frontiers | Frontiers in Public Health

TYPE Original Research
PUBLISHED 10 September 2025
DOI 10.3389/fpubh.2025.1647200

OPEN ACCESS

EDITED BY
Hyun Jung Park,
University of Pittsburgh, United States

REVIEWED BY
Debajit Karmakar,
Lakshmibai National Institute of Physical Education, India
Sohom Saha,
Lakshmibai National Institute of Physical Education, India

*CORRESPONDENCE
Tao Xie
✉ xietao834996131@163.com

RECEIVED 15 June 2025
ACCEPTED 22 August 2025
PUBLISHED 10 September 2025

CITATION
Xie T, Hao Y and Xie F (2025) A causal inference method for athletic injuries based on quantile threshold functions and latent Gaussian DAG models.
Front. Public Health 13:1647200.
doi: 10.3389/fpubh.2025.1647200

A causal inference method for athletic injuries based on quantile threshold functions and latent Gaussian DAG models

Tao Xie[1]*, Yaxian Hao[2] and Fen Xie[3]

[1]Department of Sports Science, Kyungil University, Gyeongsan, Republic of Korea, [2]School of Mathematics and Computer Science, Shanxi Normal University, Taiyuan, China, [3]Department of Data Science, City University of Hong Kong, Kowloon, Hong Kong SAR, China

Introduction: Causal inference of athletic injuries provides the critical foundations for the development of effective prevention strategies. In recent years, the directed acyclic graph model (DAG) has established itself as an indispensable tool in the study of athletic injuries.
Methods: This study proposes a quantile threshold function (QTF) and integrates it with the causal inference framework within the latent DAG model for ordinal variables. This process begins by transforming continuous variables into ordinal variables to construct a DAG, which is analyzed using the latent causal inference framework to estimate ordinal causal effects (OCE).
Results: Testing this approach on real-world data showed clear differences between groups ($F > 52{,}000$, $P < 0.05$). The analysis also revealed three direct paths and two indirect paths related to athletic injuries, based on the DAG.
Discussion: We obtained the OCE by intervening on variables that directly or indirectly influence athletic injuries. DAG path analysis further elucidated the impact of causal pathways on the risk of injury. The approach proposed in this study provides novel theoretical and methodological insights into athletic injuries and serves as a crucial basis for optimizing training programs and mitigating injury risk.

KEYWORDS

causal inference, athletic injury, directed acyclic graph, latent graphical model, ordinal causal effects

1 Introduction

Sports science research is fundamentally causal (1, 2), focusing on the mechanisms of physical activity to optimize training and strategy. Identifying determinants of athletic or team success informs performance improvement, while in health research, physical activity interventions improve fitness and well-being. Data in sports science are often highly discrete. This characteristic can be addressed through the use of scientific methods that categorize the data into distinct levels, enabling it to be treated as ordinal variables. Ordinal variables, characterized by categorical values with an inherent ranking order, are prevalent in various research fields (3). Examples include training intensity (A, B, C) and training frequency (low, medium, high). Considering the widespread occurrence of ordinal variables in sports data, investigating methods of causal effect analysis for these variables has considerable practical significance.

Causal inference aims to uncover these underlying relationships. Epidemiological research is a key area of causal inference in sports science (4). Although some advocate causal models in injury prevention (5, 6), practical applications remain limited (7). Early studies introduced graphical causal models (6, 8), but limitations in presentation and scope constrained their impact. (9) emphasized causal reasoning in strength training, although their work focused on specific issues rather than a systematic introduction. The van Mechelen sequence of prevention (10) and the Finch TRIPP framework (11) introduce measures that are likely to reduce the future risk and/or severity of athletic injuries based on causal and mechanistic understandings. However, the development of causal knowledge represents a significant challenge. To estimate a causal effect, researchers must control all major baseline variables that could influence both exposure and outcome (12, 13). However, fulfilling these conditions in real-world settings can be exceptionally challenging. Failure to control a confounding variable can lead to inaccurate conclusions about the causal relationship between variables. Randomized controlled trials (RCTs), which are considered the gold standard for causal inference, were later proposed by researchers but are challenging to implement in sports science (14), particularly in elite sports (15). Therefore, inference of causal relationships often relies on observational studies, which are prone to selection bias. The reliance on such studies, combined with the lack of robust tools and frameworks for causal inference, has hindered the advancement of causal knowledge on sports injuries and the development of effective prevention strategies. Competitive sports training often involves a high risk of sports injuries, not only affecting the overall volume of training, but may also alter training patterns and recovery strategies (16, 17).

To address some of the issues mentioned above, recent efforts have emphasized the adoption of graphical causal models in injury prevention and advocated for greater participation in causal inference research (18). Causal diagrams, including frameworks (19), models (20, 21), causal directed acyclic graphs (DAG) (22, 23), and other types of diagrams (24, 25), serve as valuable tools for organizing ideas, guiding future research, and supporting causal inference efforts. These diagrams, particularly DAG, are of significant importance in statistical analysis. In most practical situations, an appropriate causal diagram is rarely known, so methods that can learn both a network structure and its parameters from data are required. A Bayesian network is the most commonly used method for causal graph problems. When using Bayesian methods for learning, the observed data only determine the DAG describing their joint distribution up to its Markov equivalence class (26). It is crucial that each Markov equivalence class can be uniquely represented by a completed partially directed acyclic graph(CPDAG). The learning of Bayesian networks relies fundamentally on the type of data, with existing approaches focused primarily on continuous and categorical data (27, 28). However, in the field of sports training, the variables involved often include both continuous data and ordinal data. Existing methods do not fully consider the inherent ordinal nature of the data when handling ordinal data. Therefore, specialized methods are needed to calculate the causal effects between ordinal variables while fully accounting for their ordinal characteristics. Luo et al. (29) proposed an Ordinal Structural Expectation-Maximization (OSEM) algorithm based on a latent Gaussian model. This algorithm can construct an appropriate causal graph framework for ordinal data (29), providing support for a subsequent analysis of causal effects.

Realizing the existing gaps in the theoretical and practical aspects of this field, this study proposes a quantile threshold function (QTF) that transforms continuous variables into ordinal variables and ensures the consistency of the classification results while effectively preserving the ordered nature of the data. Based on data transformation, this study applies the method designed by Luo et al. (29) to construct a causal Directed Acyclic Graph (DAG). Then, using the ordinal data causal analysis algorithm proposed by (30), ordinal causal effects (OCE) between ordered variables are calculated within the framework of the latent Gaussian DAG model. These findings offer valuable insights for optimizing rehabilitation strategies and provide the critical foundations for the development of effective prevention strategies. The rest of the article is structured as follows. In Section 2, we present a summary of previous approaches, including both Gaussian DAG models and the do-operator. In Section 3, we present our original contribution with the quantile threshold function and how to evaluate ordinal causal effects by combining the algorithm of OSEM and the Latent Causal Inference Framework. In Section 4, we use real-world data to illustrate the performance of causal effect estimation with latent DAG structures. Lastly, in Section 5, we discuss the potential prospects and limitations of the research and highlight possible directions for expanding the current work.

2 Background

2.1 Gaussian DAG-models

Probabilistic graphical models, which integrate graphical structures into probabilistic inference, are widely used and effective frameworks for studying these complex systems. The concept is to factorize the joint probability distribution p for the variables $\mathbf{X} = (X_1, \cdots, X_m)^\top$ concerning a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of vertices representing the variables and $\mathcal{E}$ the set of edges encoding the independence relationships (31, 32). Bayesian networks are a special class of probabilistic graphical models, where $\mathcal{G}$ is a directed acyclic graph (DAG) or named DAG models (33, 34). The joint probability distribution p can be specified by a set of parameters θ and factorizes based on $\mathcal{G}$ as:

$$p(\mathbf{x} \mid \theta, \mathcal{G}) = p\left(x_1, \ldots, x_m \mid \theta, \mathcal{G}\right) = \prod_{i=1}^{m} p\left(x_i \mid \mathbf{x}_{pa(i)}, \theta_i, \mathcal{G}\right). \quad (1)$$

Where $\theta = \cup_{i=1}^{m} \theta_i$, $\mathbf{x}$ is a realization of $\mathbf{X}$, and we assume that the subsets $\{\theta_i\}_{i=1}^{m}$ are disjoint. Denote the parents of node i by $pa(i)$), where there is a directed edge from j to i if $j \in pa(i)$. Thus, Equation (1) can also be interpreted as stating that a variable x_i is conditionally independent of its non-descendants, given its parents $\mathbf{X}_{pa(i)}$ in $\mathcal{G}$. This is the Markov property (31).$\mathcal{B} = (\mathcal{G}, \theta)$ denotes a Bayesian network. Given a data sample $\mathcal{X}$, learning a Bayesian network, therefore, involves estimating both the network

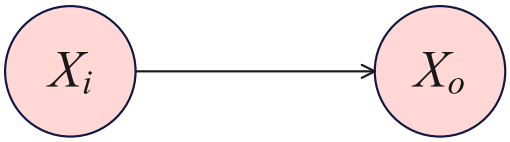

FIGURE 1
The outcome variable X_o resulting from a deterministic intervention on the intervention variable X_i.

structure $\mathcal{G}$ and θ. If the joint distribution of $\boldsymbol{X}$ is a Gaussian distribution, then

$$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{2}$$

In the case of Gaussian data, to address the uncertainty regarding the graphical structure, Maathuis et al. (35) provided lower bounds for causal effects after identifying a Markov equivalence class that is consistent with the data. By using a Bayesian approach, one can combine structure learning and effect estimation into a process that produces the posterior distribution of causal effects. A significant advantage is that this method accounts for both graphical and parameter uncertainty, as first proposed and demonstrated in a psychology application by Moffa et al. (36) for binary data.

where the matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ is symmetric, positive definite, and Markov relative to $\mathcal{G}$. Under the assumption of normality, the Gaussian DAG model is almost always faithful to the DAG within the parameter space, which means that the conditional independence relationships implied by the distribution are precisely the same as those represented by the DAG through the Markov property (37). For a Gaussian DAG model, we can rewrite the factorization in Equation 1 as Equation (3): (ϕ denotes the normal univariate density function) (33).

$$p(\mathbf{x} \mid \mathcal{G}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{m} \phi\left(x_i \mid \mu_i(\mathbf{x}_{pa(i)}), \sigma_i^2\right). \tag{3}$$

2.2 Do-operator

The do-operator is a fundamental concept in causal inference, proposed by Pearl (37). Provides a theoretical framework for quantifying the effects of interventions in a system. By modifying the observed probability distribution, the do-operator helps distinguish between correlation and causation. It is a core tool in causal inference and decision theory, allowing researchers to systematically answer the question "What would happen if?"

DAG provides an alternative approach to causal inference. Using DAG to describe the data generation process is appealing because the edges of the graph naturally represent the causal relationships between variables. Under a known DAG, the do-operator determines the causal effect of one variable on another. We have discussed the DAG model obtained by the Gaussian model, but we are interested in the causal relationships between the nodes, as shown in Figure 1.

Our goal is to determine the causal effect of variable X_{i} on the variable X_{o}. We use Pearl's do-operator to describe the effects of intervention (37), where the distribution of X_{o} under an intervention in X_{i} is generally indicated as $\mathbb{P}\left(X_{\mathrm{o}} = k \mid \mathrm{do}\left(X_{\mathrm{i}} = l\right)\right)$. Changes in the distribution or shifts in the distribution of the outcome variable across different levels of the intervention variable often serve as target estimands with practical significance (38). Evaluating and contrasting the change in the probability of X_{o} belonging to level k, when the intervention variable X_{i} is set to level l' versus level l offers a measure of the distribution shift:

$$\mathbb{P}\left[X_{\mathrm{o}} = k \mid \mathrm{do}\left(X_{\mathrm{i}} = l'\right)\right] - \mathbb{P}\left[X_{\mathrm{o}} = k \mid \mathrm{do}\left(X_{\mathrm{i}} = l\right)\right]. \tag{4}$$

for each $l \neq l'$ and $l, l' \in \{1, \ldots, L_{\mathrm{i}}\}$ and $k \in \{1, \ldots, L_{\mathrm{o}}\}$. We can evaluate ordinal causal effects (OCE) as represented by the target causal estimands in Equation (4).

3 Materials and methods

3.1 Quantile threshold function

In the data $\mathbf{X} = (X_1, \cdots, X_m)^\top$, the random variable $X_k = (x_1, x_2, \cdots, x_n)$, where $k = 1, 2, \cdots, m$, often includes both continuous data and ordinal data. Before causal analysis, data must be organized and optimized to ensure precision and reliability of the results. Inconsistent data types can have a significant impact on analysis results. For example, in regression analysis, it is necessary to standardize the data to ensure consistent units of measurement. The objective of this paper is to analyze ordinal data, so the first step is to properly transform unordered data into ordinal data, which supports the subsequent analysis. Therefore, we propose a quantile threshold function (QTF) that transforms continuous variables into ordinal variables and ensures the consistency of the classification results while effectively preserving the ordered nature of the data.

In this paper, we apply the kernel density estimation to fit the probability density of the continuous variable X_k in Equation 5. A Gaussian kernel function is chosen as the smoothing kernel in Equation 6, which has the advantage of not requiring a predefined data distribution shape. Through bandwidth adjustment, it effectively approximates unknown distributions. This method overcomes the dependence on distributional assumptions inherent in traditional parametric methods.

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right). \tag{5}$$

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right). \tag{6}$$

Where $\hat{f}(x)$ represents the probability density function, $K(u)$ is the Gaussian kernel function, and h is the bandwidth, which is a key smoothing factor in kernel density estimation. We adopt the bandwidth selection method proposed by Ripley (39) in his book *Modern Applied Statistics with S* as the criterion (page 127) (39). The definition of the quantile threshold function is as follows:

Definition 1 (quantile threshold function). The random variable $X = (x_1, x_2, \cdots, x_n)$ has a probability density function $f(x)$ and a

distribution function $F(x)$. If there exists a non-negative real-valued function $g(x)$ such that:

$$g(x) = \begin{cases} 0, & \text{if } x \leq Q_i, \\ 1, & \text{if } Q_i < x \leq Q_{i+1}, \\ 2, & \text{if } x > Q_{i+1}, \end{cases} . \tag{7}$$

where Q_i satisfies $F(Q_i) = \int_{-\infty}^{Q_i} f(x)dx = \frac{i}{4}, i \in \{1, 2\}$, then, the function $g(x)$ is called quantile threshold function

$g(x)$ is also a random variable. The value of $x \leq Q_i$ in X is defined as the lower level (assigned a value of 0), and within the $Q_i < x \leq Q_{i+1}$ interval is defined as a medium level (assigned a value of 1), and we classify values of $x > Q_{i+1}$ as a high level (assigned a value of 2). Based on the above definition, we can naturally deduce the following conclusion.

Proposition 1. The sum of the probabilities of the three categories equals 1, that is, $\sum_{j=0}^{2} \mathbb{P}[g(x) = j] = 1$

Proof. See the Supplementary material: Proof of Proposition 1. □

Definition 1 extends to multiple classification scenarios, as shown in the following Definition 2.

Definition 2. The random variable $X = (x_1, x_2, \cdots, x_n)$ has a probability density function $f(x)$ and a distribution function $F(x)$. If there exists a non-negative real-valued function $g(x)$ such that:

$$g(x) = \begin{cases} 0, & \text{if } x \leq Q_i, \\ 1, & \text{if } Q_i < x \leq Q_{i+1}, \\ 2, & \text{if } Q_{i+1} < x \leq Q_{i+2}, \\ \vdots, & \vdots, \\ n, & \text{if } x > Q_{i+n-1}, \end{cases} . \tag{8}$$

where Q_i satisfies $F(Q_i) = \int_{-\infty}^{Q_i} f(x)dx = \frac{i}{n+2}, i \in \{1, 2, \cdots, n\}$.

Proposition 2. The sum of the probabilities of the categories $n + 1$ is equal to 1, that is, $\sum_{j=0}^{n} \mathbb{P}[g(x) = j] = 1$.

Proof. See the Supplementary material: Proof of Proposition 2. □

Proposition 3. $\lim_{n\to\infty} \mathbb{P}\left[g(x) = n\right] = 0$.

Proof. See the Supplementary material: Proof of Proposition 3. □

From Proposition 3, we know that when performing ordered classification on random variables, the classification must be finite. Therefore, the function $g(x)$ achieves a smooth transition from continuous variables to ordinal variables. $g(x)$ provides a structured data representation that maintains both information retention and interpretability for subsequent analysis.

In this paper, we use $n = i = 2$ as the classification criterion. After the classification is completed, we conduct hypothesis testing on the results. This study employs the Analysis of Variance (ANOVA) method to verify the significance of differences between groups for reconstructed ordinal variables (40). Specifically, our objective is to determine whether the differences among the three groups are significant, which can be achieved by testing whether the means of each group are the same. We set the significance level at $\alpha = 0.05$. Let

$$H_0 : \mu_0 = \mu_1 = \mu_2$$

$$H_1 : \exists\, i, j \in \{0, 1, 2\} \quad \text{s.t.} \quad \mu_i \neq \mu_j$$

Where μ_0, μ_1, and μ_2 are the means of the three groups. If $p < 0.05$, then H_1 holds, indicating that there are significant differences among the three groups, which suggests that the above classification is effective. Conversely, if H_0 holds, it indicates that the classification levels are not significant.

3.2 Latent Gaussian DAG model

We obtained ordinal data $\mathbf{X}^* = (X_1, \cdots, X_m)^\top$ by the function $g(x)$. What we are interested in is the DAG that represents the relationships between ordinal variables. We introduce the Gaussian DAG-models in Subsection 2.1, but this model is built under the assumption that the data follow a Gaussian distribution. Therefore, for the construction of the DAG model for ordered data, we have used the OSEM algorithm proposed by Luo et al. (29). By assuming that each ordinal variable is obtained by marginally discretising a set of Gaussian variables, we can get the ordinality amongst the categories. And Gaussian variables jointly follow a DAG structure $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The OSEM algorithm provides a new framework that effectively learns Bayesian networks from ordinal data and captures the orderliness among categories.

Let $\mathbf{X}^*$ be a set of m ordinal variables, where X_k takes values in the collection of $\{\tau(k, 1), \tau(k, 2), \ldots, \tau(k, L_k)\}$ and $\tau(k, 1) < \tau(k, 2) < \cdots < \tau(k, L_k)$, $k = 1, \ldots, m$. We assume that the number of levels $L_k \geq 2$, therefore, each variable should at least be binary. It is typical to set $\tau(k, l) = l - 1$ for all $1 \leq l \leq L_k$, i.e. $\tau(k, 1) = 0, \tau(k, 2) = 1$, and so on. Further, we assume that each X_k is obtained by discretising an underlying Gaussian variable Y_k using the thresholds $-\infty = :\alpha(k, 0) < \alpha(k, 1) < \cdots < \alpha(k, L_k - 1) < \alpha(k, L_k) := \infty$. Let $\boldsymbol{\alpha}_i = (\alpha(k, 0), \ldots, \alpha(k, L_k))^\top$ and $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}_k\}_{k=1}^m$. Thus X_k is defined by the following rule:

$$X_k = \begin{cases} \tau(k, 1) & \text{if } Y_k \in (-\infty, \alpha(k, 1)) \\ \vdots & \\ \tau(k, L_k) & \text{if } Y_k \in [\alpha(k, L_k - 1), +\infty) \end{cases} \tag{9}$$

Of course, $\mathbf{Y} = (Y_1, \cdots, Y_m)^\top$ are unobservable, and we observe X_k obtained from the continuous variables by discretisation. The diagram in Figure 2 offers a visual depiction of the setup in an example case with a few variables.

Formally, Luo et al. (29) proposed a DAG model for ordinal variables based on latent Gaussian variables. The model makes the following optimization definition based on Equation 2. The

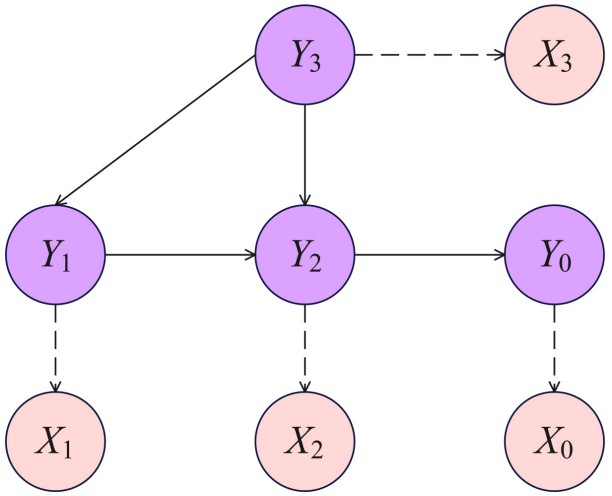

FIGURE 2
Example of latent Gaussian four-nodes DAG. Variables $X_k, k = 0, \ldots 3$ are ordinal, each obtained by discretising a latent variable Y_k with associated Gaussian parameters θ_k. Ordinal nodes are dashed for clarity.

detailed derivation process can be found in the paper (29).

$$
\begin{aligned}
&Y_k \mid \boldsymbol{y}_{\mathrm{pa}(k)}, \vartheta_k, \mathcal{G} \sim \mathcal{N}\left(\mu_k + \sum_{j \in \mathrm{pa}(k)} b_{jk}\left(y_j - \mu_j\right), v_k\right), \\
&\mathbb{P}\left(X_k = \tau(k,l) \mid Y_k = y_k, \boldsymbol{\alpha}_k\right) = \mathbb{1}\left(y_k \in [\alpha(k,l-1), \alpha(k,l)]\right), \\
&l = 1, \ldots, L_k, \\
&p(\boldsymbol{x}, \boldsymbol{y} \mid \theta, \mathcal{G}) = \prod_{k=1}^{n} \phi\left(y_k \mid \boldsymbol{y}_{\mathrm{pa}(k)}, \vartheta_k, \mathcal{G}\right) p\left(x_k \mid y_k, \boldsymbol{\alpha}_k\right),
\end{aligned}
\tag{10}
$$

The OSEM combines the multinomial probit model (41) and the structural EM algorithm of (42) to solve the problem of learning Bayesian networks from ordinal data. Specifically, the method proposes an iterative scoring and search strategy - the Ordinal Structural EM (OSEM) algorithm for learning Bayesian networks from ordinal data.

3.3 Causal effects in the latent Gaussian DAG model

Consider the general latent Gaussian DAG-model of Section 3.2. We are interested in computing the target causal estimand in Equation 4. For example, in Figure 2, the intervention variable X_i is X_1, and the outcome variable X_o is X_0. The Equation 4 can be written as Equation 11, representing the OCE on X_o of an intervention on X_i. When the intervention variable X_i is set to level l' versus level l offers a measure of the distribution shift (37):

$$
\begin{aligned}
\mathbb{P}\left[X_\mathrm{o} = \tau(\mathrm{o},k) \mid \mathrm{do}\left(X_\mathrm{i} = \tau\left(\mathrm{i},l'\right)\right)\right] - \mathbb{P}\left[X_\mathrm{o} = \tau(\mathrm{o},k)\right. \\
\left. \mid \mathrm{do}\left(X_\mathrm{i} = \tau(\mathrm{i},l)\right)\right]
\end{aligned}
\tag{11}
$$

The direct computation of Equation 11 would result in 0 for each level of the intervention and outcome variables because there is no causal path between the ordinal X_i and X_o in the DAG. But it is evident that they are causally related to each other by Y. Therefore, we can consider that if we intervene on the latent variable Y_i in a way that changes the level of its ordinal child variable X_i, and then compute the effect of this intervention on the latent parent Y_o of X_o, it is possible that the level of X_o could also change as a result. The potential change in the level of X_o resulting from an intervention on the latent parent of X_i is the OCE studied in article (30). Using the $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}_k\}_{k=1}^m$, the target causal estimand in Equation 11 on the ordinal variables can be equivalently computed as the following [(30); Definition 1,page9]:

$$
\begin{aligned}
\mathrm{OCE}_{\mathrm{io}}\left(k, l \to l'\right) = &\mathbb{P}\left[Y_\mathrm{o} \in [\alpha(\mathrm{o},k-1), \alpha(\mathrm{o},k)]\right. \\
&\left. \mid \mathrm{do}\left(Y_\mathrm{i} \in \left[\alpha\left(\mathrm{i},l'-1\right), \alpha\left(\mathrm{i},l'\right)\right]\right)\right] \\
&- \mathbb{P}\left[Y_\mathrm{o} \in [\alpha(\mathrm{o},k-1), \alpha(\mathrm{o},k)]\right. \\
&\left. \mid \mathrm{do}\left(Y_\mathrm{i} \in [\alpha(\mathrm{i},l-1), \alpha(\mathrm{i},l)]\right)\right]
\end{aligned}
\tag{12}
$$

for each $1 \leq k \leq L_\mathrm{o}, 1 \leq l, l' \leq L_\mathrm{i}$, with $l \neq l'$. The definition of OCE is anti-symmetric for the initial and end level of the intervention variable, implying that

$$
\mathrm{OCE}_{\mathrm{io}}\left(k, l \to l'\right) = -\mathrm{OCE}_{\mathrm{io}}\left(k, l' \to l\right). \tag{13}
$$

Based on the above Equation 12, the Latent Gaussian DAG-model establishes a relationship for calculating the intervention effect between ordinal variables. This allows for the computation of the OCE between variables X^*, which is equivalent to the intervention effect between variables Y. The (30) provides a detailed proof and derivation of Equation 12 in both the main text and the Appendix. Building on this result, Scauda et al. (30) proposed Proposition 5 - a method for computing OCE. The specific details can be found in Proposition 5 (Computation of the Ordinal Causal Effect) on page 12 of the (30). According to Proposition 5, we can calculate the OCE efficiently. Figure 3 illustrates the flowchart of the proposed algorithm.

4 Results

4.1 Data

Maintaining an injury-free condition is a crucial factor for success in sports. Although injuries are difficult to predict, the application of emerging technologies and data science can offer valuable insights. Even with a well-specified model, inaccurate data can compromise causal analysis. Data quality is often more critical than sample size (43), particularly in sports science, where physiological measures (e.g., maximal oxygen uptake, gene transcription activity) are inherently noisy due to biological and technical variability. Additionally, exercise intervention studies face challenges such as participant dropout, missing data, measurement errors, and inconsistencies in data processing, all of which hinder reliable causal interpretation (44).

This study utilizes a comprehensive training log dataset from Kaggle (45, 46) collected by a Dutch team in 2012–2019. This

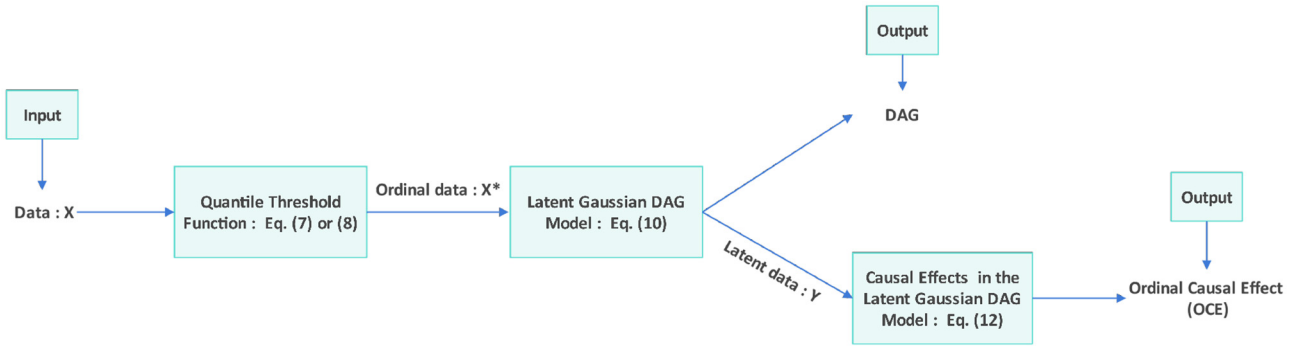

FIGURE 3
Flowchart of the algorithm: after inputting the initial data X, the QTF is applied to obtain the ordinal data X*. The Latent Gaussian DAG Model (29) identifies the latent data Y and the DAG. Then, we compute the causal effects in the Latent Gaussian DAG Model (30) to obtain the ordinal causal effect.

TABLE 1 List of Variables and Descriptions.

ID	Variable	Description
1	Sessions	Number of trainings completed
2	Totalkm	Number of kilometers covered by running
3	Kmmidlle	Number of kilometers covered in intensity zones 3 and 4
4	Kmhigh	Number of kilometers covered in intensity zone 5
5	Kmsprinting	Number of kilometers covered with sprints
6	Strengthtraining	Whether the day included a strength training session
7	PerceivedtrainingSuccess	Athlete's self-rating of how well the session went.
8	Hoursalternative	Number of hours spent on cross training
9	Perceivedexertion	Athlete's self-rating of fatigue after the session.
10	Perceivedrecovery	Athlete's self-rating of restfulness before the session.
11	Injury	Whether injured

Dataset from Kaggle: Lovdal et al. (45), Injury Prediction Dataset (45).

dataset, developed by Lövdal et al., employs machine learning to predict injuries based on data that focuses on middle- and long-distance events (800m to the marathon), with detailed performance records for 74 athletes (27 females and 47 males). The dataset adhered to the ethical principles of the *Declaration of Helsinki* and received formal approval from the ethics committee (45). This study follows the data structure established in (45), utilizing the Day approach [(45), page 1523, Table 1], with this research focusing exclusively on the data from a single day to analyze the causal effects. This dataset contains a total of 42,766 samples. Table 1 presents detailed characteristics of the variables. Variables 1, 6, and 11 are directly considered ordinal, while the remaining variables require ordinal classification using QTF.

4.2 Analysis

We simulated the probability density function for continuous variables and classified the variables based on QTF. The graph of the probability density function and its corresponding quantile ranges are presented in the Supplementary Figure 1. Taking variable 2 as an example, from Supplementary Figure 1 we can observe that Q_2 and Q_3 divide the range of the values of the variable into three parts. The portion less than Q_2 is defined as low level and assigned a value of 0. The values between Q_2 and Q_3 are categorized as medium level and assigned a value of 1. Finally, values greater than Q_3 are defined as high level and assigned a value of 2. From Supplementary Figure 1, we can see that the probability of values at the low level (green) is 0.5, the probability at the medium level (orange) is 0.25, and the probability at the high level (purple) is also 0.25. Through this method, we transform the variable from a continuous variable to an ordinal variable. Subsequently, we perform ANOVA on the three groups levels and conduct a significance test (Table 2) (40).

The mean values of the variable increase monotonically from group "0" to group "2", and the standard deviations also progressively increase. For variable 2, the means and standard deviations at different levels are as follows: In group "0", the mean is 3.186, the standard deviation is 3.991, and the sample size is 30, 437; in group "1", the mean is 12.882, the standard deviation is 0.98, and the sample size is 5, 996; in group "2", the mean is 20.016, the standard deviation is 5.2547, and the sample size is 6, 333. Other variables exhibit similar patterns, where the mean in group "0" is relatively small, while the means in groups "1" and "2" are significantly higher (e.g., for variable 5, the mean increases from 0.017 to 2.536). Several variables have extremely high F-values (e.g., 156, 092 for the variable), showing that the between-group variance is far greater than the within-group variance. The ANOVA results indicate that the differences in means across groups are statistically significant ($p < 0.05$), suggesting that the distribution of the variable differs significantly across levels and shows a clear gradient pattern (group "0" < "1" < "2"). This result implies that the method used can effectively distinguish differences between levels, leading to the rejection of the null hypothesis H_0 and the acceptance of the alternative hypothesis H_1.

TABLE 2 ANOVA for the three group levels and significance testing based on the QTF.

Ordinal	Group (mean ± SD)$_n$			F	p
	"0"	"1"	"2"		
Variable 2	$3.186 \pm 3.991_{30437}$	$12.882 \pm 0.98_{5996}$	$20.016 \pm 5.2547_{6333}$	55364	0.001
Variable 3	$0.269 \pm 1.088_{40762}$	$7.281 \pm 0.621_{1008}$	$11.298 \pm 4.244_{996}$	52351	0.001
Variable 4	$0.178 \pm 0.697_{40071}$	$5.299 \pm 0.611_{1726}$	$8.781 \pm 3.237_{969}$	77658	0
Variable 5	$0.017 \pm 0.084_{41315}$	$0.899 \pm 0.113_{764}$	$2.536 \pm 2.666_{687}$	19913	0.001
Variable 7	$0.087 \pm 0.190_{26234}$	$0.689 \pm 0.050_{8540}$	$0.850 \pm 0.063_{7992}$	102575	0.001
Variable 8	$0.050 \pm 0.020_{40181}$	$1.345 \pm 0.158_{1269}$	$2.514 \pm 1.155_{1316}$	61170	0
Variable 9	$0.047 \pm 0.067_{22291}$	$0.293 \pm 0.083_{10255}$	$0.639 \pm 0.129_{10220}$	156092	0
Variable 10	$0.055 \pm 0.074_{22844}$	$0.234 \pm 0.043_{9811}$	$0.477 \pm 0.117_{10111}$	96383	0

$p < 0.001$; The three groups of data ("0", "1", "2") are presented as mean ± standard deviation (Mean ± SD)$_n$, with n representing the sample size, and the F, p values are also listed.

In the Supplementary material, we present the ANOVA results for the three discretization methods: equal-width, equal-frequency, and k-means. Supplementary Table 1 shows the ANOVA results for equal-width. This method uses intervals of the same width, making it simple and intuitive. However, the ANOVA results indicate that its F-value is lower, suggesting the between-group differences are less significant than with the QTF method. Moreover, for outlier variables (such as Variable 4 or Variable 5), equal-width fails to reflect the data distribution accurately. Supplementary Table 2 shows the ANOVA results for equal-frequency. This method ensures that each interval contains roughly the same number of samples. When sample values are highly concentrated and repeated, the resulting zero mean and standard deviation for some groups reduces the interpretability of the data. Supplementary Table 3 presents the results of k-means clustering. The ANOVA shows large F-values for the three clusters, suggesting significant between-group differences. However, k-means is a typical data-driven method with unfixed boundaries. If clusters are ordered by their mean values to define "low," "medium," and "high" levels, the boundaries will change with the data because cluster centers are randomly initialized. This randomness weakens theoretical interpretability.

We followed the approach described in Luo et al. to derive DAG estimates from 500 bootstrap samples of the data (29, 30), utilizing the OSEM algorithm with a Monte Carlo sample size of $K = 5$ and a penalty coefficient of $\lambda = 6$. And the resulting CPDAG is shown in Figure 4.

Figure 4 illustrates the causal relationships between different variables, highlighting both direct and indirect variables surrounding the **Injury** variable. The direct influencing variable for **Injury** is **Kmsprinting**, which has a significant impact on the occurrence of **Injury**. **Kmsprinting** training typically places high stress on muscles, joints, and ligaments, especially when the intensity is excessive or recovery is insufficient, making it prone to cause tissue damage or excessive fatigue. **Kmsprinting** may lead to rapid muscle contraction and high load in a short period, thereby increasing the risk of exercise-related injuries. **Perceivedrecovery** and **Perceivedtrainingsuccess** are also direct variables influencing **Injury**. **Perceivedrecovery** plays a critical role in injury risk. When perceived recovery is poor, muscles and joints may not withstand higher loads, potentially leading to improper movement and decreased endurance, thus increasing the risk of injury. **Perceivedtrainingsuccess** directly affects injury risk. If an athlete perceives training success as high, it may indicate good physical condition and effective recovery, thereby reducing the risk of injury. Conversely, a lower perception of training success may indirectly reflect accumulated fatigue and insufficient recovery, increasing the probability of injury.

Injury's indirect influencing variables, Pathway 1 is **Totalkm** → **Sessions** → **Perceivedrecovery**→ **Injury**: Increasing total running distance can lead to higher training intensity and frequency. A rise in total distance is often accompanied by increased training frequency, which may result in insufficient recovery time. Excessive training frequency and load can compromise recovery quality, leading to poor perceived recovery and indirectly increasing the risk of injury. If increases in running distance and frequency are not balanced with adequate recovery and proper load management, the perceived recovery level may decline, significantly elevating the risk of injury.

Pathway 2 is **Strengthtraining** → **Perceivedtrainingsuccess** → **Injury**: **Strengthtraining** enhances muscle strength, joint stability, and exercise efficiency, thereby improving athletic performance and training outcomes. It also increases athletes confidence and perception of training success, which may indirectly indicate improved physical adaptation and recovery levels. **Perceivedtrainingsuccess** can reduce the risk of injury, as the body is in better physical condition, movements are more precise, and energy distribution is more efficient. Conversely, a lower perception of training success may have the opposite effect. Based on the above analysis, it is recommended to control the intensity and frequency of **Strengthtraining** to prevent muscle injuries caused by overtraining. Attention should be placed on athletes **Perceivedrecovery** and sense of **Perceivedtrainingsuccess**, and by adjusting the training plan, recovery outcomes can be effectively improved.

To visually represent the bootstrapped estimates, we present the adjacency matrices of the DAG derived using OSEM, converted to CPDAG, as a heatmap in Supplementary Figure 2 of the Supplementary material. The intensity of each cell corresponds to the frequency with which each edge appears in the bootstrapped

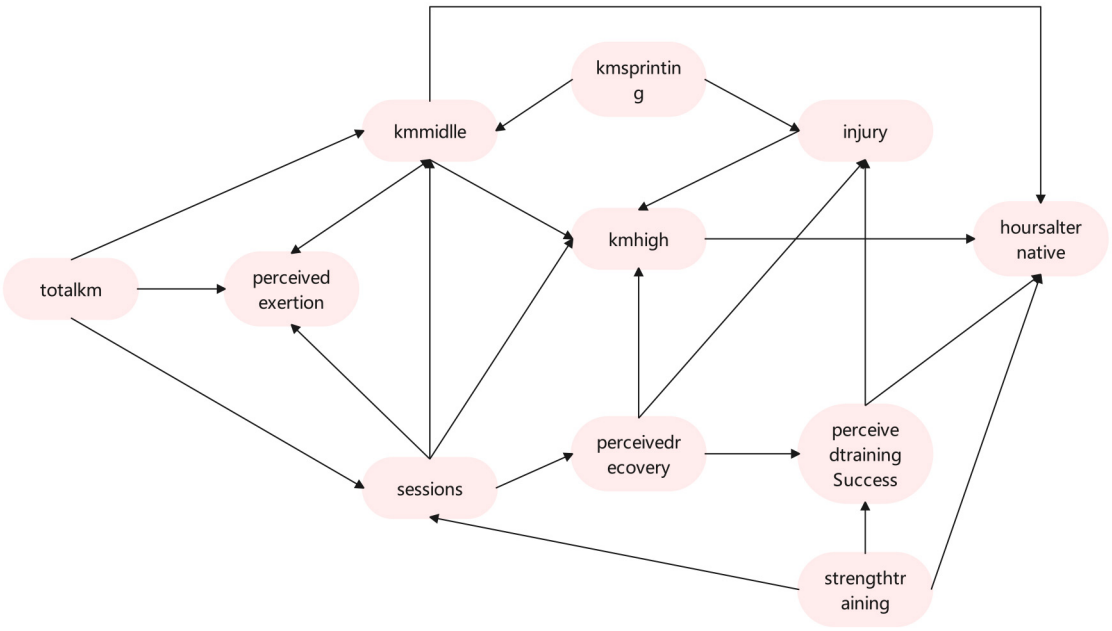

FIGURE 4
CPDAG of the hidden network structure estimated via the OSEM algorithm (29) for the ordinal data of Section 4.1.

samples. The shade in the grid indicates the proportion of times a directed edge occurs in the 500 bootstrapped CPDAG, with an undirected edge being split equally between both directions. Darker shading corresponds to a higher frequency of the respective directed edge. Additionally, we examine the causal relationship along the most frequently observed directed edge in the 500 bootstrapped CPDAG by estimating the ordinal causal direct effects of **Kmsprinting** (variable 5) on **Injury** (variable 11) within the sample's DAG. Raincloud plots, which incorporate histograms and boxplots of the estimated effects, are displayed in Figure 5.

Figure 5 shows the ordinal causal direct effects of **Kmsprinting** (variable 5) on **Injury** (variable 11). "**Injury=1**" indicates no injury, while the level of "**Injury=2**" indicates injury. We use the do-operator to describe this result.

When the "**Injury=1**", and the level of variable 5 changes from $1 \rightarrow 2$, the OCE < 0. That means $\mathbb{P}\left[\text{injury} = 1 \mid do\left(\text{kmsprinting} = 2\right)\right] - \mathbb{P}\left[\text{injury} = 1 \mid do\left(\text{kmsprinting} = 1\right)\right] < 0$. This result suggests that increasing the intensity of sprint training significantly increases the injury risk for uninjured athletes. High-intensity sprint training may lead to excessive load on muscles and joints, exceeding the body's ability to adapt, thereby increasing the likelihood of injury. Therefore, for uninjured athletes, maintaining a lower level of sprint training helps to minimize the injury risk. If the sprint training level changes from $1 \rightarrow 3$, or from $2 \rightarrow 3$, the OCE value remains negative. This finding indicates that increasing the intensity of sprint training has a significant impact on the injury probability for uninjured athletes, especially during the transitions from $1 \rightarrow 2$ and from $1 \rightarrow 3$. Among these three intervention levels, the absolute value of the OCE mean is largest for the intervention from $1 \rightarrow 3$, suggesting that transitioning directly from low to high intensity training is particularly dangerous for uninjured athletes. Therefore, training plans need to be carefully designed.

When the "**Injury=2**", and the level of variable 5 changes from $1 \rightarrow 2$, the OCE > 0. That means $\mathbb{P}\left[\text{injury} = 2 \mid do\left(\text{kmsprinting} = 2\right)\right] - \mathbb{P}\left[\text{injury} = 2 \mid do\left(\text{kmsprinting} = 1\right)\right] > 0$. When the sprint training level increases from $1 \rightarrow 2$ or $1 \rightarrow 3$, the probability of injury significantly increases. For injured athletes, continuing to increase the sprint training intensity before full recovery can worsen the existing injury or lead to incomplete recovery, which significantly increases the risk of injury. During this stage, athletes should avoid increasing sprint training intensity and prioritize basic training and recovery. When the sprint training level changes from $2 \rightarrow 3$, the OCE for both injured groups is more concentrated, with the mean close to 0, indicating that the impact of increasing from medium to high-intensity sprint training on injury is relatively small. This limited effect may be due to the athletes' adaptation to medium-intensity training and their more stable physical condition. Therefore, after sufficient medium-intensity training, gradually increasing high-intensity sprint training does not significantly increase the injury risk.

In summary, the transition from low to medium intensity ($1 \rightarrow 2$) is a critical period for injury risk, requiring close monitoring of an athlete's physical condition and recovery. After reaching medium intensity, transitioning to high intensity is relatively safe, as the body has developed some level of adaptation, resulting in a lower injury risk. For uninjured athletes, it is recommended to prioritize a progressive increase in training load and avoid a direct jump from low to high intensity ($1 \rightarrow 3$). For injured athletes, high-intensity sprint training should be strictly limited during the recovery period, and priority should be given to recovery training. High-intensity sprint training requires a focus on recovery

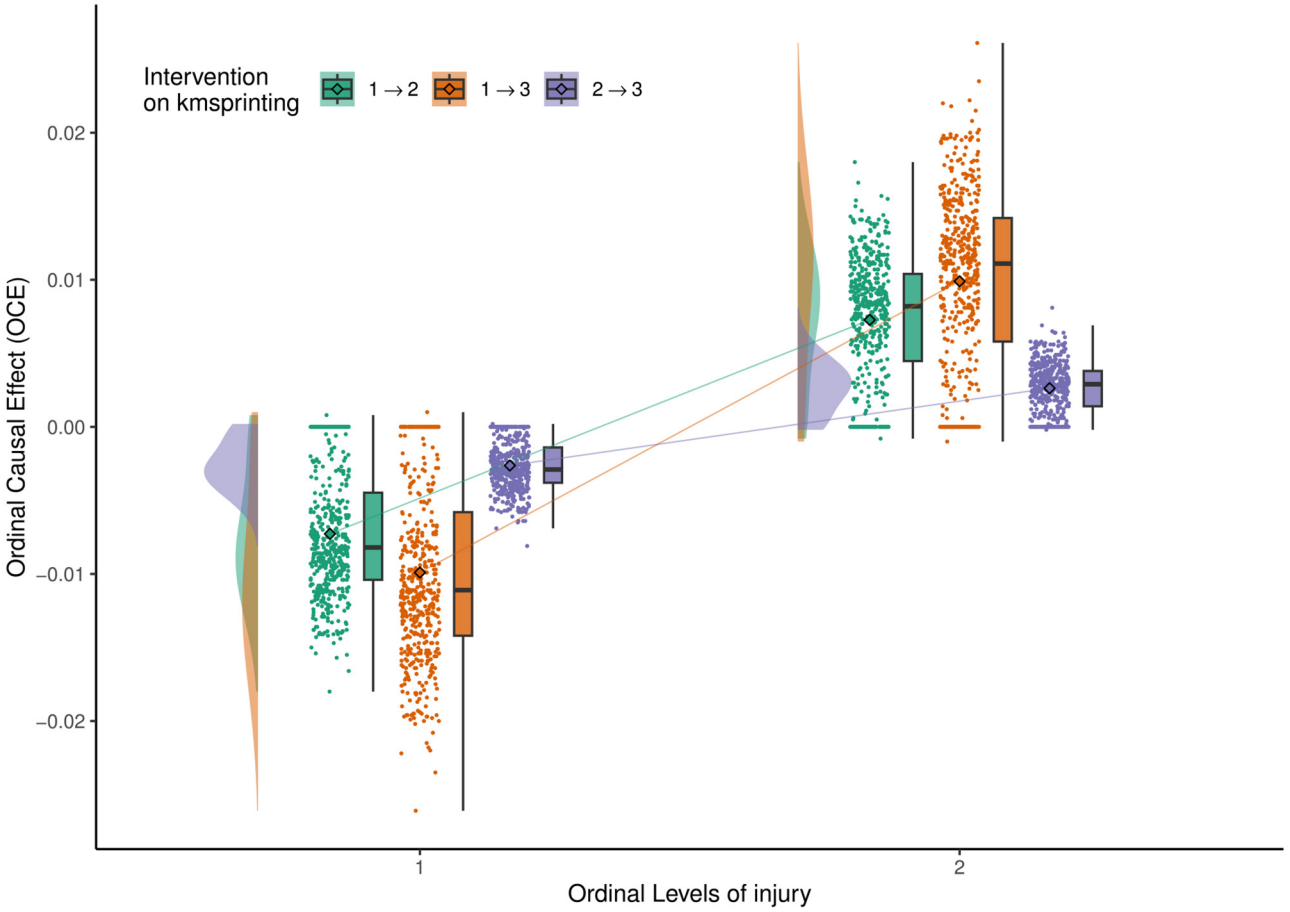

FIGURE 5
Ordinal Causal Effect of **Kmsprinting** (variable 5) on **Injury** (variable 11). The solid line connects the means of the OCEs (represented by diamond points) across different levels of the outcome variable, for each possible shift of the intervention variable.

quality after training to ensure athletes maintain a good recovery state during the gradual increase in intensity, minimizing fatigue accumulation and injury risk. The **Kmsprinting** has a significant causal effect on **Injury**. The risk from medium to high intensity (2 → 3) is relatively small, suggesting that training upgrades should be based on the athlete's adaptation. Scientifically planning sprint training intensity and pace, combined with recovery training, helps reduce injury probability while improving training effectiveness and safety.

Figure 6 shows the ordinal causal direct effects of **Kmsprinting** (variable 5) on **Kmmiddle** (variable 3) and **Kmhigh** (variable 4). The change of variable 5 from low level to medium level (1 → 2) shows that for low-level **Kmmiddle** athletes, an increase in sprinting results in a positive OCE value, indicating that improvements in sprinting may indirectly help these athletes enhance their middle-distance running ability. However, for medium and high-level **Kmmiddle** athletes, the OCE value is negative, suggesting that an increase in sprinting may decrease their performance in middle-distance running. A similar pattern applies to **Kmhigh** athletes. For low-level **Kmhigh** athletes, an increase in sprinting may improve their high-intensity running ability. In contrast, for medium and high-level athletes, an increase in sprinting may limit their high-intensity running ability.

It is worth noting that the OCE for **Kmhigh** is more scattered, indicating significant variation in the impact of sprinting on high-intensity running. The change of variable 5 from medium level to high level (2 → 3) shows that for both **Kmmiddle** and **Kmhigh**, the OCE are close to 0. This result suggests that once athletes have adapted to a certain level of sprinting load, further increases in sprint intensity have minimal effect on middle-distance running and high-intensity running performance. For athletes with weaker middle-distance and high-intensity running abilities, increasing sprint training can indirectly enhance their running abilities by improving neuromuscular adaptation and rapid power output. For athletes who already possess strong middle-distance and high-intensity running capabilities, increasing sprint training may lead to overloading or accumulation of fatigue, thereby affecting other running abilities. Once sprinting has reached a high level, athletes' bodies gradually adapt to the high-intensity load, and further increases in sprinting have limited intervention effects on other running abilities.

Therefore, we can conclude that for low-level athletes, appropriately increasing sprint training can help improve their middle-distance and high-intensity running abilities. For medium and high-level athletes, the ratio of sprinting to other running training must be balanced to avoid excessive sprint loads that

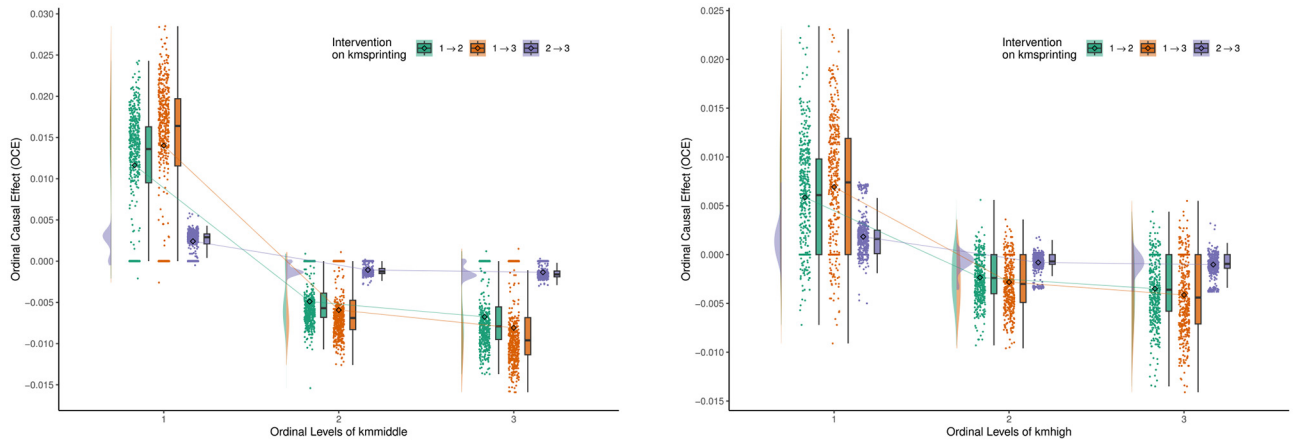

FIGURE 6
Ordinal causal effect of **Kmsprinting** (variable 5) on **Kmmiddle** (variable 3) and **Kmhigh** (variable 4).

may affect overall running performance. Regularly monitoring athletes' performance in different running abilities and adjusting the sprinting load based on data feedback will ensure maximized training effects while preventing fatigue accumulation. As sprint load increases from medium to high intensity (2 → 3), gradual adaptation should be emphasized to avoid sudden increases in training intensity that could negatively impact middle-distance or high-intensity running abilities. Through reasonable design and adjustment of sprint training, coaches and athletes can enhance specific abilities while minimizing the risk of injuries caused by overtraining, ultimately achieving comprehensive optimization of athletic performance.

Figure 7 shows the ordinal causal direct effects of other variables on **Injury** (variable 11). From Figure 7, we can observe that the OCE of **Kmmiddle**(variable 3) and **Hoursalternative**(variable 8) on injury are close to zero across various levels of change. This finding indicates that these two variables have a minimal impact on injury risk. **Kmmiddle** typically involves moderate-intensity training, and athletes' bodies are generally more adaptable, so it does not significantly increase or decrease injury risk. **Hoursalternative** mainly refers to low-intensity recovery exercises, which place minimal stress on the body, thus limiting their impact on injury risk. These two variables can be part of a stable training load to alleviate the physical stress from high-intensity training, thereby improving the overall safety of the training program.

In contrast, the changes in **Strengthtraining**(variable 6) show the most significant fluctuations in the OCE for injury, indicating a significant causal intervention effect. This finding suggests that strength training can enhance athletic performance but may also increase injury risk if the intensity is too high or the progression is too rapid. High-intensity strength training places a heavy load on muscles and joints, and without adequate recovery or adaptive training, it can lead to muscle strains, ligament injuries, and other issues. Additionally, **Strengthtraining**'s effects vary significantly among individuals, as different athletes have differing capacities to tolerate intensity and recover. This further amplifies its impact on injury risk. Therefore, strength training programs must strictly control intensity and progression, prioritizing the development of foundational strength and movement stability. This approach helps mitigate injury risk caused by excessive or improperly planned training.

The Pathway 2 analysis of the CPDAG in Figure 4 shows that **Strengthtraining**(variable 6) indirectly influences injury risk through Perceived **Perceivedtrainingsuccess**(variable 7), which supports the conclusion that their impact on injury risk is relatively limited. In practical sports, this result suggests that strength training should pay particular attention to intensity management. Proper planning of load progression is key to avoiding sports injuries, especially during high-intensity strength training, where it is important to incorporate restorative training [such as extending **Hoursalternative**(variable 8)]. **kmmiddle**(variable 3) and **Hoursalternative**(variable 8) can serve as transition phases for high-intensity training, helping athletes gradually adapt to higher loads and reduce fatigue accumulation. Combining feedback from **Perceivedrecovery**(variable 10) and **Perceivedtrainingsuccess**(variable 7) allows for the real-time optimization of training plans, ensuring a balance between performance enhancement and injury risk. Through careful design and dynamic adjustment of training plans, athletes can not only reduce the risk of training-induced injuries but also improve overall training quality. Additionally, we provide the causal direct effects of other variables in the Supplementary Figure 3 for further reference.

5 Discussion

In this work, we propose a quantile threshold function (QTF) that transforms continuous variables into ordinal variables, and ensures the consistency of the classification results while effectively preserving the ordered nature of the data. Based on data transformation, this study applies the method designed by Luo et al. (29) to construct a causal Directed Acyclic Graph (DAG). Then, using the ordinal data causal analysis algorithm proposed by Scauda et al. (30), the ordinal causal effects (OCE) between ordered variables are calculated within the framework of the latent Gaussian DAG model. These findings offer valuable insights for optimizing rehabilitation strategies

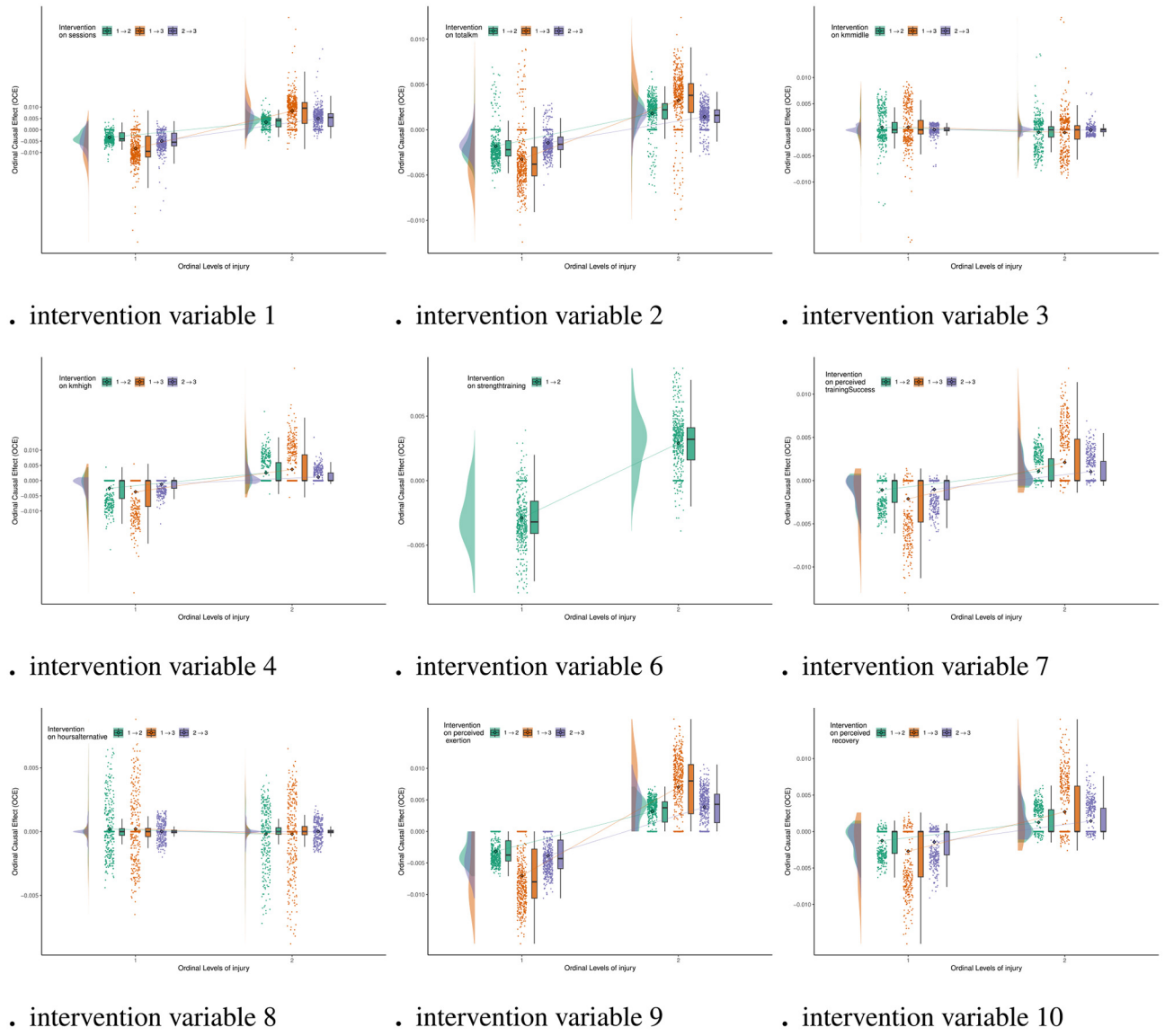

FIGURE 7
Ordinal causal effect of variable on **Injury** (variable 11).

and provide the critical foundations for the development of effective prevention strategies. The use of causal diagrams will facilitate the organization of key concepts and ideas related to athletic injury causation within a well-defined causal framework. This approach enables the exploration of specific causal links and underlying assumptions through appropriate scientific methods. The findings not only advance methodological understanding but also provide a robust theoretical foundation for optimizing rehabilitation protocols and designing effective injury prevention strategies, ultimately contributing to a more precise and thorough understanding of the mechanisms driving sports injury occurrence.

The results highlight several critical factors influencing injury risk, encompassing both measurable training loads and athlete-reported perceptions. Elevated training demands–particularly high-intensity sprinting distances–can substantially increase musculoskeletal stress and, when paired with insufficient recovery, raise the likelihood of injury. At the same time, athlete perceptions of recovery quality and training success emerged as strong indicators of physical readiness and resilience; lower scores in these measures often reflect accumulated fatigue and compromised movement control. Variations in total running volume, session frequency, and strength training intensity were also found to influence these perceptions, thereby indirectly shaping injury risk. Practical applications of these insights include regulating sprinting distances and intensities to prevent acute overload, progressively adjusting total running volume and session frequency, and structuring strength training to maximize performance benefits without inducing overtraining. Athlete-reported measures of recovery and training success can serve as low-cost, real-time indicators for fine-tuning training plans before injuries occur.

From a policy perspective, integrating both objective load metrics and subjective perception measures into institutional injury surveillance systems would enhance early detection and prevention. Sports organizations could define evidence-based thresholds for key indicators, implement mandatory periodic monitoring, and foster a training culture that prioritizes recovery alongside performance goals. The quantile threshold framework used in this study also offers practical training load monitoring guidance. By determining safe ranges for load-related variables and coupling them with perceptual feedback, coaches can detect emerging imbalances between workload and recovery, enabling timely adjustments that maintain athletes in optimal performance zones while minimizing injury risk. Finally, the findings have important implications for recovery strategies. Structured recovery programs should address both physiological restoration–through rest intervals, active recovery, sleep optimization, and nutrition–and psychological readiness, by enhancing athletes confidence and perceived training success. A dual focus on physical and perceptual recovery can improve resilience to high-intensity demands and contribute to sustained performance with lower injury incidence.

This study proposes a quantile threshold function (QTF) that transforms continuous variables into ordinal variables. Although this method offers simplicity and effectiveness, it is essential to recognize that various alternative approaches exist for converting continuous variables into ordinal variables. Methods like decision tree binning and K-Means clustering can capture complex relationships between continuous variables and other features more precisely. However, they require higher computational power and suffer from poor theoretical interpretability. Future studies could explore hybrid discretization schemes that combine the advantages of multiple methods. For example, integrating statistical techniques with machine learning algorithms may yield more flexible and efficient discretization strategies. Investigating the impact of different discretization methods on subsequent tasks, such as predictive modeling or clustering, could also provide valuable insights for practical applications. By addressing these limitations and broadening the analytical perspective, future research is expected to explore optimal strategies for transforming continuous variables into ordinal categories, thereby advancing the development of more comprehensive data preprocessing frameworks and improving model performance.

This study examines a single-day subset of the dataset, which limits the generalizability of the findings. Future work will extend the analysis to weekly datasets and employ time-series methods to capture temporal dynamics better. Moreover, although the current study focuses on single interventions, it is crucial to acknowledge that, in real-world scenarios, any exogenous intervention can simultaneously influence multiple target variables. Therefore, predicting the impact of joint interventions on an outcome variable becomes a relevant consideration. The methods proposed by (47) can be readily extended to address multiple interventions. Consequently, under the latent Gaussian model, incorporating joint interventions–similar to the approach of (48) represents a natural progression within the latent space framework. In the main text, we also mentioned that some datasets contain both ordinal and non-ordinal data. Naturally, this leads us to consider the problem of Bayesian network learning with mixed data. In particular, one may obtain a dataset with both continuous and ordinal variables by first generating a Gaussian dataset according to a DAG structure and then discretizing some of the variables while keeping others continuous. A similar learning framework may be applicable in this context. Therefore, future research could explore using such learning frameworks for causal analysis of data.

Data availability statement

The datasets for this study can be found in the [shashwatwork/injury-prediction] repository on Kaggle https://www.kaggle.com/datasets/shashwatwork/injury-prediction-for-competitive-runners. Software in the form of R code, R code is available at [https://github.com/TaoXyjammy/OrdinalEffectsSport].

Ethics statement

This study utilized a secondary dataset, and all data were strictly anonymized. The privacy and personal 638 information of the study subjects were fully protected. This research was conducted in strict accordance with the ethical guidelines set forth in the Declaration of Helsinki.

Author contributions

TX: Conceptualization, Data curation, Writing – original draft, Writing – review & editing. YH: Data curation, Investigation, Methodology, Writing – review & editing. FX: Investigation, Software, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Acknowledgments

This work was made possible through the collaboration and strong support of Kyungil University, City University of Hong Kong, and Shanxi Normal University. The authors would also like to express their sincere gratitude to LetPub for their professional language editing services. It is through the combined efforts and collaboration of all parties that this article has been successfully presented.

Conflict of interest

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

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2025.1647200/full#supplementary-material

References

1. Carlin JB, Moreno-Betancur M. On the uses and abuses of regression models: A call for reform of statistical practice and teaching. *arXiv* [Preprint] arXiv:2309.06668. (2023). doi: 10.48550/arXiv.2309.06668

2. Hernn MA, Hsu J, Healy B. A second chance to get causal inference right: a classification of data science tasks. *Chance*. (2019) 32:4249. doi: 10.1080/09332480.2019.1579578

3. Agresti A. *Analysis of Ordinal Categorical Data*. New York: John Wiley & Sons. (2010).

4. Lynch BM, Dixon-Suen SC, Ramirez Varela A, Yang Y, English DR, Ding D, et al. Approaches to improve causal inference in physical activity epidemiology. *J Phys Activity Health*. (2020) 17:80–4. doi: 10.1123/jpah.2019-0515

5. Nielsen RO, Simonsen NS, Casals M, Stamatakis E, Mansournia MA. Methods matter and the too much, too soon theory (part 2): what is the goal of your sports injury research? Are you describing, predicting or drawing a causal inference? *Br J Sports Med*. (2020) 54:1307–1309. doi: 10.1136/bjsports-2020-102144

6. Shrier I. Understanding causal inference: the future direction in sports injury prevention. *Clin J Sport Med*. (2007) 17:220. doi: 10.1097/JSM.0b013e3180385a8c

7. Rommers N, Rssler R, Shrier I, Lenoir M, Witvrouw E, DHondt E, et al. Motor performance is not related to injury risk in growing elite-level male youth football players: A causal inference approach to injury risk assessment. *J Sci Med Sport*. (2021) 24:881–5. doi: 10.1016/j.jsams.2021.03.004

8. Hopkins W. Research designs: choosing and fine-tuning a design for your study. *Sport Sci*. (2008) 12:13. Available online at: https://www.sportsci.org/2008/wghdesign.pdf

9. Steele J, Fisher J, Crawford D. Does increasing an athletes strength improve sports performance? A critical review with suggestions to help answer this, and other, causal questions in sport science. *J Trainol*. (2020) 9:20. doi: 10.17338/trainology.9.1_20

10. van Mechelen W, Hlobil H, Kemper HC. Incidence, severity, aetiology and prevention of sports injuries. A review of concepts. *Sports Med*. (1992) 14:82–99. doi: 10.2165/00007256-199214020-00002

11. Finch C. A new framework for research leading to sports injury prevention. *J Sci Med Sport*. (2006) 9:3–9. doi: 10.1016/j.jsams.2006.02.009

12. Pearl J, Mackenzie D. *The Book of Why: the New Science of Cause and Effect*. New York: Basic Books. (2018).

13. Pearl J, Causality. Models, reasoning, and inference. In: Pearl J, editor. *Econometric Theory*. Cambridge: Cambridge University Press (2000).

14. Malina D, Bothwell LE, Greene JA, Podolsky SH, Jones DS. Assessing the Gold Standard lessons from the History of RCTs. *N Engl J Med*. (2016) 374:2175–81. doi: 10.1056/NEJMms1604593

15. Bullock GS, Ward P, Hughes T, Thigpen CA, Cook CE, Shanley E. Using randomized controlled trials in the sports medicine and performance environment: is it time to reconsider and think outside the methodological box? *J Orthop Sports Phys Ther*. (2023) 53:331–4. doi: 10.2519/jospt.2023.11824

16. Bonell Monsons O, Sprri J, Gouttebarge V, Bolling C, Verhagen E. A survey on current practices, needs, responsibilities and preferences for knowledge dissemination in the field of injury and illness prevention among competitive snow sports stakeholders. *Sports Med - Open*. (2025) 11:17. doi: 10.1186/s40798-025-00818-9

17. Neumann ND, Brauers JJ, van Yperen NW, van der Linde M, Lemmink KAPM, Brink MS, et al. Critical fluctuations as an early warning signal of sports injuries? A proof of concept using football monitoring data. *Sports Med - Open*. (2024) 10:129. doi: 10.1186/s40798-024-00787-5

18. Kalkhoven JT. Athletic injury research: frameworks, models and the need for causal knowledge. *Sports Med*. (2024) 54:1121–37. doi: 10.1007/s40279-024-02008-1

19. Eisenhart M. Conceptual frameworks for research circa 1991: ideas from a cultural anthropologist; implications for mathematics education rese. In: Underhill RG, editor *Proceedings of the 13th Annual Meeting of the North American Chapter of the Psychology of Mathematics Education*. Blacksburg: Margaret Eisenhart. (1991).

20. Fried EI. Theories and models: what they are, what they are for, and what they are about. *Psychol Inq*. (2020) 31:336–44. doi: 10.1080/1047840X.2020.1854011

21. Smaldino PE. How to translate a verbal theory into a formal model. *Psychol Stud*. (2020) 65:187–96.

22. Williams TC, Bach CC, Matthiesen NB, Henriksen TB, Gagliardi L. Directed acyclic graphs: a tool for causal studies in paediatrics. *Pediatr Res*. (2018) 84:487–93. doi: 10.1038/s41390-018-0071-3

23. Shrier I, Platt RW. Reducing bias through directed acyclic graphs. *BMC Med Res Methodol*. (2008) 8:70. doi: 10.1186/1471-2288-8-70

24. McLean S. Kerhervé HA, Stevens N, Salmon PM. A systems analysis critique of sport-science research. *Int J Sports Physiol Perform*. (2021) 16:1385–92. doi: 10.1123/ijspp.2020-0934

25. Richardson TS, Robins JM. Single world intervention graphs (SWIGs): a unification of the counterfactual and graphical approaches to causality. In: *Center for the Statistics and the Social Sciences, University of Washington Series Working Paper*. (2013). p. 128.

26. Verma T, Pearl J. *Equivalence and Synthesis of Causal Models*. New York, NY, USA: Association for Computing Machinery (2022). p. 221–236.

27. Rios FL, Moffa G, Kuipers J. Benchpress: a scalable and versatile workflow for benchmarking structure learning algorithms. *arXiv* [preprint] arXiv:210703863. (2021). doi: 10.48550/arXiv.2107.03863

28. Kitson NK, Constantinou AC, Guo Z, Liu Y, Chobtham K. A survey of Bayesian Network structure learning. *Artif Intellig Rev*. (2023) 56:8721–814. doi: 10.1007/s10462-022-10351-w

29. Luo XG, Moffa G, Kuipers J. Learning Bayesian networks from ordinal data. *J Mach Learn Res*. (2021) 22:1–44. Available online at: http://www.jmlr.org/papers/volume22/20-1338/20-1338.pdf

30. Scauda M, Kuipers J, Moffa G. A latent causal inference framework for ordinal variables. *arXiv* [preprint] arXiv:250210276. (2025). doi: 10.48550/arXiv.2502.10276

31. Lauritzen SL. Graphical Models. Oxford: Oxford University Press: Clarendon Press. (1996). doi: 10.1093/oso/9780198522195.001.0001

32. Koller D, Friedman N. *Probabilistic Graphical Models: Principles and Techniques*. New York: The MIT Press. (2009).

33. Geiger D, Heckerman D. Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *Ann Stat*. (2002) 30:1412–40. doi: 10.1214/aos/1035844981

34. Pearl J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. New York: Morgan Kaufmann. (2014).

35. Maathuis MH, Kalisch M, Bühlmann P. Estimating high-dimensional intervention effects from observational data. *Ann Statist*. (2009) 37:3133–64. doi: 10.1214/09-AOS685

36. Moffa G, Catone G, Kuipers J, Kuipers E, Freeman D, Marwaha S, et al. Using directed acyclic graphs in epidemiological research in psychosis: an analysis of the role of bullying in psychosis. *Schizophr Bull*. (2017) 43:1273–9. doi: 10.1093/schbul/sbx013

37. Pearl J. *Causality*. Cambridge: Cambridge University Press. (2009).

38. Holland PW. Causal inference, path analysis and recursive structural equations models. *ETS Res Rep Ser*. (1988) 1988:i50. doi: 10.1002/j.2330-8516.1988.tb00270.x

39. Ripley BD. *Modern Applied Statistics with S*. Berlin: Springer. (2002).

40. Sthle L, Wold S. Analysis of variance (ANOVA). *Chemomet Intellig Lab Syst*. (1989) 6:259–72. doi: 10.1016/0169-7439(89)80095-4

41. Daganzo C. *Multinomial Probit: the Theory and its Application to Demand Forecasting*. Amsterdam: Elsevier. (2014).

42. Friedman N. *Learning Belief Networks in the Presence of Missing Values and Hidden Variables*. Berkeley, CA: ICML (1997). p. 125–33.

43. Meng XL. Statistical paradises and paradoxes in big data. (i): law of large populations, big data paradox, and the (2016) US Presidential Election. *Ann Appl Statist*. (2018) 12:685–726. doi: 10.1214/18-AOAS1161SF

44. Nolte S, Rein R, Quittmann OJ. Data processing strategies to determine maximum oxygen uptake: a systematic scoping review and experimental comparison with guidelines for reporting. *Sports Med*. (2023) 53:2463–75. doi: 10.1007/s40279-023-01903-3

45. Lövdal SS, Den Hartigh RJ, Azzopardi G. Injury prediction in competitive runners with machine learning. *Int J Sports Physiol Perform*. (2021) 16:1522–31. doi: 10.1123/ijspp.2020-0518

46. Tiwari S, Community K. *Injury Prediction for Competitive Runners*. San Francisco, CA: Kaggle (2021). Available online at: https://www.kaggle.com/datasets/shashwatwork/injury-prediction-for-competitive-runners (Accessed July 5, 2024).

47. Viinikka J, Hyttinen A, Pensar J, Koivisto M. Towards scalable bayesian learning of causal DAGs. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. *Advances in Neural Information Processing Systems*. Vancouver, Canada: Curran Associates, Inc. (2020). p. 6584–94.

48. Nandy P, Maathuis MH, Richardson TS. Estimating the effect of joint interventions from observational data in sparse high-dimensional settings. *Ann Statis*. (2017) 45:647–74. doi: 10.1214/16-AOS1462