in observational studies to reduce selection bias arising from non-random treatment assignment [13]. The approach estimates the probability of receiving the treatment for each individual—known as the propensity score—which is typically obtained using logistic regression or machine-learning models [14]. Individuals are then matched based on these scores so that the distribution of observed covariates becomes comparable between treatment and control groups. After matching, the treatment effect is estimated from outcome differences in the matched sample, thereby reducing confounding and improving causal validity [15].

Average Treatment Effect on the Treated (ATET)

The Average Treatment Effect on the Treated (ATET) quantifies the causal impact of treatment on individuals who actually received it [16]. It is defined as:

$$ATET = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1] \quad (1)$$

where $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes of the i -th individual under treatment and control, respectively, and D_i is a binary treatment indicator that equals 1 if individual i receives the treatment and 0 if individual i is in the control condition. In the matched sample, the ATET is estimated as the mean outcome difference between each treated individual ($D_i = 1$) and its matched control counterpart ($D_i = 0$) [17, 18]. In this study, the “treatment” corresponds to a specific physical activity–related exposure or fitness level (e.g., a given Situps or BMI category), and the ATET captures the causal effect of these exposures on the physical fitness outcome.

Covariate balance after matching was assessed using standardized mean differences (SMDs) and Chi-square tests. Absolute SMDs less than 0.1 were regarded as indicating adequate balance, whereas values between 0.1 and 0.2 were considered marginally acceptable. Chi-square tests were used to evaluate differences in categorical distributions at a significance level of $\alpha = 0.05$ [14, 19]. These

diagnostics ensure that subsequent causal estimates reflect treatment effects rather than baseline imbalances.

Sensitivity analysis

To evaluate the stability of the ATET estimates, we applied two complementary sensitivity analyses. Bootstrap confidence intervals quantify the sampling variability of the ATET estimates derived from the matched sample, whereas Rosenbaum bounds assess how strong unmeasured confounding would need to be to overturn statistical significance. Together, these approaches assess robustness to both random error and hidden bias, thereby increasing confidence in the resulting causal inferences.

Bootstrap Confidence Interval Analysis

Bootstrap is a non-parametric resampling method used to evaluate the sampling uncertainty of an estimator when the population distribution is unknown. In this study, we used bootstrap to construct 95% confidence intervals for the ATET and to assess the stability of the causal estimates obtained from PSM. After matching, let the sample contain n_M treated individuals and n_C matched controls. For each treated individual $i=1, \dots, n_M$, under the 1:1 nearest-neighbour matching used in this study, we have $n_M=n_C = n$. The point estimate of ATET is:

$$\tilde{\tau}_{ATET} = \frac{1}{n} \sum_{i=1}^n (T_i - C_i) \quad (2)$$

where T_i denotes the observed outcome of the i -th individual in the treated group (T , treated), and C_i denotes the observed outcome of the matched control corresponding to the i -th treated individual (C , control) [20]. For the bootstrap procedure, we repeatedly drew n matched pairs with replacement from the empirical distribution formed by the matched sample, and an ATET estimate was calculated for each resampled dataset. Repeating this process 1000 times produced the empirical sampling distribution $\{\tilde{\tau}_{ATET}^{(1)}, \dots, \tilde{\tau}_{ATET}^{(1000)}\}$. The 95% bootstrap confidence interval was then defined by the 2.5th and 97.5th percentiles of this distribution. When the resulting interval excluded zero, the estimated ATET was regarded as statistically significant [21, 22]. This pair bootstrap approach, while simplified, is widely accepted in applied research when combined with rigorous balance checks and sensitivity analysis. However, bootstrap methods cannot account for unmeasured confounding [23], which is addressed using Rosenbaum bounds.

Rosenbaum Bounds sensitivity analysis

To evaluate whether the estimated ATET might be influenced by unobserved confounding, Rosenbaum bounds sensitivity analysis was conducted. Because the physical fitness dataset is observational, individuals may differ in unmeasured factors—such as lifestyle, dietary habits, or exercise routines—that could jointly affect treatment assignment and outcomes. Although PSM constructs matched pairs with nearly identical observed covariates X_i , a hidden covariate Z_i may still cause residual differences in treatment probabilities. For each individual k in the matched sample, let

D_k be a binary treatment indicator, equal to 1 if the individual belongs to the treatment group and 0 if the individual belongs to the corresponding control group. For example, in the Situps analysis, $D_k = 1$ may indicate that the individual is classified in a higher Situps performance category, whereas $D_k = 0$ indicates membership in the reference Situps category. Rosenbaum bounds sensitivity analysis characterizes the possible impact of unobserved confounding using a sensitivity parameter Γ , which bounds the maximum difference in treatment probabilities within a matched pair attributable to hidden bias. Specifically, for any matched pair (i, j) with the same observed covariates $X_i = X_j$, the treatment probabilities are assumed to satisfy:

$$\frac{1}{\Gamma} \leq \frac{P(D_j=1|X_j, Z_j)}{P(D_i=1|X_i, Z_i)} \leq \Gamma \quad (3)$$

here, $P(D_i = 1|X_i, Z_i)$ denotes the probability that individual i receives the treatment given their observed covariates X_i and unobserved factor Z_i , and $(D_j = 1|X_j, Z_j)$ is defined analogously for individual j [24]. The parameter $\Gamma \geq 1$ quantifies the potential magnitude of hidden bias. When $\Gamma = 1$, matched individuals have identical treatment odds after conditioning on X_i . When $\Gamma > 1$, their odds of receiving the treatment may differ by up to a factor of Γ , implying possible unobserved confounding [25]. Thus, Equation (3) reflects the maximum allowable distortion in treatment assignment due to unmeasured variables and serves as the basis for testing the robustness of the ATET estimate. Rosenbaum further evaluates the impact of such bias on statistical significance using the Wilcoxon signed-rank test statistic $T_{Wilcoxon}$, computed from within-pair outcome differences. For a given level of hidden bias Γ , the method provides two bounds on the p-value:

$$p_{max}(\Gamma) = P(T_{Wilcoxon} \geq t_{obs} | \Gamma) \quad (4)$$

$$p_{min}(\Gamma) = P(T_{Wilcoxon} \geq t_{obs} | \frac{1}{\Gamma}) \quad (5)$$

where $T_{Wilcoxon}$ is the Wilcoxon signed-rank test statistic, t_{obs} denotes its observed value computed from the matched pairs, $p_{max}(\Gamma)$ represents the worst-case scenario where hidden bias works against the detected effect, and $p_{min}(\Gamma)$ represents the best-case scenario where hidden bias works with the effect [26, 27]. If $p_{max}(\Gamma) < 0.05$, then even under the most unfavorable hidden-bias scenario of size Γ , the treatment effect remains statistically significant.

In this study, the ATET estimate was considered robust when two conditions were met: (1) the bootstrap confidence interval excluded zero, and (2) the Rosenbaum upper-bound p-value $p_{max}(\Gamma)$ remained below 0.05. Our results showed that the ATET remained robust up to $\Gamma = 2.0$, indicating that an unmeasured confounder would need to double the odds of treatment assignment to overturn the causal conclusion—an effect size typically regarded as substantively large in observational studies.

Bayesian Networks

BNs are probabilistic graphical models that represent conditional dependencies among variables

through a directed acyclic graph [28]. Each node corresponds to a random variable, and directed edges encode parent–child relationships. The quantitative component of a BN is characterized by conditional probability tables (CPTs) [29], which define the probability distribution of each node given the states of its parents [30]. Let the set of variables be $X = \{X_1, X_2, \dots, X_n\}$, where each X_i denotes a random variable. The BN structure is represented by a DAG $G = (V, E)$, where $V = X$ is the set of nodes and E is the set of directed edges encoding parent–child relationships. Based on the local Markov property, each variable X_i is conditionally independent of all its non-descendants given its parent set $Pa(X_i)$. Under this assumption, the joint probability distribution factorizes as:

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P(X_i | Pa(X_i)) \quad (6)$$

where $Pa(X_i)$ denotes the set of parent nodes of X_i , and the product iterates over all variables in the network. The quantitative component of the BN is defined through CPTs, which specify the probability distribution of each variable given the values of its parents. Let $Val(X_i)$ denote the set of all possible values of X_i , and $Val(Pa(X_i))$ denote all possible value combinations of the parent variables $Pa(X_i)$. For a given variable X_i , a CPT entry takes the general form:

$$P(X_i = x_i | Pa(X_i) = pa_i), \forall x_i \in Val(X_i), pa_i \in Val(Pa(X_i)) \quad (7)$$

In Equation (7), x_i denotes one possible value of variable X_i , and pa_i specifies a particular configuration of parent-node values. The complete CPT of node X_i is obtained by enumerating Equation (7) over all possible pairs (x_i, pa_i) . Together, Equations (6) and (7) formalize how the BN represents a high-dimensional joint distribution through a structured set of local conditional probability relations, enabling probabilistic inference and prediction [31, 32].

PSM-guided Bayesian Networks (PSM-BN)

In this study, we constructed a Bayesian network guided by propensity score matching to incorporate empirically validated causal relationships into structure learning. The approach begins by estimating multilevel causal effects for each physical fitness–related attribute, where continuous predictors were categorized into ordered intervals and treated as discrete treatment levels. For each treatment contrast, propensity scores were estimated using machine-learning classifiers capable of modeling nonlinear covariate structures, followed by 1:1 nearest-neighbor matching with a 0.2 standard deviation caliper to ensure overlap. Covariate balance was examined using standardized mean differences and χ^2 tests, and only matched pairs within the common support region were retained. The ATET was then computed within each stratum, and statistical robustness was established through 1000 bootstrap replications and Rosenbaum sensitivity analysis. An ATET was considered credible only when the bootstrap confidence interval excluded zero and the causal conclusion remained robust under moderate levels of hidden bias. In addition, we have provided a detailed flowchart in the supplementary materials illustrating the complete process—from data preparation to matched control selection, covariate balance assessment, and final ATET estimation (Fig. S1).

These validated causal effects served as structural priors for the Bayesian network. Directed edges

were introduced when the corresponding ATETs indicated a substantial and statistically reliable causal influence. Feature–outcome relationships were incorporated when absolute ATET values exceeded 0.2, whereas feature–feature relationships required stronger evidence ($|\text{ATET}| > 0.4$) to avoid unnecessary structural density. When bidirectional effects were observed across treatment levels, the direction with the larger absolute causal effect was selected. Because each pair of variables could produce multiple stratum-specific ATETs, an edge was included only if at least one level satisfied all robustness criteria, thus ensuring that the final network reflected stable causal dependencies rather than incidental associations.

The resulting causal graph formed the structural basis of the Bayesian network implemented in GeNIe (<https://www.bayesfusion.com/genie/>), where conditional probability tables were estimated from observed frequencies of each category [33]. Compared with networks derived from traditional structure-learning algorithms—such as Peter–Clark (PC), Greedy Thick Thinning, Bayesian Search, and Naive Bayes—the PSM-BN avoided implausible dependency patterns and generated a markedly simpler and more interpretable directed acyclic graph. By grounding structure learning in robust causal inference rather than correlations alone, the PSM-BN provides a principled integration of causal estimation and probabilistic modeling, enhancing both the interpretability and predictive validity of the Bayesian network.

Experimental data

This study utilized a publicly available dataset from the official Kaggle platform. We obtained the filtered dataset from the website at <https://www.kaggle.com/datasets/kukuroo3/body-performance-data>. The dataset had been prefiltered and systematically organized by the provider, containing 13,393 independent samples. Each sample includes 11 features and a categorical label representing the individual's physical fitness level, classified into four ordered categories: A (very good), B (good), C (poor), and D (very poor). Detailed descriptions of all variables are presented in Table 1. After removing one duplicated entry, the final dataset consisted of 13,392 unique samples. Since the gender and physical fitness class variables were originally categorical, appropriate encoding was applied before analysis. For the gender variable, females (F) were encoded as 1 and males (M) as 2. For the physical fitness class, grades A and B were encoded as 0 (indicating good physical fitness), and grades C and D were encoded as 1 (indicating poor physical fitness) [34].

Preprocessing

In this study, the Local Outlier Factor (LOF) algorithm was implemented using the *LocalOutlierFactor* class in *scikit-learn*, with the number of neighbors set to 20 and the contamination parameter set to 'auto'. A total of 671 outliers were detected. After removing these samples, the dataset size decreased from 13,392 to 12,721. This step reduced the potential influence of anomalous observations and improved the robustness of subsequent model construction and causal inference

analyses. Detailed mathematical definitions and formulas of the LOF method are provided in the Supplementary Materials.

Table 1 Description of the meaning of the variables in the dataset

Characteristic	Total (N=12721)	A (n=3244)	B (n=3308)	C (n=3291)	D (n=2878)	A/B (n=6552)	C/D (n=6169)
Gender - F	4449	1411 (43.5%)	1152 (34.8%)	1065 (32.4%)	821 (28.5%)	2563 (39.1%)	1886 (30.6%)
Gender - M	8272	1833 (56.5%)	2156 (65.2%)	2226 (67.6%)	2057 (71.5%)	3989 (60.9%)	4283 (69.4%)
Age (years)	36.2 ± 13.3	34.9 ± 12.8	36.9 ± 13.6	36.4 ± 13.6	36.7 ± 13.2	35.9 ± 13.2	36.5 ± 13.4
Height (cm)	168.9 ± 8.0	168.0 ± 7.6	168.7 ± 8.0	169.4 ± 8.3	169.7 ± 8.2	168.4 ± 7.8	169.5 ± 8.3
Weight (kg)	67.5 ± 11.5	64.6 ± 10.3	66.7 ± 10.8	66.9 ± 10.7	72.2 ± 12.8	65.6 ± 10.6	69.4 ± 12.0
Body fat percentage (%)	22.8 ± 6.8	20.5 ± 6.4	21.9 ± 6.6	22.5 ± 6.1	26.6 ± 6.6	21.2 ± 6.5	24.5 ± 6.7
Diastolic(mmHg)	78.9 ± 10.5	77.9 ± 10.3	78.6 ± 10.6	78.7 ± 10.5	80.4 ± 10.4	78.3 ± 10.5	79.5 ± 10.5
Systolic(mmHg)	130.2 ± 14.3	129.2 ± 14.4	130.6 ± 14.3	129.9 ± 14.3	131.2 ± 14.1	129.9 ± 14.4	130.5 ± 14.2
Grip force(kg)	37.4 ± 10.3	38.8 ± 10.7	38.0 ± 10.3	36.8 ± 10.1	35.9 ± 9.9	38.4 ± 10.5	36.4 ± 10.0
Sit and reach(cm)	15.4 ± 7.9	21.3 ± 5.1	17.4 ± 4.8	14.3 ± 5.9	7.9 ± 9.0	19.3 ± 5.4	11.3 ± 8.2
Situps(counts)	40.8 ± 13.3	48.1 ± 10.5	42.8 ± 11.6	39.0 ± 12.5	32.3 ± 13.5	45.4 ± 11.4	35.9 ± 13.4
Standing broad jump(cm)	192.9 ± 37.2	203.5 ± 35.0	195.9 ± 35.7	189.8 ± 38.2	181.0 ± 36.2	199.7 ± 35.5	185.7 ± 37.6

Categorical Encoding of Continuous Variables

To estimate causal effects across different ranges of physical and physiological measurements, continuous feature variables were categorized into discrete intervals prior to PSM analysis [35]. Several basic indicators were combined to generate clinically meaningful measures: height and weight were used to compute the Body Mass Index (BMI), and systolic and diastolic blood pressure were integrated to derive the Mean Arterial Pressure (MAP). Physical fitness-related indicators were further divided into four or five ordinal levels based on the National Physical Fitness Standards or relevant clinical guidelines [36-38]. In addition, to account for physiological differences between males and females, certain variables were categorized according to gender-specific reference values. The corresponding interval definitions for males (M) and females (F) are summarized in Table 2.

Table 2 Division intervals for each characteristic variable

Features		Interval partitioning and corresponding category encoding					
		1	2	3	4	5	6
Age	M&F	<30 (47.26%)	[30,40) (19.09%)	[40,50) (13.8%)	[50,60] (13.75%)	≥60 (6.11%)	
BMI	M&F	<18.5 (2.37%)	[18.5,24) (55.99%)	[24,27.9) (35.58%)	≥27.9 (6.06%)		
Gender	M&F	Females (34.97%)	Males (65.03%)				

Body fat	M	<14 (15.21%)	[14,20) (35.77%)	[20,25) (30.69%)	≥25 (18.33%)		
percentage (%)	F	<18 (3.71%)	[18,25) (27.44%)	[25,30) (32.39%)	≥30 (36.46%)		
MAP	M&F	<70 (0.46%)	[70,80) (7.36%)	[80,90) (22%)	[90,100) (34.89%)	[100,110) (24.44%)	≥110 (10.85%)
Grip force	M	<40 (31.24%)	[40,50) (51.76%)	[50,60) (15.93%)	≥60 (1.06%)		
(kg)	F	<20 (7.31%)	[20,25) (34.59%)	[25,30) (39.78%)	≥30 (18.32%)		
Sit and reach	M&F	<10 (21.22%)	[10,15) (21.55%)	[15,20) (28.56%)	[20,25) (20.66%)	≥25 (8%)	
(cm)							
Situps	M	<25 (5.27%)	[25,35) (15.64%)	[35,50) (43.45%)	[50,60) (28.76%)	≥60 (6.88%)	
(count)	F	<15 (10.16%)	[15,25) (20%)	[25,35) (26.34%)	[35,45) (27.17%)	≥45 (16.32%)	
Standing	M	<200 (32.35%)	[200,230) (41.65%)	[230,260) (23.67%)	≥260 (2.33%)		
broad jump							
(cm)	F	<150 (37.4%)	[150,180) (45.09%)	[180,200) (14.43%)	≥200 (3.08%)		

Interpretable Machine Learning

To evaluate the consistency between the causal effects identified through PSM and the predictive relevance of variables in machine learning models, four widely used classifiers—Random Forest, XGBoost, Support Vector Machine (SVM), and LightGBM—were trained to predict individuals with poor physical fitness. Two complementary interpretation methods were applied: SHAP and model-based feature importance. SHAP provides a unified, game-theoretic framework to quantify the contribution of each feature to individual-level predictions [39], whereas model-based feature importance summarizes the overall influence of each variable on predictive performance. For SHAP analysis, standard Python libraries (including pandas, scikit-learn, xgboost, lightgbm, matplotlib, and shap) were used for data preprocessing, model training, and SHAP value computation and visualization [40]. Model-based feature importance was obtained directly from the corresponding model implementations in scikit-learn, xgboost, and lightgbm. For the SVM model, the absolute values of the linear-kernel coefficient weights were used as feature importance. For the tree-based models (Random Forest, XGBoost, and LightGBM), feature importance scores were extracted using their built-in feature importance functions. To facilitate cross-model comparison, all importance values were normalized to the [0,1] range. Bar plots were generated to illustrate how the importance rankings varied across algorithms, thereby highlighting which physical fitness indicators exert the strongest predictive influence under different machine learning models [12, 41, 42].

Results

Causal Effects of Individual Predictors on Physical Fitness

Table 3 presents the estimated ATET values for each feature at different levels of physical fitness. We also report changes in SMDs before and after matching for selected matched groups to illustrate the covariate balance achieved during PSM. Full SMDs comparisons for all covariates are provided in the supplementary materials (Table S8–S14).

Sit and reach showed a strong positive causal effect on poor physical fitness. Compared with the reference group (Category 3), individuals with the lowest flexibility (Category 1) had an ATET of +0.34 (95% CI: [0.31, 0.37]), and Category 2 vs. 3 yielded an ATET of +0.13 (95% CI: [0.09, 0.17]), indicating a higher probability of poor physical fitness at lower flexibility levels. At higher flexibility levels, the effects were smaller or reversed: Category 4 vs. 3 produced an ATET of –0.06 (95% CI: [–0.08, –0.03]), and Category 5 vs. 3 produced –0.04 (95% CI: [–0.07, 0.00]); the latter comparison was not statistically significant because the confidence interval included zero.

Situps also emerged as an important causal contributor. The lowest performance group (Category 1 vs. 3) had an ATET of +0.34 (95% CI: [0.30, 0.38]), and Category 2 vs. 3 showed a similar effect (ATET = +0.22; 95% CI: [0.18, 0.26]). In contrast, better Situps performance was associated with substantially lower risk: Category 5 vs. 3 had an ATET of –0.25 (95% CI: [–0.29, –0.20]), and Category 4 vs. 3 had an ATET of –0.20 (95% CI: [–0.23, –0.17]), indicating a clear graded relationship between core endurance and physical fitness.

For Age, a distinct but more complex pattern was observed. Relative to the youngest group (<30 years; Category 1), individuals aged 50–59 (Category 4) had an ATET of –0.34 (95% CI: [–0.43, –0.25]), and those aged ≥60 (Category 5) had an ATET of –0.51 (95% CI: [–0.63, –0.40]), suggesting a lower estimated probability of poor physical fitness in older adults within this sample. A moderate effect was observed for the 40–49 age group (Category 3 vs. 1; ATET = –0.19; 95% CI: [–0.23, –0.14]). For the 30–39 group (Category 2 vs. 1), the ATET was –0.10 (95% CI: [–0.13, 0.07]), and this comparison was not statistically significant.

Table 3 Estimation of causal effects of each characteristic variable at different levels and physical fitness, as well as covariate balance and sensitivity analyses

Treatment variant	n_M	n_C	SMDs	χ^2 -tests	CSR	CSR		Sensitivity Analysis	
						coverage	ATET	Bootstrap.	p-value
						(%)		95% CI	($\Gamma = 2.0$)
BMI = 1vs2	221	221	<0.1	> 0.05	[0.00, 0.68]	73.40%	0.04	[-0.05, 0.13]	<0.0001
BMI = 3vs2	3163	3163	<0.1	> 0.05	[0.01, 0.95]	69.90%	0.04	[0.02, 0.07]	<0.0001
				MAP = 0.02					
BMI = 4vs2	379	379	<0.2	Situps = 0.013	[0.00, 0.96]	49.20%	0.11	[0.04, 0.17]	<0.0001
Body fat percentage = 1vs2	1196	1196	<0.1	> 0.05	[0.02, 0.76]	84%	-0.04	[-0.08, -0.01]	<0.0001
Body fat percentage = 3vs2	1583	1583	<0.1	> 0.05	[0.11, 0.96]	39.80%	0.05	[0.01, 0.08]	<0.0001
Body fat percentage = 4vs2	152	152	<0.2	> 0.05	[0.05, 0.99]	4.80%	0.09	[-0.02, 0.20]	<0.0001
MAP = 1vs4	58	58	<0.2	> 0.05	[0.01, 0.12]	98.30%	-0.1	[-0.29, 0.07]	<0.0001
MAP = 2vs4	933	933	<0.1	> 0.05	[0.03, 0.61]	99.70%	0.02	[-0.03, 0.06]	<0.0001
MAP = 3vs4	2794	2794	<0.1	> 0.05	[0.13, 0.71]	99.80%	0.01	[-0.01, 0.04]	<0.0001
MAP = 5vs4	1552	1552	<0.1	> 0.05	[0.15, 0.67]	49.90%	0.04	[-0.00, 0.07]	<0.0001
MAP = 6vs4	1373	1373	<0.1	> 0.05	[0.04, 0.71]	99.50%	0.01	[-0.02, 0.05]	<0.0001
Sit and reach = 1vs3	933	933	<0.1	> 0.05	[0.02, 0.94]	34.60%	0.34	[0.31, 0.37]	<0.0001
Sit and reach = 2vs3	1369	1369	<0.1	> 0.05	[0.15, 0.77]	49.90%	0.13	[0.09, 0.17]	<0.0001
Sit and reach = 4vs3	2610	2610	<0.1	> 0.05	[0.14, 0.76]	99.30%	-0.06	[-0.08, -0.03]	<0.0001
Sit and reach = 5vs3	1001	1001	<0.1	> 0.05	[0.04, 0.79]	98.30%	-0.04	[-0.07, -0.00]	<0.0001
Age = 2vs1	2423	2423	<0.1	> 0.05	[0.06, 0.74]	99.80%	-0.1	[-0.13, 0.07]	<0.0001
Age = 3vs1	777	777	<0.2	> 0.05	[0.02, 0.85]	44.30%	-0.19	[-0.23, -0.14]	<0.0001
Age = 4vs1	171	171	<0.1	> 0.05	[0.04, 0.94]	9.80%	-0.34	[-0.43, -0.25]	0.063
Age = 5vs1	73	73	<0.1	> 0.05	[0.07, 0.92]	9.50%	-0.51	[-0.63, -0.40]	<0.0001
Situps = 1vs3	658	658	<0.1	> 0.05	[0.01, 0.90]	74.10%	0.34	[0.30, 0.38]	<0.0001
Situps = 2vs3	980	980	<0.1	> 0.05	[0.02, 0.89]	44.90%	0.22	[0.18, 0.26]	<0.0001
Situps = 4vs3	1764	1764	<0.1	> 0.05	[0.02, 0.87]	49.10%	-0.2	[-0.23, -0.17]	<0.0001
				BMI = 0.028					
Situps = 5vs3	629	629	<0.1	BFP = 0.047	[0.01, 0.93]	48.60%	-0.25	[-0.29, -0.20]	<0.0001
Grip force = 1vs2	2908	2908	<0.1	> 0.05	[0.05, 0.88]	100%	0.1	[0.08, 0.13]	<0.0001
Grip force = 3vs2	2765	2765	<0.1	> 0.05	[0.04, 0.84]	89.50%	-0.02	[-0.05, -0.00]	<0.0001
Grip force = 4vs2	339	339	<0.1	MAP=0.044	[0.00, 0.93]	37.50%	-0.06	[-0.12, 0.00]	<0.0001
Standing broad jump = 1vs2	635	635	<0.2	Gender = 0.042	[0.05, 0.98]	14.60%	0.04	[-0.01, 0.09]	<0.0001
Standing broad jump = 3vs2	2062	2062	<0.1	> 0.05	[0.01, 0.85]	79.30%	0	[-0.02, 0.03]	<0.0001
Standing broad jump = 4vs2	271	271	<0.2	> 0.05	[0.01, 0.69]	82.10%	0.03	[-0.03, 0.09]	<0.0001

Note: n_M refers to the number of samples in the treatment group, and n_C refers to the number of samples in the control group. "All" indicates that for each group, all covariates have standardized mean differences SMDs less than 0.2 or chi-square test p-values greater than 0.05. CSR represents the region of common support where the propensity scores of the treatment and control groups overlap. CSR

coverage (%) refers to the proportion of samples retained within the common support region. 95% CI refers to the confidence interval obtained through bootstrapping. *P*-value refers to the lower-bound *p*-value of the Rosenbaum sensitivity analysis when the sensitivity parameter Γ is set to 2.0, with values remaining below 0.05. BFP stands for Body Fat Percentage.

Feature Importance and SHAP Analysis for Model Interpretability

Figure 1 presents the SHAP values and feature importance scores. Across all models, Sit and reach, Situps, and Age were consistently identified as the most influential predictors, indicating a strong contribution to physical fitness classification. In contrast, BMI, Body fat percentage, and MAP were ranked as weaker predictors, suggesting a relatively smaller impact on model performance. For comparability, all feature importance values were normalized.

The distributions of SHAP values in the Random Forest, XGBoost, and LightGBM models showed that Sit and reach, Situps, and Age had the largest mean absolute SHAP values and substantial variability across individuals, confirming their dominant predictive role. Although the global feature importance rankings suggest that MAP, BMI, and Body fat percentage have relatively small effects, their SHAP patterns indicate that higher values of these features are associated with an increased predicted risk of poor physical fitness. This is consistent with their multi-level causal effect estimates obtained from the PSM analysis. Overall, the interpretability results support the conclusion that Age, Situps, and Sit and reach are the primary predictive features, whereas the remaining variables play a comparatively minor role.

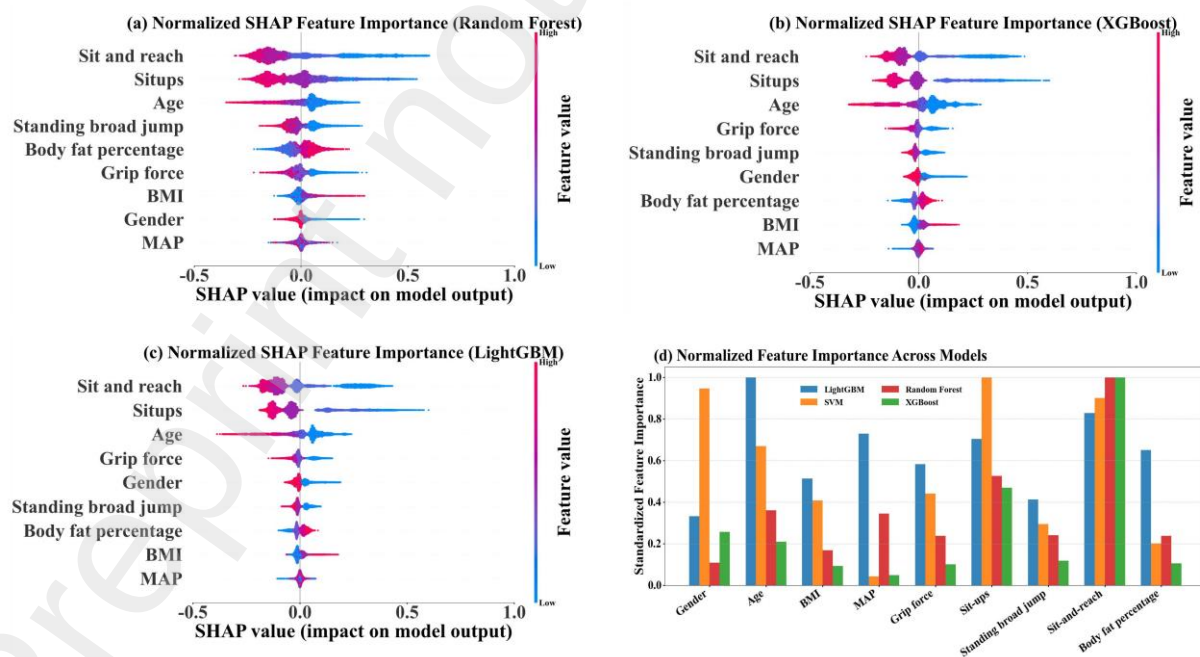


Fig. 1 Feature SHAP values and Feature Importance. (a) shows the feature SHAP values of the Random Forest Model; (b) shows the feature SHAP values of the XGBoost Model; (c) shows the feature SHAP values of the LightGBM model; (d) shows the Feature Importance of all machine learning models. In the SHAP analysis, each feature is represented by points, where distance from zero indicates the feature's impact on the prediction—positive values (right) increase the likelihood of class 1 (poor fitness), and negative values (left) decrease it. Point color reflects feature value (blue = low, red = high). In the feature importance plot, bar height shows each feature's contribution to prediction, but not the direction, which requires SHAP interpretation.

Causal Relationships Among Predictor Variables

To construct the directed edges of the BN, we first estimated the pairwise causal effects among predictor variables using PSM. The procedures for matching, ATET calculation, and sensitivity analysis were identical to those used in the earlier analysis of relationships between predictors and physical fitness outcomes. Table 4 summarizes the causal effects between Situps and the other feature variables. The PSM-derived ATET values indicated strong causal effects of Situps on six variables: Sit and reach, Standing broad jump, Body fat percentage, MAP, Grip force, and BMI.

For example, in the relationship between Situps and Sit and reach, lower Situps performance corresponded to significant negative ATET values (Category 1 vs. 3: ATET = -0.52 , 95% CI [-0.64 , -0.41]; Category 2 vs. 3: ATET = -0.37 , 95% CI [-0.44 , -0.30]). In contrast, higher Situps performance had positive causal effects on Standing broad jump (Category 5 vs. 3: ATET = $+0.45$, 95% CI [0.39 , 0.50]). These robust and directionally consistent ATET estimates directly informed the construction of the PSM-BN.

Directed edges were included only when the causal effects were statistically significant and consistent across levels. In cases of bidirectional causal effects, the direction of the edge was determined based on the stronger ATET magnitude. This process ensures that the PSM-BN captures the underlying causal structure among predictor variables rather than relying solely on statistical associations. The complete set of pairwise causal estimates is provided in the supplementary materials (Tables S1–S7).

Table 4 Causal relationships among feature variables

	n_M	n_C	SMDs	χ^2 -tests	CSR	CSR coverage (%)	ATET	Sensitivity Analysis	
								Bootstrap	p-value
								95% CI	($\Gamma=2.0$)
Situps→Sit and reach									
Situps = 1vs3	873	873	<0.1	> 0.05	[0.01, 0.85]	98.31%	-0.52	[-0.64, -0.41]	<0.0001
Situps = 2vs3	2172	2172	<0.1	> 0.05	[0.03, 0.88]	99.45%	-0.37	[-0.44, -0.30]	<0.0001
Situps = 4vs3	2852	2852	<0.1	> 0.05	[0.04, 0.87]	79.49%	0.37	[0.31, 0.43]	<0.0001
Situps = 5vs3	1150	1150	<0.1	> 0.05	[0.01, 0.94]	88.80%	0.39	[0.30, 0.49]	<0.0001
Situps→Standing broad jump									
Situps = 1vs3	786	786	<0.1	> 0.05	[0.01, 0.91]	88.51%	-0.16	[-0.20, -0.11]	<0.0001
Situps = 2vs3	1631	1631	<0.1	> 0.05	[0.03, 0.87]	74.68%	-0.13	[-0.17, -0.10]	<0.0001
Situps = 4vs3	1960	1960	<0.1	> 0.05	[0.02, 0.88]	54.63%	0.23	[0.19, 0.27]	<0.0001
Situps = 5vs3	1282	1282	<0.1	> 0.05	[0.01, 0.86]	99%	0.45	[0.39, 0.50]	<0.0001
Situps→Body fat percentage									
Situps = 1vs3	702	702	<0.1	> 0.05	[0.01, 0.90]	79.05%	0.34	[0.25, 0.42]	<0.0001
Situps = 2vs3	2172	2172	<0.1	> 0.05	[0.03, 0.88]	99.45%	0.22	[0.17, 0.27]	<0.0001
Situps = 4vs3	3555	3555	<0.1	> 0.05	[0.02, 0.88]	99.08%	-0.12	[-0.17, -0.08]	<0.0001
Situps = 5vs3	1139	1139	<0.1	> 0.05	[0.00, 0.93]	87.95%	-0.14	[-0.21, -0.07]	<0.0001
Situps→MAP									
Situps = 1vs3	880	880	<0.1	> 0.05	[0.01, 0.92]	99.09%	-0.22	[-0.31, -0.12]	<0.0001
Situps = 2vs3	2181	2181	<0.1	> 0.05	[0.02, 0.88]	99.86%	0.02	[-0.05, 0.09]	<0.0001
Situps = 4vs3	3006	3006	<0.1	> 0.05	[0.02, 0.87]	83.78%	0.02	[-0.04, 0.08]	<0.0001
Situps = 5vs3	1277	1277	<0.1	> 0.05	[0.01, 0.95]	98.61%	0.14	[0.05, 0.23]	<0.0001
Situps→Gripforce									
Situps = 1vs3	264	264	<0.1	> 0.05	[0.01, 0.89]	28.73%	-0.12	[-0.25, 0.02]	<0.0001
Situps = 2vs3	2169	2169	<0.1	> 0.05	[0.02, 0.87]	99.31%	-0.02	[-0.07, 0.02]	<0.0001
Situps = 4vs3	2499	2499	<0.1	> 0.05	[0.03, 0.89]	69.64%	0.16	[0.11, 0.21]	<0.0001
Situps = 5vs3	1260	1260	<0.1	> 0.05	[0.01, 0.92]	97.30%	0.24	[0.18, 0.31]	<0.0001
Situps→BMI									
Situps = 1vs3	880	880	<0.1	GF<0.01	[0.01, 0.91]	99.09%	-0.03	[-0.09, 0.03]	<0.0001
Situps = 2vs3	869	869	<0.1	> 0.05	[0.03, 0.88]	40.34%	-0.07	[-0.14, -0.01]	<0.0001
Situps = 4vs3	1949	1949	<0.1	> 0.05	[0.02, 0.87]	54.32%	0.04	[0.00, 0.08]	<0.0001
Situps = 5vs3	621	621	<0.1	> 0.05	[0.01, 0.93]	47.95%	0.07	[0.01, 0.14]	<0.0001

Note: n_M refers to the number of samples in the treatment group, and n_C refers to the number of samples in the control group. "All" indicates that for each group, all covariates have standardized mean differences SMDs less than 0.2 or chi-square test p-values greater than 0.05. CSR represents the region of common support where the propensity scores of the treatment and control groups overlap. CSR coverage (%) refers to the proportion of samples retained within the common support region. 95% CI refers to the confidence interval obtained through bootstrapping. P-value refers to the lower-bound p-

value of the Rosenbaum sensitivity analysis when the sensitivity parameter Γ is set to 2.0, with values remaining below 0.05. GF stands for Grip force.

Comparison of PSM-Guided Bayesian Network Construction and Traditional Structure Learning Algorithms in Performance

Figure 2 presents the BN structures generated using several structure learning algorithms, including the Greedy Thick Thinning algorithm (GTT-BN, Fig. 2a), the Bayesian Search algorithm (BS-BN, Fig. 2b), the PC algorithm (PC-BN, Fig. 2c), the Naive Bayes algorithm (NB-BN, Fig. 2d), and the PSM-guided Bayesian network framework (PSM-BN, Fig. 2e). The networks produced by these algorithms differed substantially in both the total number of edges and the specific directed relationships among variables.

All BN models were evaluated using five-fold cross-validation. As shown in Fig. 2f, the PSM-BN achieved an average AUC of 0.888, exceeding the performance of other models, whose AUC values ranged from 0.845 (PC-BN) to 0.861 (GTT-BN). PSM-BN also attained the highest overall classification accuracy (82.3%), while the remaining models achieved accuracies below 80%.

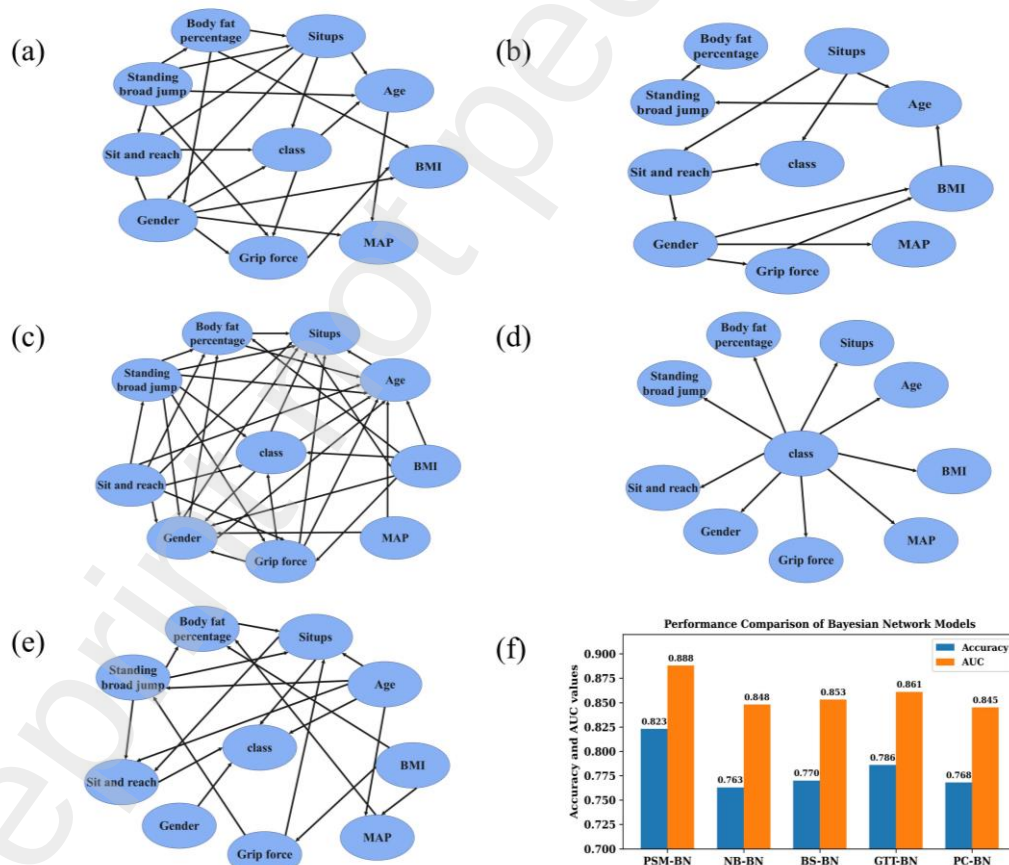


Fig. 2 Construction of different Bayesian networks and comparison of classification performance of the BNs. (a) shows the Bayesian network generated by the Greedy Thick Thinning algorithm (GTT-

BN); (b) shows the Bayesian network generated by the Bayesian Search algorithm (BS-BN); (c) shows the Bayesian network generated by the PC algorithm (PC-BN); (d) shows the Bayesian network generated by the Naive Bayes algorithm (NB-BN); (e) shows the Bayesian network constructed based on the causal effect paths identified by PSM (PSM-BN); (f) shows AUC scores and classification accuracies of different Bayesian networks.

Risk Assessment and Conditional Probability Estimation

We also used the trained Bayesian network to simulate hypothetical health-related scenarios. Specifically, we examined high-risk individuals with poor physical fitness (class = 1) under different levels of key influencing factors. As shown in Table 5, when sit and reach was fixed at a moderate level (Category 3), several combinations of Situps, Age, and Gender resulted in a predicted probability of poor physical fitness exceeding 90%. In total, 13 such high-risk combinations were identified. Notably, among individuals younger than 40 years, even average performance in both Situps and Sit and reach was associated with a high likelihood of poor physical fitness. Moreover, none of the high-risk combinations included Situps at Category 5 (excellent), indicating that excellent Situps substantially reduce the probability of poor physical fitness.

Table 5 Given a fixed Sit and reach level, we identify combinations of Gender, Age, and Situps performance under which the probability of poor physical fitness exceeds 90%

Situps(count)	Age(years)	Gender	Worst physical health
<15	< 30	Female	95.8%
<25	< 30	Male	92.8%
<15	[30,40)	Female	98.9%
<25	[30,40)	Male	97.2%
[15,25)	< 30	Female	92.9%
[25,35)	< 30	Male	90%
[15,25)	[30,40)	Female	97.5%
[25,35)	[30,40)	Male	97.7%
[25,35)	< 30	Female	97.4%
[35,50)	< 30	Male	97.4%
[35,50)	[30,40)	Male	95.2%
[35,45)	< 30	Female	91.7%
[50,60)	< 30	Male	95.7%

Discussion

To our knowledge, this study is the first to construct a BN using a systematic PSM-based causal framework without relying on expert-specified structures. It analyzes multilevel causal effects among variables, revealing complex and potentially bidirectional relationships. Based on these causal links, we built a BN for risk assessment and prediction in public health, which outperformed traditional structure-learning algorithms. Integrating machine learning interpretability enhances model transparency and credibility. Our PSM-based ATET estimation additionally applies categorical encoding to continuous variables by grouping them into intervals, enhancing covariate balance and offering a novel framework for causal inference.

Methodological considerations

Compared with traditional BN structure learning algorithms, the PSM-BN framework offers clear methodological advantages by grounding edge selection in causal reasoning rather than in purely correlation-driven procedures [43]. As illustrated in Figure 2, traditional algorithms frequently produced implausible or reversed dependencies—such as edges from the outcome variable to its predictors—or yielded overly complex graphs that undermine interpretability and require ad hoc expert correction. These issues introduce subjectivity and weaken the credibility of BN-based risk assessment. In contrast, the PSM-BN approach combines multi-level PSM with machine learning interpretability techniques to guide network construction based on causal relationships, reducing dependence on purely correlation-based algorithms and expert experience, while deeply integrating causal inference with probabilistic prediction. This framework unifies risk factor identification, feature selection, causal analysis, probabilistic prediction, and individualized risk assessment, providing a novel analytical system for public health and health research.

The proposed framework offers three major methodological advantages. First, by using robust PSM-derived causal effects to guide network construction, the approach minimizes correlation-driven bias and avoids implausible or overly complex structures, substantially improving interpretability without requiring expert-imposed constraints. Second, integrating SHAP-based interpretability with PSM strengthens the reliability of identified risk factors: SHAP provides efficient, model-agnostic feature insights, while PSM supplies explicit causal direction and effect magnitudes, together mitigating the “black box” limitation of conventional machine-learning models [44]. Third, the multilevel estimation of ATETs—enabled by categorizing continuous variables and applying stepwise matching—supports granular causal assessment and improves covariate balance. This design enhances the precision of risk evaluation and ensures broad applicability across diverse data types and domains.

Causal Effect Evaluation Based on Machine Learning Interpretability and PSM

Table 3 summarizes the multilevel ATET estimates together with balance diagnostics and sensitivity analyses, confirming that the causal effects identified by PSM are both statistically robust and

substantively meaningful. These estimates directly informed the structure of the PSM-BN: features with larger ATET magnitudes were linked to the outcome, and feature–feature edges were introduced only when a higher causal threshold was met. Most comparisons achieved excellent covariate balance (SMDs < 0.1; χ^2 , $p > 0.05$), and Rosenbaum bounds further supported the stability of the causal conclusions. As a result, the constructed network consistently highlighted Sit and reach, Situps, and Age as dominant predictors and revealed clear dose–response patterns, where stronger causal effects corresponded to more influential graph connections.

Using both PSM and machine learning interpretability methods (e.g., SHAP values), we consistently identified that Situps, Age, and Sit and reach are the primary causal drivers of physical health, which aligns closely with previous research in exercise science. For example, Araújo, et al. [45] conducted a large-scale prospective cohort study in a middle-aged population and found a significant inverse relationship between body flexibility and the risk of natural death, indicating that flexibility indicators, such as the Sit and reach, are strongly associated with overall health. Dong, et al. [46] included eight randomized controlled trials involving a total of 169 athletes and found that core strength (e.g., Situps) has a positive impact on overall athletic performance. Thus, our research uniquely integrates rigorous causal inference methods with machine learning validation, significantly strengthening the causal claims.

In addition, for example, the causal relationship between Situps and physical health (Table 3), the ATET values gradually decrease as the Situps values increase, indicating that higher Situps performance corresponds to better physical health. Comparing this with the SHAP results in Figure 2, we observe that in different models, low Situps values (blue points) are mostly located to the right of the zero line, with positive SHAP values, indicating a higher probability of being predicted as having poorer physical health; whereas high Situps values (red points) are concentrated to the left of the zero line, with negative SHAP values, showing a significantly lower probability of being predicted as physically unhealthy. This result further confirms the complementary and mutually validating relationship between machine learning interpretability techniques and causal inference. Interpretability techniques can visually show the importance and direction of a feature’s impact on model predictions, but they cannot quantify the effect or reveal multi-level changes [47].

Performance Evaluation of PSM-BN

Compared with traditional structure learning algorithms, PSM-BN offers a more concise and interpretable network structure while reducing reliance on expert knowledge. Applied to physical health data, PSM-BN achieved a prediction accuracy of 82.3% and an AUC of 88.8%, outperforming conventional BN and demonstrating excellent predictive performance. These results validate our core idea: using PSM-identified causal effect pathways to guide BN structure learning can significantly enhance model generalizability, interpretability, and overall performance. Moreover, this confirms that our integrated framework—combining causal inference, machine learning interpretability techniques,

and BNs—is both reasonable and effective, providing a novel analytical perspective for risk assessment and probabilistic prediction in public health and health research.

Furthermore, we applied PSM-BN to assess physical health risks, highlighting its practical advantages. Through multivariable combination analysis (Table 5), we identified 13 high-risk feature combinations (probability > 90%). Notably, individuals with excellent core muscle strength (Situps, Category 5) did not appear in any high-risk combination, underscoring its protective role. Conversely, males aged 30–40 still exhibited a high risk even at moderate Situps levels, indicating that this population requires particular attention. Since this study focuses primarily on methodology, the practical risk assessment considered only a single scenario. Given the large sample size, we set the initial probabilities in the CPT based on observed frequencies of each feature category [48]. In cases of limited or uneven sample distributions, Bayesian smoothing can be applied to adjust initial probabilities, thereby improving model stability and generalizability [49, 50].

Limitations

Despite the advantages of this study, several limitations remain. First, the study was applied only to a single physical health dataset, and the dataset contained relatively limited features, without considering important factors such as smoking status, exercise amount, and lifestyle habits [51]. Future research should expand the dataset to validate the generalizability of the method. Second, the use of cross-sectional data limits the inference of temporal causal relationships, and longitudinal data are needed to capture dynamic changes [52]. Finally, when determining the directed edges of the BN using PSM causal effects, there is no standardized threshold for ATET, which involves a certain degree of subjective choice. Future studies should focus on establishing standardized criteria for determining ATET thresholds.

Conclusion

In summary, this study proposed a PSM-BN data analysis framework that integrates propensity score matching, machine learning interpretability analyses, and BNs, and applied it to a physical fitness dataset. The results indicate that Sit-and-reach, Situps performance, and Age are key features with robust causal effects on physical fitness. Furthermore, the multi-level causal effects among variables not only help to understand complex causal relationships but also provide foundational support for constructing a causally informed BN, resulting in predictive accuracy that is markedly superior to traditional modeling approaches. This method not only enhances the predictive capability of BNs but also improves model interpretability and credibility, offering a novel analytical framework for public health and wellness research.

Supplementary Information

This study includes a supplementary material that describes the PSM results among the remaining feature variables.

Acknowledgements

We would like to express our gratitude to all the staff and participants.

Authors' contributions

Yuanhang Zhou: Writing- original draft, Methodology, Validation, Visualization. **Lulu Geng:** Methodology, Validation, Visualization. **Jinjie Chen:** Methodology, Investigation, Visualization. **Jiaoyang Zhang:** Methodology, Validation. **ShengGuo:** Validation. **Jienuan Lai:** Investigation, Visualization. **Fei Wang:** Conceptualization, Writing-review& editing, Funding acquisition. **X. San Liang:** Writing-review& editing, Funding acquisition. **Tao Zhang:** Writing- original draft, Methodology, Writing - review & editing.

Funding

This work was supported by the Project of Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai) (#SML2023SP203).

Data and Code availability statement

The data that support the findings of this study are openly available at <https://www.kaggle.com/datasets/kukuroo3/body-performance-data>. The main function of the code is provided in the code supplementary materials.

Declarations

Consent for publication

All authors have read and agreed to the publication of this manuscript.

Competing interests

The authors declare that they have no conflict of interest.

Ethics approval and consent to participate

The data in this study are from public data, which complies with the CC0 1.0 agreement and no ethical statement is applicable.

References

- [1] Z. Sadeghi *et al.*, "A review of Explainable Artificial Intelligence in healthcare," *Computers and Electrical Engineering*, vol. 118, p. 109370, 2024/08/01/ 2024, doi: <https://doi.org/10.1016/j.compeleceng.2024.109370>.
- [2] I. D. Mienye *et al.*, "A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges," *Informatics in Medicine Unlocked*, vol. 51, p. 101587, 2024/01/01/ 2024, doi: <https://doi.org/10.1016/j.imu.2024.101587>.
- [3] N. K. Kitson, A. C. Constantinou, Z. Guo, Y. Liu, and K. Chobtham, "A survey of Bayesian Network structure learning," *Artificial Intelligence Review*, vol. 56, no. 8, pp. 8721–8814, 2023.
- [4] J. M. Ordovas *et al.*, "A Bayesian network model for predicting cardiovascular risk," *Computer methods and programs in biomedicine*, vol. 231, p. 107405, 2023.
- [5] J. Zhao, H. Cui, G. Wang, J. Zhang, and R. Yang, "Risk assessment of safety level in university laboratories using questionnaire and Bayesian network," *Journal of Loss Prevention in the Process Industries*, vol. 83, p. 105054, 2023.
- [6] C. Glymour, K. Zhang, and P. Spirtes, "Review of causal discovery methods based on graphical models," *Frontiers in genetics*, vol. 10, p. 524, 2019.
- [7] A. Singh, K. Hosanagar, and A. Gandhi, "Machine learning instrument variables for causal inference," in *Proceedings of the 21st ACM Conference on Economics and Computation*, 2020, pp. 835–836.
- [8] E. J. T. Tchetgen, C. Park, and D. B. Richardson, "Universal difference-in-differences for causal inference in epidemiology," *Epidemiology*, vol. 35, no. 1, pp. 16–22, 2024.
- [9] M. J. Vowels, "SLEM: Machine learning for path modeling and causal inference with super learner equation modeling," *arXiv preprint arXiv:2308.04365*, 2023.
- [10] K. Narita, J. d. D. Tena, and C. Detotto, "Causal inference with observational data: A tutorial on propensity score analysis," *The Leadership Quarterly*, vol. 34, no. 3, p. 101678, 2023.
- [11] B. Lee, N.-e. Kim, S. Won, and J. Gim, "Propensity score matching for comparative studies: a tutorial with R and Rex," *Journal of Minimally Invasive Surgery*, vol. 27, no. 2, p. 55, 2024.
- [12] H. Wang, Q. Liang, J. T. Hancock, and T. M. Khoshgoftaar, "Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods," *Journal of Big Data*, vol. 11, no. 1, p. 44, 2024.
- [13] E. A. Stuart, "Matching methods for causal inference: A review and a look forward," *Statistical science: a review journal of the Institute of Mathematical Statistics*, vol. 25, no. 1, p. 1, 2010.
- [14] P. C. Austin, "An introduction to propensity score methods for reducing the effects of confounding in observational studies," *Multivariate behavioral research*, vol. 46, no. 3, pp. 399–424, 2011.
- [15] S. Kanamori and M. Takeuchi, "Treatment effect estimation using the propensity score in clinical trials with historical control," *BMC Medical Research Methodology*, vol. 24, no. 1, p. 47, 2024.
- [16] N. Greifer and E. A. Stuart, "Choosing the causal estimand for propensity score analysis of observational studies," *arXiv preprint arXiv:2106.10577*, 2021.

- [17] P. J. Schulte and E. J. Mascha, "Propensity score methods: theory and practice for anesthesia research," *Anesthesia & Analgesia*, vol. 127, no. 4, pp. 1074–1084, 2018.
- [18] V. S. Harder, E. A. Stuart, and J. C. Anthony, "Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research," *Psychological methods*, vol. 15, no. 3, p. 234, 2010.
- [19] Q.-Y. Zhao, J.-C. Luo, Y. Su, Y.-J. Zhang, G.-W. Tu, and Z. Luo, "Propensity score matching with R: conventional methods and new features," *Annals of translational medicine*, vol. 9, no. 9, p. 812, 2021.
- [20] W. Pan and H. Bai, "Propensity score interval matching: using bootstrap confidence intervals for accommodating estimation errors of propensity scores," (in eng), *BMC Med Res Methodol*, vol. 15, p. 53, Jul 28 2015, doi: 10.1186/s12874-015-0049-3.
- [21] C. Rothe, "Robust confidence intervals for average treatment effects under limited overlap," *Econometrica*, vol. 85, no. 2, pp. 645–660, 2017.
- [22] S. F. Mokhtar, Z. M. Yusof, and H. Sapiri, "Confidence intervals by bootstrapping approach: a significance review," *Malaysian Journal of Fundamental and Applied Sciences*, vol. 19, no. 1, pp. 30–42, 2023.
- [23] G. W. Imbens and D. B. Rubin, *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.
- [24] A. Assima, G. Zanello, and M. Smale, "Articulating fertilizer subsidy effects on women's diet quality by food supply source in Mali," *CABI Agriculture and Bioscience*, vol. 3, no. 1, p. 42, 2022.
- [25] H. S. Twabi, S. O. Manda, and D. S. Small, "Assessing the effects of maternal HIV infection on pregnancy outcomes using cross-sectional data in Malawi," *BMC Public Health*, vol. 20, no. 1, p. 974, 2020.
- [26] S. Heng and D. S. Small, "Sharpening the Rosenbaum sensitivity bounds to address concerns about interactions between observed and unobserved covariates," *Statistica Sinica*, vol. 31, pp. 2331–2353, 2021.
- [27] X. Nie, G. Imbens, and S. Wager, "Covariate balancing sensitivity analysis for extrapolating randomized trials across locations," *arXiv preprint arXiv:2112.04723*, 2021.
- [28] N. Babakov, A. Sivaprasad, E. Reiter, and A. Bugarín-Diz, "Reusability of Bayesian Networks case studies: a survey," *Applied Intelligence*, vol. 55, no. 6, p. 417, 2025.
- [29] H. Li, X. Ren, and Z. Yang, "Data-driven Bayesian network for risk analysis of global maritime accidents," *Reliability Engineering & System Safety*, vol. 230, p. 108938, 2023.
- [30] P. Laitila and K. Virtanen, "Advancing construction of conditional probability tables of Bayesian networks with ranked nodes method," *International Journal of General Systems*, vol. 51, no. 8, pp. 758–790, 2022.
- [31] Y. Yildiz, D. B. A. Gülşen, U. Sevilmiş, and O. B. Öztürk, "Evaluating athletic mental energy analysis: a novel approach using fuzzy-based Bayesian networks," *BMC Sports Science, Medicine and Rehabilitation*, vol. 17, no. 1, p. 283, 2025.
- [32] Y. Li, X. Chen, Y. Wang, J. Hu, Z. Shen, and X. Ding, "Application of group LASSO regression based Bayesian networks in risk factors exploration and disease prediction for acute kidney injury in hospitalized patients with hematologic malignancies," *BMC nephrology*, vol. 21, no. 1, p. 162, 2020.
- [33] J. M. Ordovás *et al.*, "A Bayesian network model for predicting cardiovascular risk," *Computer methods and programs in biomedicine*, vol. 231, p. 107405, 2023.
- [34] N. Kosaraju, S. R. Sankepally, and K. Mallikharjuna Rao, "Categorical data: Need, encoding, selection of encoding method and its emergence in machine learning models—a practical review study on heart disease prediction dataset using pearson correlation," in *Proceedings of International Conference on Data Science and Applications: ICDSA 2022, Volume 1*, 2023: Springer, pp. 369–382.

- [35] N. S. Hejazi, D. Benkeser, I. Diaz, and M. J. van der Laan, "Efficient estimation of modified treatment policy effects based on the generalized propensity score," *arXiv preprint arXiv:2205.05777*, 2022.
- [36] F. Zhang *et al.*, "Physical fitness reference standards for Chinese children and adolescents," *Sci Rep*, vol. 11, no. 1, p. 4991, 2021.
- [37] R. M. Orr, R. Lockie, G. Milligan, C. Lim, and J. Dawes, "Use of physical fitness assessments in tactical populations," *Strength & Conditioning Journal*, vol. 44, no. 2, pp. 106–113, 2022.
- [38] T. A. Miller, *NSCA's Guide to Tests and Assessments*. Human Kinetics, 2012.
- [39] Y. Gebreyesus, D. Dalton, S. Nixon, D. De Chiara, and M. Chinnici, "Machine learning for data center optimizations: feature selection using Shapley additive exPlanation (SHAP)," *Future Internet*, vol. 15, no. 3, p. 88, 2023.
- [40] A. V. Ponce-Bobadilla, V. Schmitt, C. S. Maier, S. Mensing, and S. Stodtmann, "Practical guide to SHAP analysis: explaining supervised machine learning model predictions in drug development," *Clinical and Translational Science*, vol. 17, no. 11, p. e70056, 2024.
- [41] J. T. Hancock, T. M. Khoshgoftaar, and Q. Liang, "A problem-agnostic approach to feature selection and analysis using shap," *Journal of Big Data*, vol. 12, no. 1, p. 12, 2025.
- [42] C. Molnar, G. König, B. Bischl, and G. Casalicchio, "Model-agnostic feature importance and effects with dependent features: a conditional subgroup approach," *Data Mining and Knowledge Discovery*, vol. 38, no. 5, pp. 2903–2941, 2024.
- [43] Y. Lee and E. L. Ogburn, "Network dependence can lead to spurious associations and invalid inference," *Journal of the American Statistical Association*, vol. 116, no. 535, pp. 1060–1074, 2021.
- [44] W. Y. Ng, L. R. Wang, S. Liu, and X. Fan, "Causal SHAP: Feature Attribution with Dependency Awareness through Causal Discovery," *arXiv preprint arXiv:2509.00846*, 2025.
- [45] C. G. S. Araújo *et al.*, "Reduced Body Flexibility Is Associated With Poor Survival in Middle-Aged Men and Women: A Prospective Cohort Study," *Scandinavian journal of medicine & science in sports*, vol. 34, no. 8, p. e14708, 2024.
- [46] K. Dong, T. Yu, and B. Chun, "Effects of core training on sport-specific performance of athletes: a meta-analysis of randomized controlled trials," *Behavioral Sciences*, vol. 13, no. 2, p. 148, 2023.
- [47] X. Huang and J. Marques-Silva, "On the failings of Shapley values for explainability," *International Journal of Approximate Reasoning*, vol. 171, p. 109112, 2024.
- [48] K. Drury, M. J. Barons, and J. Q. Smith, "Structural Refinement of Bayesian Networks for Efficient Model Parameterisation," *arXiv preprint arXiv:2510.00334*, 2025.
- [49] L. Azzimonti, G. Corani, and M. Zaffalon, "Hierarchical estimation of parameters in Bayesian networks," *Computational Statistics & Data Analysis*, vol. 137, pp. 67–91, 2019.
- [50] C. Cooper, G. I. Webb, and D. F. Schmidt, "Efficient Parameter Estimation for Bayesian Network Classifiers using Hierarchical Linear Smoothing," *arXiv preprint arXiv:2505.23320*, 2025.
- [51] Y. Kim, "The effects of smoking, alcohol consumption, obesity, and physical inactivity on healthcare costs: a longitudinal cohort study," *BMC Public Health*, vol. 25, no. 1, p. 873, 2025.
- [52] S. Toh and M. A. Hernán, "Causal inference from longitudinal studies with baseline randomization," *The international journal of biostatistics*, vol. 4, no. 1, 2008.