Artificial Intelligence In Medicine 173 (2026) 103350

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Research paper

Integrating probabilistic trees and causal networks for clinical and epidemiological data

Sheresh Zahoor [a,*], Pietro Liò [b], Gaël Dias [c], Mohammed Hasanuzzaman [d]

[a] Munster Technological University, Rossa Ave, Bishopstown, Cork, Ireland
[b] Department of Computer Science and Technology, University of Cambridge, The Old Schools, Trinity Ln, Cambridge, United Kingdom
[c] Université Caen Normandie, ENSICAEN, CNRS, Normandie Univ, GREYC UMR6072, F-14000 Caen, France
[d] The School of Electronics, Electrical Engineering and Computer Science, Queens University Belfast, University Road, Belfast, United Kingdom

ARTICLE INFO

Keywords:
Causal Bayesian networks
Healthcare
Probability trees
Causal inference

ABSTRACT

Healthcare decision-making requires not only accurate predictions but also insights into how factors influence patient outcomes. While traditional machine learning (ML) models excel at predicting outcomes, such as identifying high-risk patients, they are limited in addressing "what if" questions about interventions. This study introduces the Probabilistic Causal Fusion (PCF) framework, which integrates Causal Bayesian Networks (CBNs) and Probability Trees (PTrees) to extend beyond predictions. PCF leverages causal relationships from CBNs to structure PTrees, enabling both the quantification of factor impacts and the simulation of hypothetical interventions. The framework is evaluated on three clinically diverse, real-world datasets, MIMIC-IV, Framingham Heart Study, and BRFSS (Diabetes), demonstrating consistent predictive performance comparable to conventional ML models, while offering enhanced interpretability and causal reasoning capabilities. In contrast to conventional approaches focused solely on prediction, PCF offers a unified framework for prediction, intervention modelling, and counterfactual analysis, forming a holistic toolkit for clinical decision support. To enhance interpretability, PCF incorporates sensitivity analysis and SHapley Additive exPlanations (SHAP). Sensitivity analysis quantifies the influence of causal parameters on outcomes such as Length of Stay (LOS), Coronary Heart Disease (CHD), and Diabetes, while SHAP highlights the importance of individual features in predictive modelling. This dual-layered interpretability offers both macro-level insights into causal pathways and micro-level explanations for individual predictions. By combining causal reasoning with predictive modelling, PCF bridges the gap between clinical intuition and data-driven insights. Its ability to uncover relationships between modifiable factors and simulate hypothetical scenarios provides clinicians with a clearer understanding of causal pathways. This approach supports more informed, evidence-based decision-making, offering a robust framework for addressing complex questions in diverse healthcare settings.

1. Introduction

Effective healthcare requires not just accurate predictions but also a deeper understanding of the factors that drive patient outcomes. While traditional machine learning (ML) models are proficient at identifying patterns and forecasting risks, such as predicting which patients are more likely to develop a condition, they often fall short in evaluating how specific interventions might alter these outcomes. This predictive focus limits their utility in addressing causal questions, such as estimating the impact of treatments or other modifiable factors on patient trajectories. To bridge this gap, there is a growing need for approaches that go beyond prediction, enabling the quantification of causal effects and simulation of intervention outcomes.

Causal ML addresses these gaps by estimating treatment effects and answering counterfactual questions, such as "How would a patient's outcome change if a different treatment were administered?" Unlike traditional ML, which focuses on correlations, causal ML is built on the foundation of causal inference [1], enabling a deeper understanding of relationships and supporting evidence-based decision-making [2]. For instance, traditional ML might predict a patient's likelihood of developing diabetes [3], but causal ML can estimate how that likelihood would change under specific interventions, such as a lifestyle modification or a new medication [4]. These capabilities are particularly valuable in healthcare, where understanding cause–effect relationships is critical for developing targeted interventions [5].

[*] Corresponding author.
E-mail address: sheresh.zahoor@mymtu.ie (S. Zahoor).

https://doi.org/10.1016/j.artmed.2026.103350
Received 22 January 2025; Received in revised form 18 December 2025; Accepted 5 January 2026
Available online 6 January 2026

To support both causal reasoning and practical interpretability, we propose the Probabilistic Causal Fusion (PCF) framework, which combines Causal Bayesian Networks (CBNs) with ensembles of Probability Trees (PTrees). CBNs model dependencies between variables using directed acyclic graphs (DAGs) [1], providing a principled foundation for causal inference. While CBNs are well-suited for modelling joint distributions, their global structure can make patient-specific reasoning paths difficult to interpret in practice.

To address this, PCF incorporates Probability Trees, interpretable, rule-based models that align more naturally with clinical workflows [6,7]. Their hierarchical paths provide clear, context-specific decision rules (e.g., "BMI > 30 → HbA1c > 6.5% → High ICU stay risk"), which clinicians can easily follow and validate. These paths also capture context-specific dependencies that are challenging to express compactly in CBNs, as shown in staged tree learning [8]. In addition to their transparency, PTrees are computationally efficient and well-suited to ensemble learning strategies such as bootstrap aggregation, which improves robustness and generalisability. The main limitation of PTrees, however, lies in their dependence on predefined variable orderings, a process that traditionally relies on expert input and introduces subjectivity, inefficiency, and inconsistency.

This reliance on expert-defined sequences not only introduces bias but also limits scalability. To overcome this, PCF leverages the causal structure learned from CBNs to inform variable ordering, an essential step in guiding PTree construction. While CBNs can provide a topological ordering from the learned graph structure, recent works show that structure learning algorithms are sensitive to dataset column order, leading to instability in learned graphs [9]. To mitigate this, PCF aggregates outputs from multiple CBNs learned under varying conditions and derives a consensus topological ordering based on stable, frequently occurring edges. This data-driven ordering aligns PTree decision paths with inferred causal dependencies, reducing reliance on expert specification and preserving clinical interpretability. Clinicians may still refine the ordering locally if desired, but without needing to manually reconstruct the full structure. The result is a hybrid approach that balances automated discovery with domain expertise, enhancing both robustness and transparency in complex, high-dimensional datasets such as electronic health records.

This hybrid approach is particularly valuable in clinical settings, where observational data are abundant and experimental studies are often impractical. Unlike Randomised Controlled Trials (RCTs), which are costly and time-consuming, PCF enables causal inference and intervention simulation directly from observational data. By moving beyond purely predictive models and towards transparent causal reasoning, PCF supports clinicians in exploring not just what might happen, but how and why, and how it could be changed. These capabilities are made possible by PCF's hierarchical design, which brings together the global structure-learning strength of CBNs with the local interpretability of PTrees.

The hierarchical architecture of PCF enhances its clinical relevance by combining the complementary strengths of CBNs and PTrees. CBNs are used to uncover dependencies among clinical variables, such as comorbidities and temporal trends, while PTrees translate these dependencies into interpretable, cohort-specific decision rules. This layered design supports the integration of heterogeneous data sources, including bedside monitoring and structured electronic health records, into a unified framework for causal reasoning. Rather than claiming to recover true causal structure with certainty, PCF aligns the structure of Probability Tree models with the conditional dependencies inferred by CBNs, dependencies that, under standard assumptions, encode hypothesised causal relationships through patterns of conditional independence. In doing so, the framework reduces the need for manually defined variable hierarchies and improves the consistency of modelling across settings. While domain expertise remains important, PCF helps mitigate uncalibrated subjectivity by integrating data-driven structure learning with opportunities for expert refinement, thereby enhancing both reproducibility and clinical interpretability.

Traditionally, constructing PTrees has been an iterative process reliant on domain experts to define variable orderings [7], which introduces challenges such as:

1. Subjectivity: Expert knowledge may be incomplete or biased, leading to suboptimal structures.
2. Inefficiency: Manual construction is time-intensive, especially for complex datasets.
3. Data Inconsistency: Sole reliance on expert-defined structures can overlook key relationships present in the data.

By employing an ensemble learning approach, PCF mitigates these limitations, offering a robust and generalisable solution that reduces overfitting and enhances performance.

Beyond its core modelling capabilities, the PCF framework also lays the foundation for a centralised causal knowledge repository, an initiative designed to address several persistent challenges in healthcare analytics. In particular, it responds to the difficulty many smaller institutions face in building robust causal models from limited local data, and to the fragmented nature of causal discovery across the healthcare sector. By enabling the sharing of pre-trained PCF models, alongside their underlying causal graphs across organisations, the framework supports a more collaborative and equitable approach to model development. This allows smaller centres to build upon validated causal structures and adapt them to local contexts, rather than starting from scratch. For example, a large urban hospital (Hospital A) might train a PCF model on its extensive ICU dataset, producing a validated causal graph and ensemble of Probability Trees. A smaller regional hospital (Hospital B), with limited local data, could then reuse this pre-trained PCF model as a foundation for its own decision support system. By fine-tuning the model with local data or adapting selected components, Hospital B can deploy a robust, interpretable tool without the need for full retraining, thereby accelerating implementation and improving accessibility in resource-constrained settings.

Crucially, PCF's modular architecture, separating structure learning via CBNs from outcome modelling via PTrees, makes it uniquely well-suited to this kind of distributed refinement. Over time, such a system could accelerate the translation of causal insights into practice, promote greater methodological consistency, and support the development of generalisable models that are responsive to the needs of diverse healthcare environments.

In this work, we leverage the PCF framework to support prediction, intervention, and counterfactual analysis in three distinct clinical contexts. First, we aim to predict and identify factors associated with the length of stay in the Intensive Care Unit (ICU). Second, the framework is applied to assess the risk of Chronic Heart Disease (CHD) and investigate potential modifiable factors that could mitigate its progression. Finally, we utilise the framework to predict the risk of diabetes and analyse the influence of various risk factors on its onset.

The main contributions of this work are summarised as follows:

- We propose a framework, Probabilistic Causal Fusion (PCF), that combines Causal Bayesian Networks (CBNs) with ensembles of Probability Trees (PTrees) to enable predictions, interventions, and counterfactual analysis in healthcare. This integration addresses the limitations of traditional PTree construction, such as reliance on domain expertise for variable ordering, by leveraging causal relationships identified through CBNs.
- The use of an ensemble of PTrees improves predictive performance and robustness. By balancing the trade-off between bias and variance, the ensemble approach mitigates risks of overfitting or underfitting and enhances generalisability.
- Methodological refinements are introduced in computing transition probabilities within the PTree framework. Specifically, data is partitioned using empirical marginal probabilities, while causal relationships derived from CBNs inform split decisions, enabling more data-driven and effective model construction.

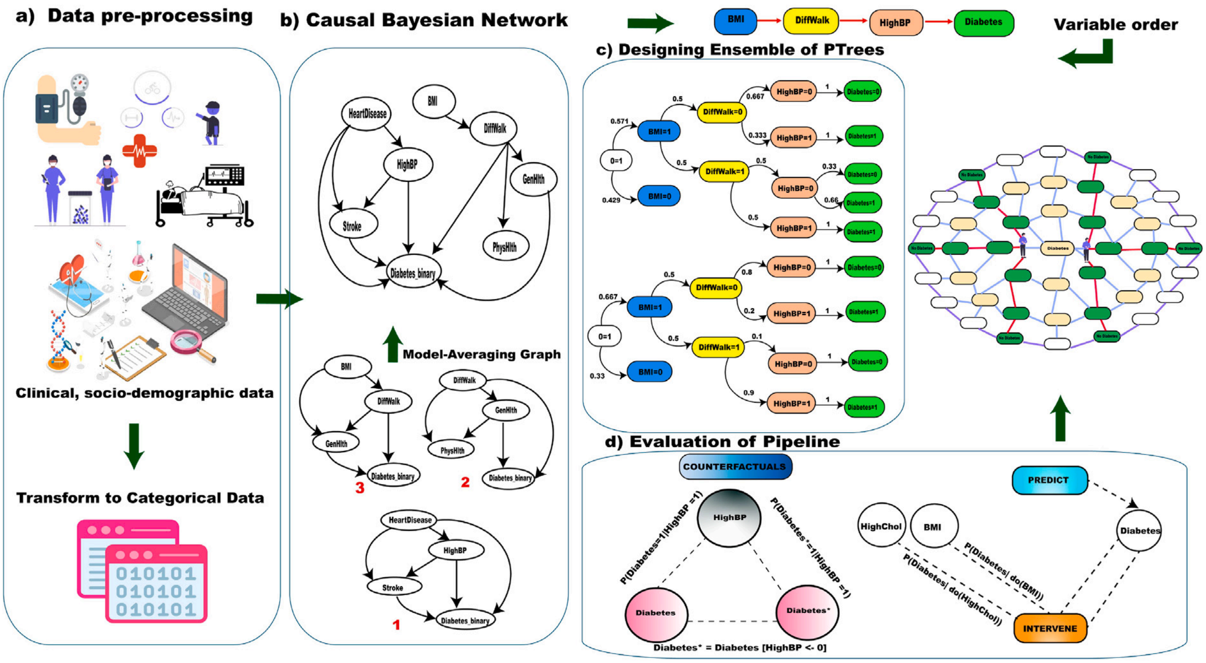

Fig. 1. Different steps involved in the PCF framework. The first module addresses data pre-processing to shape the input required for the CBN. The next module involves generating individual CBNs and creating a model-averaging graph. Subsequently, the ensemble of PTrees is developed based on the variable order from the model-averaging graph. The final module involves evaluating the overall performance of PCF.

- The applicability and utility of the PCF framework are demonstrated through its application to multiple real-world healthcare datasets.

Existing causal inference methods each offer valuable capabilities: model-agnostic estimators excel at capturing treatment effect heterogeneity, deep learning approaches effectively handle high-dimensional confounding, and structural causal models provide a principled foundation for reasoning across interventions and counterfactuals. However, these approaches often fall short in clinical decision support due to limitations in interpretability, scalability, or support for path-specific and symbolic reasoning. PCF addresses this gap by integrating the graphical structure of Causal Bayesian Networks with the local interpretability of Probability Trees, enabling transparent, patient-level inference. By supporting interventional and counterfactual analysis within a scalable, modular framework, PCF aims to provide clinicians with both the rigour and clarity needed for practical causal reasoning.

The structure of the paper is as follows. Section 2 provides an overview of the relevant literature, establishing the context and motivation for this work. Section 3 describes the proposed framework in detail, including the steps involved in its construction and implementation. Section 4 presents the application of the framework to three distinct clinical case studies, while Section 5 discusses the results obtained from these applications. Finally, Section 6 provides conclusive remarks and directions for future work.

2. Relevant works

2.1. Traditional ML vs. Causal ML

Traditional ML has become a cornerstone in healthcare for tasks such as risk prediction, patient stratification, and outcome forecasting [10,11]. These models are highly effective at identifying patterns and correlations, enabling predictions such as the likelihood of disease onset or hospital readmissions [12–14]. However, traditional ML lacks the ability to answer counterfactual questions or estimate treatment effects, as it is inherently focused on predictive accuracy rather than causal reasoning [15,16].

In contrast, causal ML seeks to address "what if" questions by estimating the causal effect of interventions on outcomes. For instance, rather than simply predicting the probability of diabetes onset, causal ML can assess how this probability might change if a patient adopts a new medication or lifestyle modification [17,18]. These capabilities enable healthcare providers to explore counterfactual scenarios and make data-driven decisions that go beyond prediction to inform treatment planning and resource allocation.

Causal ML requires additional considerations compared to traditional ML, such as the need to account for unobservable outcomes and confounding variables. For example, the "fundamental problem of causal inference" [1,19] states that only factual outcomes under a given treatment are observed, while counterfactual outcomes remain unobserved. Estimating causal effects, therefore, requires more than just assumptions; it depends on having access to a sufficient set of measured covariates that enable identification, for instance, through adjustment sets that block all backdoor paths between treatment and outcome. Moreover, causal models must consider both direct and mediated (indirect) effects, as interventions may propagate through the system in complex ways. For example, a smoking cessation program might influence diabetes risk indirectly through changes in body mass index (BMI), necessitating a structured causal framework to capture such dependencies.

2.2. Applications of CBNs in healthcare

CBNs are a widely used tool in causal ML for modelling relationships among variables through a directed acyclic graph (DAG) structure. CBNs allow for the incorporation of prior knowledge and probabilistic reasoning, making them particularly effective for understanding complex dependencies in healthcare data. For instance, Rajendran et al. [20] employed CBNs to integrate risk factor analysis in breast cancer research, facilitating early detection and risk stratification. Shahmirzalou et al. [21] applied CBNs to recurrent breast cancer data to predict survival outcomes and guide personalised treatment strategies. Similarly, Jang et al. [22] leveraged CBNs to model risks associated with radiation therapy, offering support for personalised oncology care.

CBNs have also been applied to explore relationships between cardiovascular risk factors and related conditions, as demonstrated by Ordovas et al. [23]. These examples highlight the versatility of CBNs in identifying risk factors, modelling disease progression, and simulating the effects of interventions. However, while CBNs excel at representing causal relationships, their construction often requires significant domain expertise and computational resources, which may limit scalability.

2.3. Probability Trees (PTrees) in causal modelling

Probability Trees (PTrees) offer an intuitive and sequential representation of probabilistic relationships. In the context of this work, we adopt a representation where nodes correspond to variable instantiations and branches are labelled by events or outcomes, aligning more closely with a Moore-style structure in automata theory. This design choice supports clearer causal interpretation by associating each decision point (node) with a specific variable and deferring probabilistic branching to the edges. While alternative formulations exist, such as Mealy-style trees where transitions depend on both current states and input labels, our choice is motivated by the need to represent sequential dependencies and interventions transparently. This structure is particularly suitable for modelling healthcare processes, where decisions unfold over time and are influenced by evolving clinical states [6].

Despite their simplicity and expressiveness, PTrees have received relatively less attention in the machine learning literature compared to CBNs or structural causal models. Ambags et al. [7] proposed a hybrid approach combining probabilistic fuzzy decision trees with causal reasoning, applying it in two medical case studies to demonstrate its potential for real-world applications. Traditional PTree construction, however, often depends on domain expertise to determine variable orderings. While expert input is valuable, such reliance can introduce subjectivity and inefficiency, and expert-defined orderings may not always align with empirical dependencies in the data.

The Probabilistic Causal Fusion (PCF) framework addresses these challenges by integrating ensembles of PTrees with causal orderings derived from CBNs. These orderings are not treated as absolute ground truth but as a data-driven basis that can be refined through calibration with clinical expertise. This hybrid design promotes consistency and reproducibility, while ensemble methods reduce overfitting and enhance robustness across heterogeneous patient cohorts.

2.4. Advances in causal ML for healthcare

Recent advances in causal ML demonstrate its potential for improving healthcare decision-making by estimating treatment effects and simulating counterfactual scenarios. These methods have been applied to a variety of clinical challenges, from estimating the effects of medication adherence on diabetes progression to predicting survival probabilities under different oncology treatment plans [2,5]. By integrating causal ML techniques into clinical workflows, researchers aim to bridge the gap between prediction and actionable insights, enabling data-driven strategies for improving patient outcomes.

Building on this foundation, emerging work is exploring how causal reasoning can be integrated with high-capacity deep learning models. Nan et al. [24] demonstrate how visual diagnostic patterns used by expert pathologists can be captured by neural networks in a causally interpretable manner, improving transparency in high-dimensional histopathology tasks. Complementary advances in causal representation learning [25] propose methods for isolating relevant features and estimating counterfactual outcomes in complex image-based settings, offering pathways to more generalisable and robust clinical models.

However, causal ML comes with its own challenges. Assumptions about unmeasured confounding, model scalability, and the reliability of observational data remain critical concerns [26]. Addressing these limitations through frameworks like PCF, which combine causal reasoning with robust predictive modelling, represents an important step towards leveraging causal ML in diverse healthcare contexts.

2.5. Comparison with existing causal inference methods

Contemporary causal inference approaches can be broadly classified into model-agnostic estimators, deep representation learning models, and structural causal models, each with distinct strengths and limitations. The PCF framework is designed to bridge methodological and practical gaps across these paradigms.

First, non-parametric ITE estimators such as Causal Forests [27] and Bayesian Additive Regression Trees (BART) [28,29] provide flexible estimation of treatment heterogeneity under the unconfoundedness assumption. While effective at modelling complex treatment-response relationships, these methods lack model-inherent support for explicit counterfactual reasoning or principled interventional analysis (e.g., via do-calculus).

Second, deep causal representation learning methods, including TARNet and Counterfactual Variational Autoencoders (CEVAE) [30,31], mitigate confounding by learning latent representations. These models are particularly suited for high-dimensional observational data, but they often operate as black boxes, offering limited interpretability and no explicit support for tracing causal pathways or executing symbolic interventions. Moreover, their architectures are typically restricted to binary or single-shot treatments, limiting their applicability in scenarios involving multi-valued or sequential interventions, as commonly encountered in clinical settings.

Third, Structural Causal Models (SCMs) and CBNs [1,32] provide a rigorous formalism for causal reasoning across all three rungs of Pearl's causal hierarchy. However, their expressiveness comes at a cost: SCMs typically require predefined structural equations or strong assumptions, such as causal sufficiency and faithfulness, which can be difficult to validate or scale to high-dimensional, mixed-type datasets. The PCF framework also inherits these assumptions through its reliance on CBNs, but it avoids the need for structural equations by using a modular, tree-based construction that improves computational tractability and interpretability.

Building on this foundation, PCF combines the structural guidance of CBNs with a probabilistic tree-based architecture that preserves conditional dependencies while enabling efficient computation of interventional distributions, $P(Y \mid do(X))$. Unlike SCMs, it does not require full specification of structural functions, and it supports counterfactual inference via a twin-tree construction [6], facilitating patient-level "what if" reasoning in a transparent format. Through sparsity constraints and pathwise interpretability, PCF retains the semantic clarity of graphical models while offering the scalability and flexibility demanded by real-world healthcare applications.

In essence, PCF occupies a practical middle ground between symbolic and statistical causal methods, combining the interpretability of graphical models with the non-parametric adaptability of tree-based learners, while addressing core limitations of existing approaches in clinical decision support.

3. Methodology-PCF framework

We introduce PCF, a hybrid framework that integrates CBNs with ensembles of PTrees to support prediction, intervention modelling, and counterfactual analysis. Traditional PTree construction often depends on expert-defined variable orderings and empirical transition estimates, an approach that is labour-intensive, prone to variation across studies, and difficult to standardise. PCF addresses these limitations by using structural dependencies derived from CBNs to generate a data-driven variable ordering that reflects conditional independencies and provides a reproducible scaffold for tree construction. Importantly, this ordering defines a partial hierarchy rather than a rigid sequence: it constrains only those variables for which directional dependencies are inferred, while leaving flexibility for reordering or excluding variables that are weakly connected or clinically irrelevant. This design enhances interpretability and computational efficiency without compromising causal consistency. Clinicians may still refine the ordering locally to enhance interpretability, but without the need to redesign the global structure.

In parallel, PCF addresses several practical challenges of CBNs. While CBNs offer a principled framework for capturing causal dependencies, their global structures can be difficult to interpret at the patient level and are sensitive to variability in structure learning. PCF mitigates these issues by averaging across multiple candidate graphs to highlight stable dependencies, and by embedding these dependencies into PTrees, where they appear as transparent, rule-based pathways. This hybrid design improves interpretability and robustness, while ensemble learning further strengthens predictive performance by balancing bias and variance across heterogeneous datasets. Although PCF does not resolve the fundamental problem that no algorithm can guarantee full recovery of the true causal structure, it provides a more stable and computationally efficient scaffold for building clinically meaningful models. Fig. 1 outlines the key stages of the PCF pipeline, including data preprocessing, CBN generation, model-averaging graph construction, and PTree ensemble development.

3.1. Causal Bayesian Network (CBN) construction

PCF begins by constructing CBNs to uncover causal relationships among variables and identify directed pathways influencing the target outcome Y. These causal structures form the foundation for building PTrees, which translate causal dependencies into explicit decision paths. Formally, let $\mathcal{V} = \{X_1, X_2, \dots, X_n, Y\}$ denote the set of observed variables, with Y as the outcome of interest. A CBN is defined as a DAG $G = (\mathcal{V}, \mathcal{E})$, where nodes represent variables and edges $(X_i \to X_j) \in \mathcal{E}$ encode potential causal influences. The joint distribution over variables factorises according to the DAG structure as defined in Eq. (1), where $\mathrm{Pa}(X_i)$ denotes the set of parent variables of X_i in G. The resulting topology defines a partial ordering that reflects the inferred causal hierarchy.

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

CBNs were chosen as the foundation of PCF due to their theoretical capacity to encode and reason about causal dependencies under well-established assumptions, including causal sufficiency, the Markov condition, and faithfulness. Their DAG structure allows for explicit representation of conditional independencies, enabling interventional and counterfactual inference through graphical separation and do-calculus.

The construction of the CBN proceeds through three stages: structure learning, model averaging and topological sorting.

Structure learning. The construction of the CBN begins with structure learning, implemented through a suite of established algorithms: Hill Climbing (HC) [33,34], TABU Search [35], SaiyanH [36], Model-Averaging Hill Climbing (MAHC) [37], and Greedy Equivalence Search (GES) [38]. Each of these algorithms explores candidate Directed Acyclic Graphs (DAGs) in order to identify the network structure G that maximises a predefined scoring function $S(G \mid \text{data})$. Formally, the optimisation problem can be expressed as:

$$G = \arg\max_{G} S(G \mid \text{data}), \tag{2}$$

where the score S evaluates the goodness of fit between the candidate graph and the observed data.

Structure learning is performed using the open-source Bayesian network structure learning system Bayesys [39], which supports target-aware search, focusing the discovery process on variables with direct influence over the outcome Y. The selected algorithms are chosen for their complementary strengths: HC is computationally efficient but prone to local optima; TABU introduces diversification strategies to escape local minima; SaiyanH integrates heuristic biasing; MAHC stabilises solutions through repeated sampling; and GES scales well to high-dimensional data, though it could be less effective on small samples. Also, these algorithms are selected for their capability to incorporate the target variable during model construction and to handle diverse knowledge approaches, including direct relationships and forbidden edges [40]. While the selected score-based algorithms learn DAGs that serve as approximations of the underlying causal structure under standard assumptions (causal sufficiency, faithfulness, and acyclicity), they do not provide formal statistical guarantees. These methods are, however, effective for constructing interpretable causal scaffolds to support downstream counterfactual analysis. Incorporating statistical testing to validate or refine edge inclusion represents an important direction for future enhancement of the framework.

Given the inherent uncertainty in edge direction, especially in the absence of domain priors, we apply a model averaging strategy to mitigate algorithmic bias. By generating multiple candidate structures and identifying recurring patterns across them, we construct a consensus graph highlighting stable causal relationships. This ensemble-based approach provides a more reliable and interpretable foundation for downstream PTree construction.

Model averaging. To integrate the outputs from multiple structure learning algorithms, we construct an averaged network G_{avg} using Bayesys. This consensus graph captures only the most consistently inferred relationships, reducing the influence of algorithm-specific variance and improving structural reliability. Edges are included in G_{avg} based on their frequency of occurrence across the candidate graphs, subject to the following criteria:

a. **Directed Edges**: Add directed edges $e = (u, v)$ to G_{avg} starting with the edges that occur most frequently across input graphs, ensuring no cycles are formed:

$e \in G_{\text{avg}}$ if $\mathrm{freq}(e) >$ threshold and no cycle

If adding an edge e would create a cycle, reverse the edge:

$e \to e^{-1}$ if e forms a cycle

b. **Undirected Edges**: Add undirected edges $e = \{u, v\}$ to G_{avg}, skipping those already added as directed edges:

$e \in G_{\text{avg}}$ if $\mathrm{freq}(e) >$ threshold

c. **Cycle Handling**:Add directed edges from the cycle-inducing edge set C:

$e \in G_{\text{avg}}$ if $e \in C$ and occurs frequently

The model-averaging procedure draws on ensemble methods in causal discovery, using a majority-vote rule to decide edge inclusion and orientation while keeping the graph acyclic. Instead of relying on a single algorithm, it highlights dependencies that recur across several candidate graphs, producing a representative structure. The resulting consensus DAG therefore reflects recurring patterns across multiple, and often incompatible, models, and aligns with strategies used in other relevant studies [41–43]. Although no single procedure can guarantee recovery of the full causal data-generating process, the consensus graph preserves the strongest and most consistently inferred dependencies, thereby retaining meaningful causal content. This approach reduces algorithm-specific variability, improves robustness, and provides a stable foundation for downstream PTree construction. Building on this foundation, the PCF framework extends the practical utility of CBNs in two key ways. First, it stabilises structural learning through model averaging, as outlined above, ensuring that only reliably inferred dependencies inform downstream inference. Second, it overlays this consensus structure with a Probability Tree, enabling modular and interpretable simulations of interventional and counterfactual scenarios. While PCF does not resolve challenges such as unmeasured confounding, an inherent limitation of observational causal inference, it offers a systematic mechanism for incorporating domain expertise post hoc. This allows for refinement of the causal structure through informed reordering or exclusion of variables. In this respect, PCF does not aim to expand formal identifiability, but rather to enhance the framework's practical applicability through improved robustness, interpretability, and clinician-guided adaptability.

***Topological sorting.*.** Once the consensus graph G_{avg} is constructed, we derive a topological ordering π such that:

$$\pi = \text{topological\_sort}(G_{\text{avg}}) \tag{3}$$

This results in a partial order over the variable nodes, reflecting the causal dependencies encoded in the DAG: each variable appears after its parents in the ordering. Among different available approaches [44, 45], we employ Kahn's algorithm for its efficiency and suitability for moderately sized DAGs.

By deriving the ordering from data rather than relying solely on expert-defined hierarchies, this step reduces subjectivity and preserves subtle yet meaningful dependencies captured during structure learning. Although clinicians may refine the ordering locally to enhance interpretability, PCF adheres to the partial ordering constraints inferred from the causal structure, ensuring that any such modifications remain consistent with the DAG. The variable order π serves as the backbone for PTree construction, guiding branching decisions and contributing to both predictive accuracy and interpretability of the final ensemble model.

Sensitivity analysis. Sensitivity analysis is performed to assess the responsiveness of nodes to changes in their parent and ancestor nodes [46]. For a node X_i with parent X_j, sensitivity S is defined as:

$$S = \frac{\partial P(X_i)}{\partial \theta_{X_j}} \tag{4}$$

where θ_{X_j} represents the parameters in the Conditional Probability Table (CPT) of X_j. High sensitivity indicates that small changes in θ_{X_j} result in significant changes in the posterior distribution of X_i, suggesting a strong dependency. Conversely, low sensitivity implies that large changes in θ_{X_j} have minimal impact on X_i's distribution, indicating a weak dependency.

The posterior probability T of the selected state of the target node, given the parameter p, is represented by the following general linear rational functional form:

$$T = \frac{a \cdot p + b}{c \cdot p + d} \tag{5}$$

The sensitivity analysis algorithm calculates the coefficients a, b, c, and d. The derivative, which is the basic measure of sensitivity, is given by:

$$D = \frac{a \cdot d - b \cdot c}{(c \cdot p + d)^2} \tag{6}$$

The denominator is positive, indicating that the sign of the derivative is constant for all values of p. By substituting 0 and 1 for p (noting that p is a probability), we can calculate the range within which the posterior will change if p is modified across its entire range, defined by:

$$p_1 = \frac{b}{d} \tag{7}$$

$$p_2 = \frac{a+b}{c+d} \tag{8}$$

Sensitivity analysis is crucial in understanding the stability and robustness of the model. It helps identify the most influential parameters in the network, guiding targeted interventions and enhancing the interpretability of the model. We use the GeNIe BN software [47] to perform this analysis.

3.2. Probability tree construction

Building on the causal structure defined by the model-averaged graph G_{avg}, we construct PTrees through a three-step process designed to support interpretable and data-efficient prediction. PTrees translate the causal relationships identified by CBNs into explicit, rule-based paths that enable instance-level reasoning. While CBNs excel at probabilistic inference, their global structure can obscure the logic behind individual predictions. PTrees, by contrast, provide a transparent framework that clinicians can interpret. Formally, each node n in a PTree is defined as a tuple $n = (u, S, C)$, where u is a unique identifier, S is a list of variable assignments, and C is an ordered set of transitions $\{(p_m, m)\}$, with $p_m \in [0, 1]$ denoting the transition probability to child node m. These probabilities satisfy $\sum p_m = 1$. The root node has no parent; leaves have empty transition sets. A complete path from root to leaf constitutes a full realisation, whose joint probability is computed as the product of transitions along that path:

$$P(\text{realisation}) = \prod_{i=1}^{k} p_i. \tag{9}$$

At each node, the variables in S are assigned concrete values, conditioning all subsequent transitions. To avoid exponential tree growth, we employ (1) the CBN-derived topological ordering π to guide variable splits, (2) pruning based on transition probabilities to remove low-support branches, and (3) ensemble learning on bootstrapped samples to maintain tractability while improving generalisability. This combination balances interpretability, robustness, and scalability, key issues for real-world clinical decision support.

The process for building PTrees comprises three main steps: creating the tree from the input data, ensemble learning step and the prediction process.

Create tree from data. The process commences with Algorithm 1 (CREATE_TREE_FROM_DATA), which outlines the construction of a PTree from a given dataset and the variable order derived from the previously learned CBN. The ordering π ensures that nodes are expanded in a sequence consistent with the conditional dependencies encoded in the DAG G.

a. **Root Level:** A PTree T is initialised from dataset D using the variable order pi. At the root, data are partitioned according to the target variable Y, and empirical marginal probabilities are computed as:

$$p(v) = \frac{\text{Count}(v)}{\text{Total Samples}} \tag{10}$$

where $p(v)$ denotes the marginal probability for value v of Y. This initial partitioning establishes a probabilistic baseline for subsequent splits.

b. **Transition Probabilities:** For each subsequent node, conditional probabilities are calculated using the parent–child relationships defined in the CBN:

$$P(X_i \mid \text{Pa}(X_i)) = \frac{\text{Count}(X_i, \text{Pa}(X_i))}{\text{Count}(\text{Pa}(X_i))} \tag{11}$$

where $\text{Pa}(X_i)$ are the parent variables of X_i in G. This differs from conventional PTrees, where splits depend solely on local frequency counts. By constraining probability estimation to parent–child relationships encoded in the DAG, the PTree explicitly incorporates causal dependencies inferred from the CBN. This ensures that branches reflect not just statistical associations but dependencies consistent with the learned causal structure.

c. **Pruning:** To prevent overfitting, branches with probabilities below a pruning threshold θ are removed. Rather than fixing θ globally, we optimise it for each dataset using pruning-curve analysis. This involves evaluating predictive performance across a range of θ values and selecting the point that balances model complexity with generalisation. In this way, pruning retains branches that provide meaningful conditioning for subsequent variables, while discarding those that contribute little beyond noise.

Ensemble learning. A key innovation of PCF is the use of ensemble learning to enhance robustness and generalisability. A single PTree may overfit to local data patterns or reflect sampling variability, particularly in high-dimensional healthcare datasets. By combining multiple PTrees trained on different data partitions, the ensemble reduces variance while preserving the interpretability of individual trees.

The ensemble process is implemented through the ENSEMBLE_PROBABILITY_TREES function, which follows a three-step strategy:

a. **Data Splitting:** The dataset D is partitioned into k disjoint subsets $\{D_1, D_2, \ldots, D_k\}$, through stratified folds to maintain class balance. This splitting procedure is repeated over multiple runs with different random seeds, ensuring each tree in the ensemble is trained on slightly varied distributions. This replication introduces diversity among trees and helps mitigate overfitting, while still maintaining representative class proportions in each fold.

b. **Tree Construction:** For each subset D_i, a PTree T_i is constructed using Algorithm 1 and the CBN-derived variable order π:

$$T_i = \text{CREATE\_TREE\_FROM\_DATA}(D_i, \pi).$$

Because each tree shares the same causal ordering but is trained on different subsets, the ensemble captures cohort-specific variations while maintaining structural consistency across models.

c. **Model Aggregation:** The root nodes of the ensemble are stored in a list `ptrees` for inference. Predictions from individual trees are later aggregated, allowing the ensemble to smooth out idiosyncratic errors from individual PTrees and yield more stable outputs.

Prediction process. Predictions are generated by aggregating outputs from all PTrees in the ensemble, a strategy that reduces variance while retaining interpretability.

a. **Individual Predictions:** For each PTree T_i, a prediction $\hat{y}_i$ is obtained by traversing the tree according to the attribute values of the input instance:

$$\hat{y}_i = \text{Predict}(T_i, \text{instance}). \tag{12}$$

Each tree reflects both the causal ordering derived from the CBN and the statistical patterns in its training subset D_i. Algorithm 2 is used to compute the conditional probability of a class given the observed feature conditions, ensuring that local predictions are consistent with the inferred causal dependencies.

b. **Aggregation:** Predictions are combined across the k ensemble members to obtain a consensus estimate:

$$\hat{y}_{\text{avg}} = \frac{1}{k}\sum_{i=1}^{k} \hat{y}_i. \tag{13}$$

This averaging smooths out biases introduced by individual PTrees and provides a more calibrated probability estimate, particularly important in heterogeneous healthcare data where single-model predictions may be unstable.

c. **Threshold Classification:** A final class label is assigned by comparing $\hat{y}_{\text{avg}}$ against a threshold τ:

$$\text{Class} = \begin{cases} \text{Positive} & \text{if } \hat{y}_{\text{avg}} > \tau, \\ \text{Negative} & \text{otherwise.} \end{cases}$$

The threshold τ is not fixed a priori but can be tuned through cross-validation to maximise application-specific performance metrics, or adjusted post hoc to prioritise sensitivity over specificity in high-risk clinical screening tasks.

This ensemble-based inference procedure balances bias and variance across partitions, produces calibrated probability estimates, and offers an interpretable mechanism for both prediction and uncertainty quantification.

SHapley additive explanations (SHAP). To enhance model interpretability and elucidate feature importance, SHAP [48] was integrated into the framework. SHAP values provide a unified measure of feature importance, enabling us to understand the contribution of each feature to the model's predictions. This enhances the transparency and trustworthiness of our model's outputs. Rooted in cooperative game theory, SHAP values offer a method to attribute the difference between the prediction for a specific instance and the average prediction to individual features. SHAP values adhere to local accuracy, missingness, and consistency, ensuring reliable and interpretable explanations. Mathematically, for a model f and an instance x, the SHAP value ϕ_i for feature i is calculated as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)]$$

where:

- N is the set of all features.
- S is a subset of N that excludes feature i.
- $f(S)$ is the prediction for the instance with only the features in S.

This formula calculates the average marginal contribution of feature i across all possible feature subsets, ensuring fair and comprehensive feature importance attribution.

To facilitate SHAP analysis, predictions from the ensemble of PTrees were encapsulated in a wrapper compatible with the SHAP framework. A background dataset was generated using k-means clustering on the training data to provide a reference point for SHAP value calculations. SHAP values were computed for a subset of the test data using the Kernel Explainer to balance computational efficiency and accuracy

4. Case studies

The proposed framework addresses several limitations of traditional prediction models by offering a multifaceted approach for clinicians. It facilitates the identification of causal relationships between variables, enables predictive modelling, and supports the exploration of potential interventions and counterfactual scenarios. This combination provides clinicians with a more comprehensive understanding of the data while acknowledging the inherent challenges of causal analysis.

We evaluated the framework using multiple real-world healthcare datasets to assess its applicability and generalisability across diverse clinical contexts. The first dataset was MIMIC-IV, a collection of electronic health records from critical care settings, where the objective was to predict the length of stay in the Intensive Care Unit (ICU). The second dataset was the Framingham Heart Study, which focuses on cardiovascular disease (CVD) and its risk factors, trends over time, and familial patterns. Finally, the Diabetes dataset from BRFSS-2015 was used to analyse risk factors and predict the likelihood of diabetes onset. These case studies were intentionally selected to span distinct domains, acute care, chronic cardiovascular conditions, and metabolic disease, demonstrating the framework's robustness and generalisability beyond a single clinical setting. This diversity enables an assessment of PCF's adaptability to varied healthcare challenges and strengthens its relevance for broader medical applications.

To ensure valid causal inference, the PCF framework operates under three foundational assumptions. First, it assumes causal sufficiency, that is, relevant confounders affecting both treatment and outcome are either directly observed or adequately captured within the CBN. Second, the framework relies on the faithfulness assumption, which posits that observed statistical dependencies in the data reflect the underlying causal structure. Third, it assumes that the CBN itself is a valid representation of the data-generating process, whether specified by experts or learned from data. These assumptions are necessary for the PCF model to support interpretable interventional and counterfactual analyses that aim to capture plausible causal mechanisms rather than spurious associations.

Algorithm 1 Create Tree from Data

```
Require: $father\_node$: Node, $data$: DataFrame, $variable\_order$: List,
    $level$: Integer, $pruning\_threshold$: Float
Ensure: Root node of the decision tree ($father\_node$)
function CREATE_TREE_FROM_DATA($father\_node$, $data$, $variable\_order$,
    $level$, $pruning\_threshold$)
    $current\_variable \leftarrow variable\_order[0]$
    $current\_data \leftarrow data[current\_variable]$
    $class\_nodes \leftarrow$ empty list
    for $val$ in unique values of $current\_data$ do
        Create class node with ID, level, statements, and no children
        Append class node to $class\_nodes$
    end for
    $total\_samples \leftarrow$ total number of samples in $current\_data$
    for each $class\_node$ and $val$ in $class\_nodes$ do
        $is\_root \leftarrow$ Check if $father\_node$ is root
        $val\_count \leftarrow$ Count of $val$ in $current\_data$
        if $is\_root$ then
            Calculate transition probability based on occurrences
            if $transition\_prob \geq pruning\_threshold$ then
                Insert transition probability into $father\_node$
            end if
        else
            Calculate transition probability based on parent's state
            if $transition\_prob \geq pruning\_threshold$ then
                Insert transition probability into $father\_node$
            end if
        end if
    end for
    for each $class\_node$ and $val$ in $class\_nodes$ do
        Get next data for $val$
        Get next variable order
        if next variable order is not empty then
            Recursively call $create\_tree\_from\_data$
        end if
    end for
    return $father\_node$
end function
```

Algorithm 2 Conditional Probability Calculation

```
function CONDITIONALPROBABILITY($self$,$input\_condition$)
    if $input\_condition$ is empty then
        return 0.0
    end if
    $cut\_disease \leftarrow$ self.prop('target_variable')
    $combined\_cut \leftarrow$ None
    for all ($var, val$) in $input\_condition$ do
        $cut \leftarrow$ self.prop(var + ' = ' + val)
        if $combined\_cut$ is None then
            $combined\_cut \leftarrow cut$
        else
            $combined\_cut \leftarrow combined\_cut \wedge cut$
        end if
    end for
    $disease\_see \leftarrow$ self.see($combined\_cut$)
    $probability \leftarrow disease\_see.prob(cut\_disease)$
    return $probability$
end function
```

4.1. Length of stay in ICU case study

The intensive care unit (ICU) stands as a vital line of defence for critically ill patients, offering specialised care to prevent deterioration from severe illness or injury [49,50]. However, the ever-increasing demand for ICU beds threatens this critical service. The imbalance between ICU capacity and patient needs has significant consequences for patient outcomes, public health, and even socio-economic factors [51,52]. Therefore, optimising ICU resource allocation and planning for future needs necessitates interpretable models that facilitate counterfactual analysis for informed decision-making, ultimately ensuring optimal care for critically ill patients.

4.1.1. MIMIC-IV

In this study, we use the MIMIC-IV version 2.2 database [53], which includes patients admitted to the BETH Israel Deaconess Medical Center during the period 2008–2019. The data contains multiple dimensions, from administrative data to laboratory results and diagnoses. We employed preprocessing techniques as described in [54] to ensure consistency and comparability with existing literature. The cohort included all patients with at least one ICU visit. However, certain subsets of patients were excluded: those who died during their ICU stay, those who returned to the ICU within 48 h of discharge, those with an LOS greater than 21 days, and those with an LOS of less than one day.

The exclusion of patients who returned to the ICU within 48 h was motivated by the focus of this analysis on understanding factors influencing the initial ICU stay and its length. Rapid readmissions often reflect distinct cases with underlying complexities such as incomplete recovery or premature discharge, which could introduce confounding factors. Similarly, patients with extremely long LOS (greater than 21 days) were excluded to avoid the influence of outliers, which could disproportionately impact model performance. Patients with an LOS of less than one day were excluded because the data collected during the first 24 h was used for modelling, making such cases incomplete for analysis. These exclusions ensure that the cohort is representative of the broader ICU patient population, allowing for more generalisable findings. Future work could investigate the effects of these exclusions on model performance by reintroducing these subsets for a comparative analysis.

To transform the length-of-stay task into a classification problem, we categorised LOS into clinically meaningful groups: short stays (1–4 days) and long stays (greater than 4 days). This categorisation was guided by the 75th percentile of LOS distribution (Q3 = 4.0) in the dataset, as described in [54], and reflects thresholds commonly used in critical care practice.

4.2. Heart disease case study

Despite significant advancements in healthcare, CHD remains a leading cause of global mortality, accounting for 17.9 million deaths in 2019 as reported by the World Health Organization (WHO) [55]. While accurate prediction of future risk is undeniably crucial, medical experts increasingly recognise the limitations of solely relying on such prognostic models. To optimise patient care, a deeper understanding of the factors influencing individual susceptibility to CHD is paramount. This necessitates the development of intelligent systems that can not only predict future risk but also explore the potential impact of interventions and counterfactual analysis.

4.2.1. Framingham data

In this study, we use the Framingham heart disease dataset includes over 4238 records and 15 attributes [56]. The goal of the dataset is to predict whether the patient has 10-year risk of future CHD. The initial preprocessing steps involved converting the numerical variables in the dataset into categorical variables. Given that the variables pertain to health-related data, specific ranges were meticulously considered during this conversion process. Numerical data representing health metrics such as blood pressure, cholesterol levels, or body mass index were categorised into clinically relevant ranges indicative of different health conditions or risk levels. By transforming numerical data into categorical form based on meaningful health-related ranges, the dataset became better suited for subsequent analysis and interpretation within the context of healthcare applications.

4.3. Diabetes case study

Despite the existence of preventative measures, diabetes remains a significant global health burden [57]. Characterised spectrum of devastating complications, it necessitates a multifaceted approach that transcends traditional risk prediction. While accurate future risk prediction remains valuable for preventative strategies, a deeper understanding of modifiable factors influencing individual susceptibility is paramount, especially considering the potential for early intervention to reduce diabetes-related mortality [58]. This necessitates the development of robust computational models capable of not only predicting future risk but also exploring the potential impact of various interventions through counterfactual analysis. Such models could empower clinicians by enabling the exploration of "what-if" scenarios: investigating how a patient's risk profile might change with different lifestyle modifications or therapeutic interventions, ultimately leading to tailored preventative strategies and optimised patient care.

4.3.1. Diabetes data

Data was obtained from the Behavioural Risk Factor Surveillance System (BRFSS), which is the primary system of health-related telephone surveys that collect state data on risk behaviours, chronic health conditions, and use of preventative treatments amongst U.S. residents [59]. The survey started in 1984 and currently performs over 400,000 adult interviews each year, making it the world's largest continuously conducted health survey system. This survey data provides a dataset that could be used to analyse and forecast diabetes risk variables. We utilised the BRFSS-2015 dataset, which included 253,680 health assessments.

5. Evaluation and discussion of the results

The evaluation process begins in Section 5.1 with an analysis of the varying outcomes derived from the sensitivity analysis. In Section 5.2, we assess the predictive performance of the PCF model, comparing it against a range of benchmark models, including both interpretable and non-interpretable methods, across all three datasets. Additionally, this section explores model interpretability using SHAP. Section 5.3 then examines the effects of potential interventions through interventional analysis. Lastly, Section 5.4 investigates counterfactual analysis to provide further insights.

5.1. Interpretation of sensitivity analysis

Sensitivity analysis is a crucial step in our framework to understand the influence of various parameters on the target variable. The diverse outcomes from our sensitivity analysis provide valuable insights into the multifaceted factors influencing LOS, CHD, and Diabetes, as depicted in Fig. 2. The colour of the bars indicates the direction of change in the target state, with red representing a negative impact and green representing a positive impact. For LOS, factors such as first care unit admission, patient's verbal communication ability, and specific diagnoses (e.g., Respiratory system, Circulatory system) show high sensitivity, indicating their collective substantial impact on LOS. In the context of CHD, the absence of diabetes and hypertension was found to significantly reduce the risk. Additionally, other significant factors include being a non-smoker with normal systolic blood pressure and specific education levels. These findings highlight the combined effect of lifestyle and socio-economic factors on CHD risk. For Diabetes, the sensitivity analysis demonstrates the complex interplay between hypertension, cholesterol levels, BMI, and other health indicators. The figure reveals that individuals with hypertension have the highest sensitivity value, indicating that high blood pressure significantly increases the risk of developing diabetes. Additionally, other influential factors include high cholesterol, elevated BMI, and the presence of heart disease. These findings underscore the complex interplay of multiple health conditions in determining diabetes risk, highlighting the necessity of addressing various health parameters simultaneously to effectively manage and prevent diabetes.

Table 1
Results using SMOTE for MIMIC-IV.

Algorithm	Accuracy	Specificity	Sensitivity	AUC-ROC
Ensemble Algorithms				
GB	73.71	77.96	57.91	67.94
XGB	73.92	79.14	54.54	66.84
Adaboost	75.07	79.78	57.57	68.67
RF	75.64	83.22	47.47	65.35
PCF	**73.21**	**77.69**	**56.56**	**67.13**
Other Algorithms				
SVM	73.07	76.42	60.60	68.51
KNN	71.14	77.87	46.12	62.00
Interpretable Algorithms				
DT	79.42	88.84	44.44	66.64
LR	70.14	73.70	56.90	65.30
PTree	80.29	91.11	40.06	65.59

5.2. Prediction

This section evaluates the predictive capabilities of PCF by applying it to three clinical datasets. We conduct a comprehensive assessment of its performance by juxtaposing its outcomes against those attained by established benchmark methodologies, comprising Logistic Regression (LR) [60], Decision Tree (DT) [61], Random Forest (RF) [62], Support Vector Machine (SVM) [63], K-Nearest Neighbours (KNN) [64,65], Gradient Boosting (GB) [66,67], eXtreme Gradient Boosting (XGB) [68], and Adaptive Boosting (Adaboost) [69]. Noteworthy among these benchmarks are LR, DT, and KNN, esteemed for their interpretability, which facilitates the elucidation of their decision-making mechanisms. Additionally, given the ensemble nature of our approach involving PTrees, a comparison with ensemble techniques such as GB, XGB, Adaboost and RF is warranted.

The dataset undergoes stratification-based partitioning into training and testing subsets to ensure their representativeness. Subsequently, the performance of each model is assessed utilising diverse metrics such as accuracy, specificity, sensitivity, and the Area Under the Receiver Operating Characteristic Curve (AUC–ROC). Predictions are made based on a default threshold, with the potential for adjustment in scenarios characterised by resource constraints, thereby prioritising cases of utmost urgency.

The performance metrics for each dataset are presented in Tables 1, 2 and 3. It is well-documented that class imbalance within datasets can significantly impact the evaluation of machine learning algorithms. To mitigate this potential bias and ensure a fair comparison across all models, various techniques for handling class imbalance were employed. In the case of the MIMIC-IV and Framingham heart datasets, the SMOTE (Synthetic Minority Over-Sampling Technique) [70] oversampling technique yielded superior results. Conversely, ADASYN (Adaptive Synthetic Minority Oversampling Technique) [71] demonstrated the best performance when applied to the diabetes dataset. This finding suggests that the most effective class imbalance handling technique may vary depending on the specific characteristics of the data and the machine learning models being evaluated.

Tables 1, 2, and 3 present the performance evaluation of the PCF framework compared to benchmark methodologies across the three datasets. Notably, PCF achieves results that are largely comparable to established ensemble-based and interpretable models, balancing specificity and sensitivity while maintaining competitive predictive accuracy.

On the MIMIC-IV dataset, PCF achieves an accuracy of 73.21% and an AUC–ROC of 67.13%, performing similarly to ensemble-based methods such as Gradient Boosting (73.71% accuracy) and Adaboost (75.07% accuracy). Although DT achieves a higher accuracy of 79.42%, PCF demonstrates better sensitivity (56.56%) than simpler models like KNN, which achieves only 46.12%. Among ensemble methods, PCF

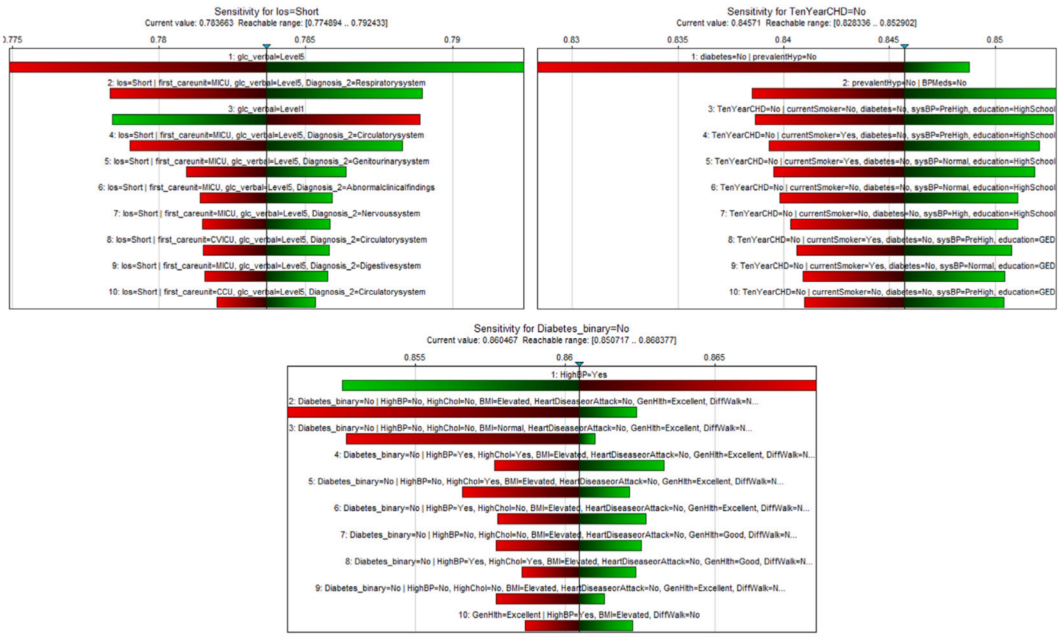

Fig. 2. Sensitivity Analysis for LOS, Diabetes, and Framingham datasets.

Table 2
Results using SMOTE for Framingham heart data.

Algorithm	Accuracy	Specificity	Sensitivity	AUC-ROC
Ensemble Algorithms				
GB	63.08	64.81	53.48	59.15
XGB	63.91	68.01	41.08	54.54
Adaboost	66.03	69.26	48.06	58.66
RF	67.09	72.73	35.65	54.19
PCF	**66.98**	**69.68**	**51.93**	**60.80**
Other Algorithms				
SVM	69.22	74.26	41.08	57.67
KNN	76.76	87.62	16.27	51.95
Interpretable Algorithms				
DT	64.03	66.06	52.71	59.38
LR	61.55	63.14	52.71	57.92
PTree	65.57	70.23	39.53	54.88

Table 3
Results using ADASYN for diabetes data.

Algorithm	Accuracy	Specificity	Sensitivity	AUC-ROC
Ensemble Algorithms				
GB	70.78	70.02	75.66	72.84
XGB	69.71	71.34	59.25	65.30
Adaboost	71.35	71.09	73.01	72.05
RF	73.42	77.45	47.61	62.53
PCF	**73.64**	**73.41**	**75.13**	**74.27**
Other Algorithms				
SVM	69.71	68.62	76.71	72.67
KNN	79.00	85.86	35.26	60.56
Interpretable Algorithms				
DT	62.92	60.19	80.42	70.31
LR	70.71	70.27	73.54	71.90
PTree	72.36	71.12	52.82	61.97

effectively balances specificity and sensitivity, avoiding extremes like RF, which prioritises specificity at the expense of sensitivity.

On the Framingham dataset, PCF achieves an AUC–ROC of 60.80%, outperforming most ensemble methods while maintaining competitive accuracy at 66.98%, compared to RF (67.09%) and Adaboost (66.03%). Although KNN achieves the highest accuracy at 76.76%, its significantly lower sensitivity (16.27%) highlights its limitations in handling balanced classification scenarios. PCF's balanced trade-off between specificity and sensitivity makes it particularly well-suited for datasets requiring nuanced predictions.

On the diabetes dataset, PCF achieves the highest AUC–ROC among all models (74.27%) and an accuracy of 73.64%, comparable to ensemble methods such as RF (73.42%) and higher than DT (62.92%). While KNN achieves the highest accuracy (79.00%), its specificity (85.86%) comes at the cost of sensitivity (35.26%). PCF balances both metrics effectively, achieving 73.41% specificity and 75.13% sensitivity, demonstrating its robustness across diverse datasets.

To confirm that PCF's observed performance was not the result of random variation, we conducted a non-parametric permutation test with 1000 iterations. In all datasets, the test yielded $p < 0.001$, indicating that PCF's accuracy is statistically significant and highly unlikely to have occurred under random label assignments. This strengthens the validity of the reported results and supports the reliability of PCF as a predictive model.

Overall, while PCF does not consistently outperform simpler models such as DT or KNN in terms of accuracy, it offers balanced performance across key metrics and demonstrates robustness across datasets. Its ability to maintain competitive predictive performance while also uncovering causal relationships and supporting intervention modelling sets it apart from purely predictive methodologies.

5.2.1. *Interpretability with SHAP*

The SHAP plot, shown in Fig. 3, interprets the influence of each feature on predictions for LOS, Coronary Heart Disease (CHD), and Diabetes. For LOS, features such as creatinine, first care unit, and urea nitrogen have high SHAP values, indicating their strong influence on prolonged ICU stays. The Glasgow Coma Scale (glc_verbal) score and elevated white blood cells, along with specific diagnoses (e.g., respiratory and circulatory system issues), also significantly impact LOS predictions. In CHD predictions, critical features include current smoking status, systolic blood pressure (sysBP), and glucose levels, which are known risk factors for heart disease. High cholesterol (totChol), the presence of diabetes, socio-economic factors, and lifestyle choices such as education level and physical activity further influence CHD risk. For Diabetes, key contributors include high blood pressure

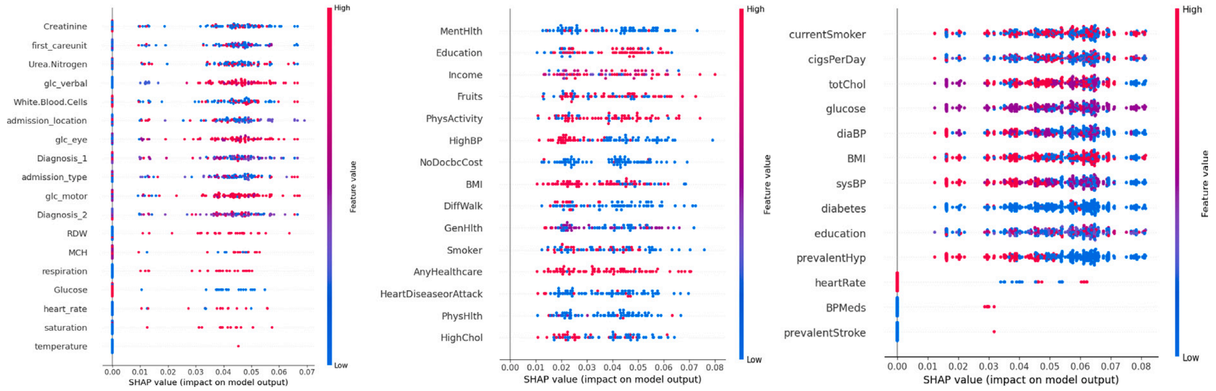

Fig. 3. SHAP plot showing feature impacts on predictions for LOS, CHD, and Diabetes.

(HighBP), elevated BMI, and high cholesterol levels, which are essential components of metabolic syndrome. General health status (GenHlth) and difficulty walking (DiffWalk) also play significant roles, along with the presence of heart disease and levels of physical activity. The SHAP plots reveal that LOS is heavily influenced by clinical indicators related to critical health conditions and specific ICU units. CHD risk is predominantly affected by cardiovascular risk factors, lifestyle choices, and socio-economic factors, while Diabetes risk is determined by metabolic health markers, overall health, and physical activity. These insights validate the results from sensitivity analysis and provide a detailed understanding of feature contributions, helping identify key areas for intervention and improve clinical decision-making processes by emphasising the most impactful factors for each health condition. By combining sensitivity analysis for understanding variable influence within the CBN and SHAP values for detailed prediction explanations, our framework ensures robust causal inference and clear, actionable insights for clinical decision-making.

5.3. Intervention

In the context of PCF, intervention involves the strategic modification of transition probabilities to ensure a specific event occurs with certainty (probability of 1). This approach allows for the exploration of conditional probabilities represented as $P(A \mid do(B))$, indicating the probability of event A occurring given that event B is enforced. Conceptually, an intervention reflects an externally imposed change, such as adjusting a patient's physiological or treatment variable, and estimates the resulting shift in outcome probabilities. For instance, $P(\text{recovery} \mid do(\text{early discharge}))$ represents the likelihood of recovery if early discharge were implemented, irrespective of factors that would normally influence discharge timing.

Unlike in CBNs, interventions in PCF are more general and do not require unique value assignments to manipulated random variables. Instead, the impact of an intervention depends on a critical set, defined as the minimal subset of nodes or branches in the probability tree whose transition probabilities must be modified to make the target event occur with certainty. In practical terms, the critical set identifies precisely where in the tree the intervention must act to achieve the desired causal outcome, while leaving unaffected branches unchanged.

Formally, let $\mathcal{T}$ denote a probability tree over variables $\mathcal{V}$, and let $E \subseteq \mathcal{V}$ be the event to be enforced. The *critical set* $C \subset \mathcal{T}$ is defined as the minimal collection of decision contexts at which modifying transition probabilities is sufficient to ensure $P(E) = 1$ under the intervened tree. The intervention proceeds as follows: (i) traverse the tree to locate all branches where E is not satisfied; (ii) identify the minimal set C of nodes where changes can block these violating paths; (iii) set the transition probabilities leading to incompatible paths to zero; and (iv) renormalise the remaining transitions locally to preserve probabilistic consistency.

All simulated interventions are restricted to clinically meaningful and actionable modifications. Only modifiable variables, such as heart rate, respiration rate, or laboratory measures, are considered. This ensures that the scenarios produced by PCF correspond to interventions that are both realistic and interpretable within real-world healthcare contexts.

MIMIC-IV:. To elucidate the impact of key physiological parameters on ICU length of stay (LOS), an interventional analysis was conducted. Eight critical factors were examined to assess the PCF model's ability to replicate established causal relationships. The primary objective was to assess PCFs ability to replicate established causal relationships between these parameters and LOS.Box plots (Fig. 4) were used to visualise the distribution of los across different intervention groups. In these plots, red signifies an increased probability of los exceeding 4 days (los = 1), while green signifies a decrease.

Existing literature establishes a link between bradycardia (heart rate ≤ 60 beats per minute) and extended ICU stays due to underlying medical conditions requiring further investigation or treatment [72]. To explore this relationship within our model, interventional analysis was conducted on heart rate. Simulating bradycardia ($do(heart\_rate = 0)$) significantly increased the likelihood of extended ICU stays ($los = 1$), consistent with established medical knowledge. However, the relationship between heart rate and length of stay is more complex. Similar trends were observed for heart rates above 60 bpm ($do(heart\_rate = 1)$), though to a lesser extent, and a slight decrease in $los = 1$ was noted for even higher heart rates ($do(heart\_rate = 2)$). This highlights the need to consider multiple physiological and clinical variables beyond heart rate.

In literature, it is established that low levels of urea nitrogen (UN) are not typically concerning, often associated with low protein intake [73]. Similar findings are observed in our model, where manipulating UN levels to be low ($do(Urea\_Nitrogen = 0)$) results in a decrease in the proportion of patients with extended ICU stays ($los = 1$). However, as the intervention values increase ($do(Urea\_Nitrogen = 1)$ and $do(Urea\_Nitrogen = 2)$), a trend emerges indicating a potential rise in the probability of $los = 1$. While this increase is subtle, it is visually detectable by the red colour in the box plots.

Literature indicates that elevated Red Cell Distribution Width (RDW) is closely associated with increased risk of cardiovascular morbidity and mortality in patients with previous myocardial infarction, potentially leading to prolonged hospital stays [74]. Our model's interventional analysis, where RDW is manipulated to be high ($do(RDW = 2)$), shows a corresponding increase in the percentage of patients with extended ICU stays ($los = 1$), consistent with existing literature.

Lower levels of creatinine are often related to muscle loss and severe liver disease. Patients experiencing significant muscle mass loss in the first week of ICU admission are at higher risk of extended stays [75]. This aligns with our findings, where intervening to set low creatinine

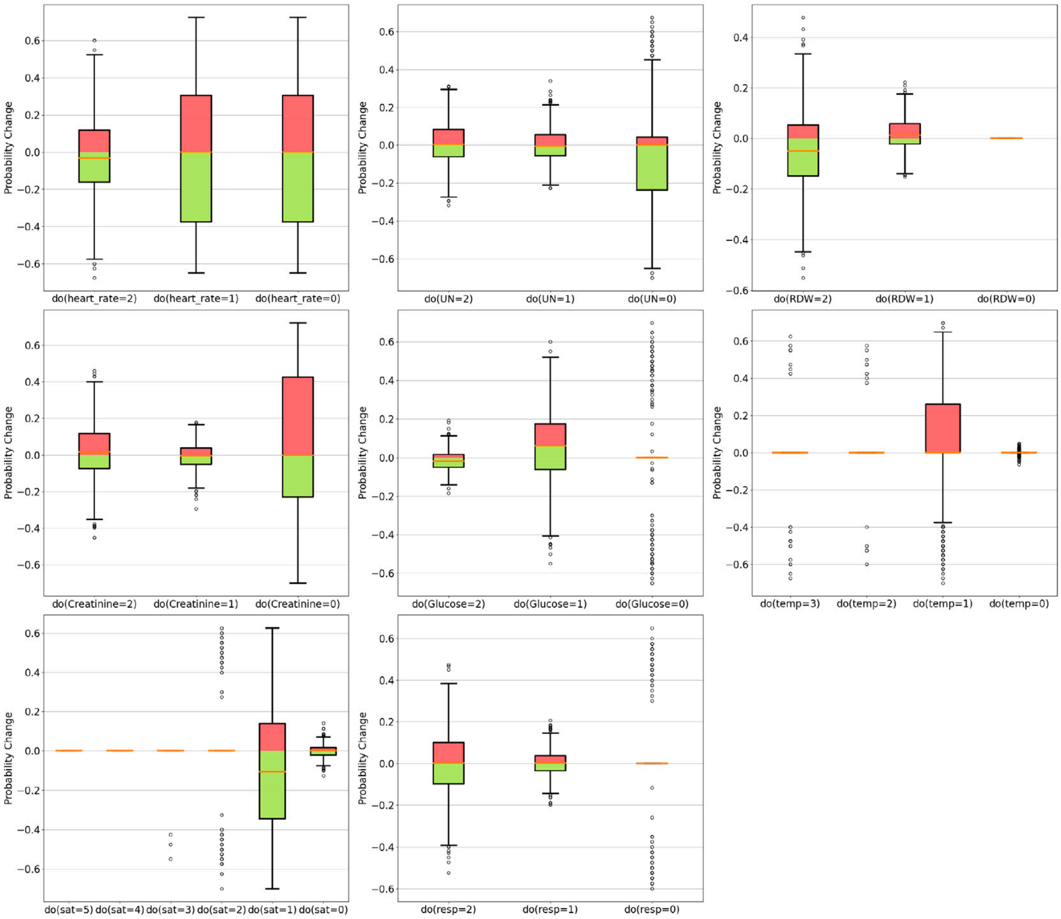

Fig. 4. Probability change of los given interventions on heart_rate, Urea_Nitrogen (UN), RDW, Creatinine, Glucose, temperature (temp), saturation (sat), and respiration rate (resp).

levels ($do(Creatinine = 0)$) notably increases the likelihood of extended ICU stays ($los = 1$).

High respiration rate, indicative of tachypnea in adults, is characterised by a respiratory rate exceeding 20 breaths per minute and often requires further assessment, leading to prolonged hospital stays [76]. In our model, intervening to elevate the respiration rate ($do(respiration = 2)$) resulted in an increased probability of extended ICU stays ($los = 1$).

Fever is a common issue in ICU patients and often necessitates diagnostic tests and procedures, which significantly prolongs the stay [77]. Consistent with this, our model shows that intervening to indicate mild fever ($do(temperature = 1)$) also leads to an elevated probability of extended ICU stays ($los = 1$).

Inadequate oxygen saturation ($do(saturation = 1)$) has a varied effect on $los = 1$, while other saturation levels show no discernible impact. Manipulating glucose levels reveals an inverse response, suggesting these variables indirectly influence LOS through intermediary factors rather than exerting a direct effect.

Framingham heart data:. This section explores the impact of various health factors on CHD risk through intervention analysis, aiming to assess the PCF's ability to replicate established causal relationships between these parameters and CHD. As illustrated in 5, our findings underscore significant alterations in $P(TenYearCHD = 1)$ across different interventions, shedding light on the intricate interplay between these health factors and CHD risk.

Existing literature highlights both systolic and diastolic hypertension as independent risk factors for adverse cardiovascular events [78]. Our analysis corroborates this, demonstrating that interventions on systolic blood pressure, such as $do(sysBP = 3)$, result in an increased probability of $CHD = 1$. Similarly, interventions on diastolic blood pressure ($do(diaBP = 1)$, $do(diaBP = 2)$ and $do(diaBP = 3)$) also heighten the likelihood of $CHD = 1$, reinforcing the significant impact of blood pressure levels on cardiovascular health.

Additionally, glucose metabolism plays a critical role in cardiovascular health, as deviations from normal glucose levels can lead to adverse outcomes [79]. Our model confirms this relationship, showing that interventions altering glucose levels, such as $do(glucose = 0)$ and $do(glucose = 2)$, significantly increase the probability of $CHD = 1$. These findings underscore the critical role that glucose regulation plays in cardiovascular risk management.

Raised total cholesterol levels are well-documented as a significant risk factor for coronary heart disease (CHD) [80]. In our model, interventions on total cholesterol ($do(totChol = 1)$ i.e levels ≥ 200) were found to increase the probability of $CHD = 1$, reinforcing the established link between elevated cholesterol and heightened CHD risk.

Literature highlights that asymptomatic bradycardia may influence heart disease risk due to underlying autonomic or cardiovascular issues [81]. Our intervention analysis, which simulates bradycardia through interventions on heart rate ($do(heartRate = 0)$), reveals a marked increase in the probability of $CHD = 1$.

Smoking has been highlighted as a leading risk factor for heart disease [82]. Our model's interventions demonstrate that smoking 6–10 cigarettes per day ($do(cigsPerDay = 2)$) and more than 11 cigarettes per day ($do(cigsPerDay = 3)$) significantly increase the probability of $CHD = 1$. These findings underscore the substantial impact of smoking on coronary heart disease risk.

Research indicates that higher education levels can lead to substantial health benefits [83]. Our model corroborates these findings,

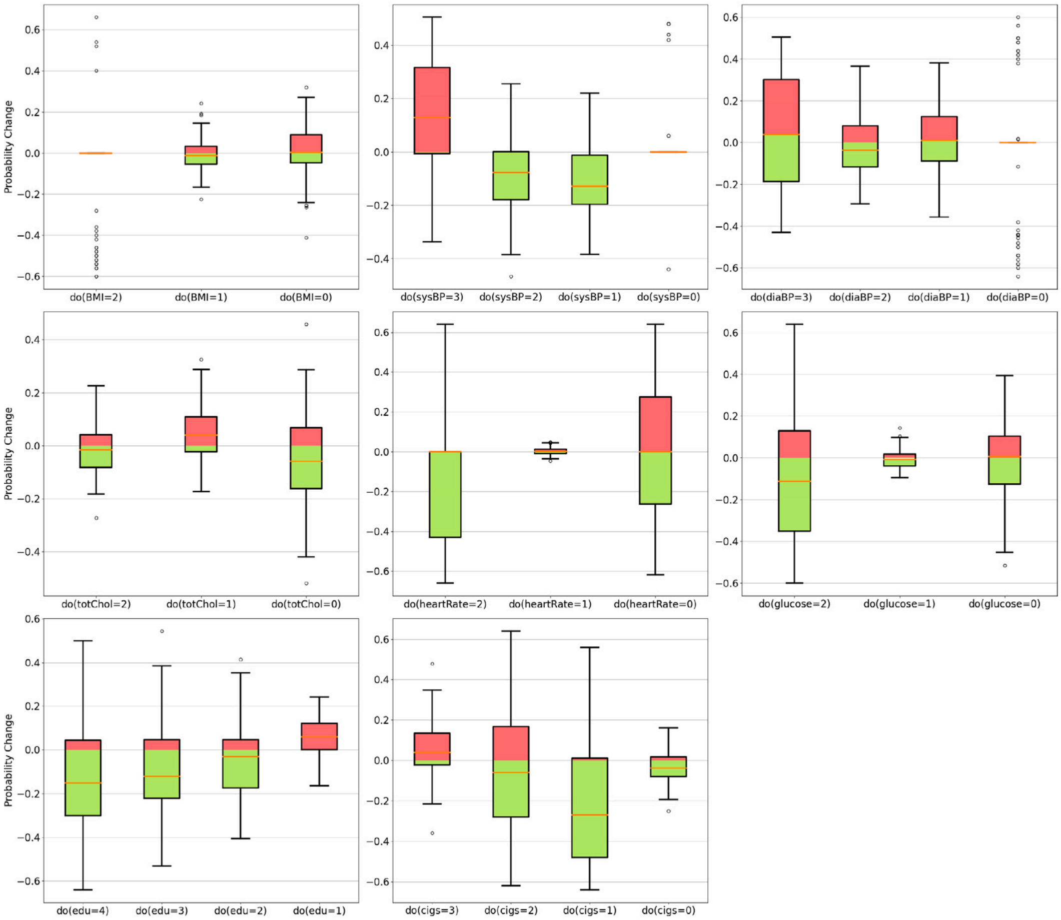

Fig. 5. Probability change of TenYearCHD given interventions on sysBP, diaBP, totChol, BMI, education, glucose, heartRate and cigsPerDay.

demonstrating that higher levels of education significantly decrease the probability of $CHD = 1$.

The relationship between BMI and CHD is often characterised as inconsistent and complex [84]. Our findings support this observation, as BMI interventions did not produce interpretable results. This ambiguity may be attributed to the intricate interplay of metabolic factors that extend beyond body mass alone, suggesting that BMI might not be a straightforward predictor of CHD risk. The lack of a clear relationship underscores the need for a more nuanced understanding of how metabolic factors contribute to CHD.

Diabetes:. This section explores the influence of various health factors on the likelihood of developing diabetes through intervention analysis. As depicted in Fig. 6, illustrates the significant changes in diabetes risk ($P(Diabetes = 1)$) following interventions on the selected variables.

Hypertension and hyperlipidemia are well-established predictors of diabetes risk, as highlighted in existing literature [85,86]. Our analysis confirms this relationship, showing that high blood pressure ($do(HighBP = 1)$) increases diabetes probability, while its absence ($do(HighBP = 0)$) decreases the risk. Similarly, elevated cholesterol levels ($do(HighChol = 1)$) are linked to a higher likelihood of diabetes, whereas normal cholesterol levels ($do(HighChol = 0)$) reduce the risk.

Body Mass Index (BMI) is another significant risk factor for diabetes [87]. Our findings indicate that maintaining a normal BMI ($do(BMI = 0)$) lowers the probability of diabetes, while a BMI of 40 or more ($do(BMI = 2)$) substantially raises this probability. This suggests that keeping a BMI between 0–24 mitigates diabetes risk, whereas higher BMI levels considerably elevate it.

Maintaining a healthy lifestyle is crucial for diabetes prevention [88]. Our analysis demonstrates that individuals with excellent general health ($do(GenHlth = 1)$) have a lower risk of developing diabetes compared to those in good ($do(GenHlth = 2)$) or poor health ($do(GenHlth = 3)$). Additionally, regular physical activity ($do(PhysActivity = 1)$) significantly reduces diabetes risk compared to a sedentary lifestyle ($do(PhysActivity = 0)$).

Education also plays a positive role in diabetes management and complication prevention [89]. Our results indicate that individuals with limited education ($do(education = 1)$) face a higher diabetes risk, while those with some secondary education ($do(education = 2)$) show mixed outcomes. Notably, higher education levels ($do(education = 3)$) are significantly associated with reduced diabetes risk.

The link between diabetes and heart disease is well-documented [90]. Our analysis supports this connection, as the absence of heart disease ($do(HeartDisease = 0)$) decreases diabetes risk, whereas its presence ($do(HeartDisease = 1)$) significantly increases it. This finding underscores the interconnected nature of these conditions, highlighting the need for integrated healthcare strategies.

To establish the empirical relevance of the selected intervention variables, we performed Chi-Square tests of independence with Benjamini–Hochberg false discovery rate (FDR) correction across all three datasets. All variable–outcome pairs exhibited statistically significant associations after multiple comparison adjustment ($q < 0.001$).

In the Framingham dataset, traditional cardiovascular risk factors, including blood pressure, cholesterol, glucose, smoking, and body mass index (BMI), were strongly associated with CHD risk ($\chi^2 = 16.9$–149.1). In the MIMIC-IV critical care dataset, physiological markers such as vital signs and laboratory values showed robust associations with ICU length of stay ($\chi^2 = 62.1$–360.3). The Diabetes dataset similarly confirmed well-established relationships between metabolic factors, health behaviours, and diabetes status ($\chi^2 = 55.5$–522.6).

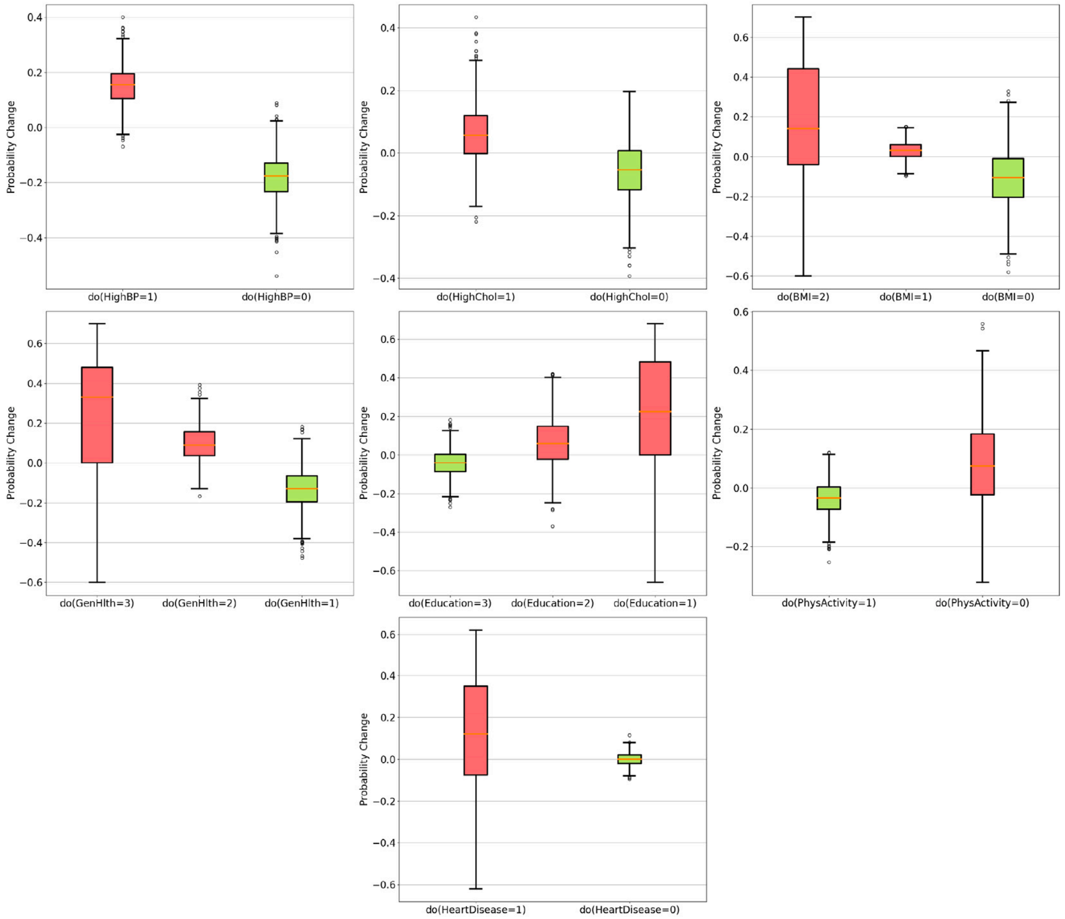

Fig. 6. Probability change of Diabetes given interventions on sysBP, diaBP, totChol, BMI, education, glucose, heartRate and cigsPerDay.

These results validate the empirical grounding of our intervention variables across diverse clinical domains, reinforcing the theoretical and causal relevance of the PCF framework.

5.4. *Counterfactuals*

Counterfactuals are alternative scenarios or outcomes that could have occurred if certain events or variables had been different. In healthcare, they are used to ask questions such as what would have happened if a different course of action had been taken or if specific variables had changed. Counterfactual reasoning enables clinicians to explore the probabilities associated with an "alternate reality," distinguishing between the indicative (events that actually occurred) and the subjunctive (events that could have occurred under different circumstances).

In this study, we investigate two types of counterfactual scenarios to assess the impact of clinician insights on our model, as described in Sections 5.4.1 and 5.4.2.

5.4.1. *Reordering variables based on domain knowledge*

We investigate the impact of modifying variable order within the PCF framework to illustrate its potential benefits as a proof of concept, rather than an implementation of clinician-directed decisions. While the CBN provides a robust foundational structure, our intent is to demonstrate how the model could be enhanced by integrating clinician-informed causal relationships. The core principle of our model is adaptability; it combines empirical foundations provided by the CBN with potential clinical insights. While the CBN's causal structure establishes the initial framework, we explore how clinician adjustments to the variable order might impact predictive precision. These adjustments are permitted within this partial ordering, that is, reordering is allowed only among variables not causally linked in the DAG. This maintains the integrity of the underlying causal assumptions while enabling clinically meaningful refinements.

By granting clinicians the ability to adjust variable sequences, we create an inclusive environment where domain-specific knowledge can shape and refine the model's architecture. This approach aims to balance the structured scaffolding provided by the CBN with the nuanced insights derived from clinical acumen, ultimately enriching the model's interpretability and operational efficacy in real-world healthcare settings. This fosters the exploration of counterfactual scenarios, assessing the impact of incorporating clinician insights on model performance and decision-making outcomes.

Soliciting specific feedback on variables, assumptions, and potential causal relationships enhances the interpretability, relevance, and trustworthiness of our model, ensuring alignment with clinical expertise and practice. Through this refinement, we navigate diverse scenarios or "what-if" queries related to the model's operation under distinct conditions, including modifications of causal trajectories informed by clinical expertise.

Fig. 7 illustrates how large hospitals, equipped with extensive datasets, can develop CBNs to represent causal relationships and generate pre-trained PCF models. These models could then be shared with smaller hospitals to support knowledge transfer and collaborative decision-making. Realising this vision of a centralised causal knowledge repository, however, requires addressing several practical challenges. Data sharing must comply with strict privacy regulations (e.g., HIPAA, GDPR), effective exchange depends on standardisation of clinical data across institutions, and strong governance is needed to ensure that shared models remain transparent, validated, and regularly updated.

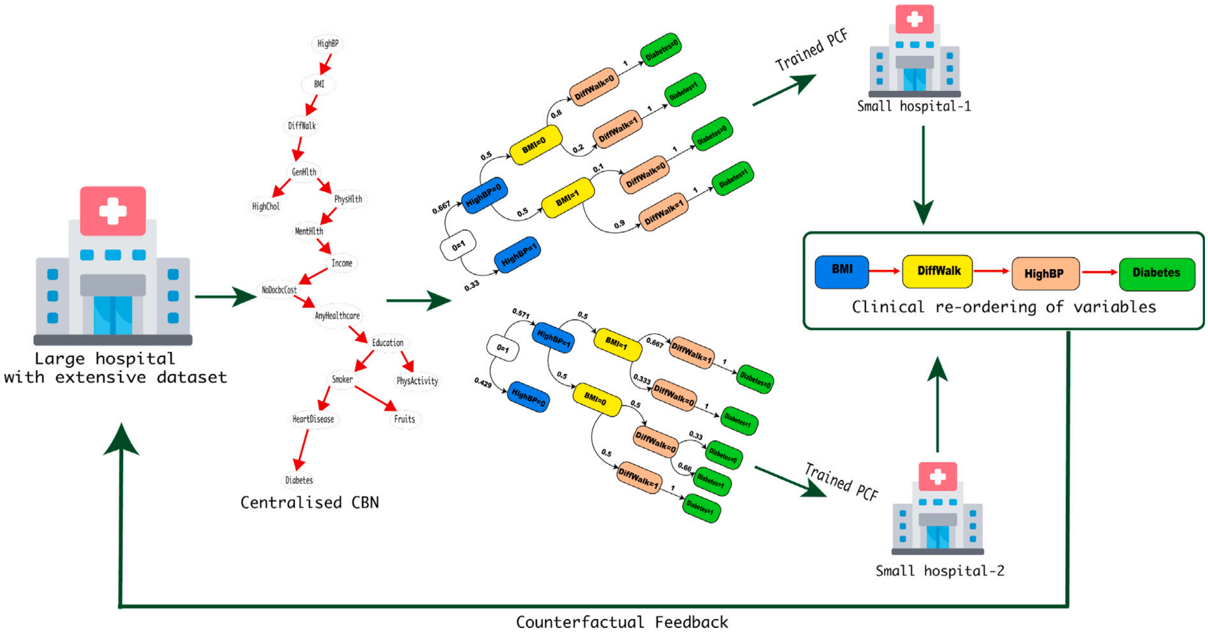

Fig. 7. The process of developing and sharing pre-trained PCFs by large hospitals with extensive datasets. The feedback loop illustrates how re-ordering variables integrates clinical insights into the Centralised CBN, enhancing collaborative decision-making in healthcare.

Table 4
Results using SMOTE for MIMIC-IV after changing the order.

Algorithm	Accuracy	Specificity	Sensitivity	AUC-ROC
PTree	80.29	91.11	40.06	65.59
PCF	72.43	78.33	50.50	64.41

Table 5
Results using SMOTE for framingham data after changing the order.

Algorithm	Accuracy	Specificity	Sensitivity	AUC-ROC
PTree	64.98	69.40	40.31	54.85
PCF	66.39	69.12	51.16	60.14

Table 6
Results using ADASYN for diabetes data after changing the order.

Algorithm	Accuracy	Specificity	Sensitivity	AUC-ROC
PTree	73.29	78.93	38.14	58.54
PCF	73.71	75.45	62.88	69.17

While these barriers are substantial, advances in federated learning and privacy-preserving computation provide potential pathways towards implementation.

***MIMIC-IV:*.** The initial variable order provided by the CBN positioned “Diagnosis_2” earlier in the sequence, suggesting its early influence on the outcome variable. However, the counterfactual adjustment hypothesised that strategically reordering the variables might enhance the model’s capacity to learn causal relationships and predict los more accurately. This adjustment reflected the potential causal flow where “Diagnosis_2” is informed by preceding laboratory tests. Therefore, we reordered the variables to place “Diagnosis_2” later in the sequence, just before the outcome variable (los).

Table 4 summarises the performance of the PCF and PTree methods following the variable order change. The table reveals that while the rearrangement did not alter the accuracy of the PTree model, it resulted in a slight decrease in PCF performance. This outcome might be attributed to the sensitivity of the PCF model to variable order, as it relies on an intricate interplay of causality among variables. The reordering may have disrupted previously established dependencies, highlighting the complex interactions inherent in the data. Such changes underline the importance of carefully considering the sequence of variables in models sensitive to causal relationships.

***Framingham heart data:*.** The original variable order provided by the CBN included ‘BPMeds’, ‘prevalentHyp’, ‘heartRate’, ‘prevalentStroke’, ‘diabetes’, ‘sysBP’, ‘totChol’, ‘glucose’, ‘diaBP’, ‘BMI’, ‘education’, ‘currentSmoker’, ‘cigsPerDay’, and ‘TenYearCHD’. Subsequently, a counterfactual adjustment was made, rearranging certain variables to create a revised order: ‘BPMeds’, ‘prevalentHyp’, ‘diabetes’, ‘glucose’, ‘heartRate’, ‘sysBP’, ‘diaBP’, ‘BMI’, ‘education’, ‘totChol’, ‘prevalentStroke’, ‘currentSmoker’, ‘cigsPerDay’, and ‘TenYearCHD’.

Table 5 presents the performance evaluation results for the PCF and PTree methods following the variable order modification. The analysis indicates almost similar performance for both the PCF and PTree method (slightly lower than original). This slight decrease can be attributed to the fact that the original order might have implicitly captured relevant relationships for CHD prediction better than the counterfactual order.

***Diabetes:*.** The original order of variables provided by the CBN was as follows: ‘HighBP’, ‘BMI’, ‘DiffWalk’, ‘GenHlth’, ‘PhysHlth’, ‘HighChol’, ‘MentHlth’, ‘Income’, ‘NoDocbcCost’, ‘AnyHealthcare’, ‘Education’, ‘Smoker’, ‘PhysActivity’, ‘HeartDiseaseorAttack’, ‘Fruits’, ’Diabetes_binary’. The counterfactual order maintained the overall structure but changed the position of a few variables and the new order became: ‘GenHlth’,‘BMI’,‘DiffWalk’,‘PhysHlth’, ‘PhysActivity’, ‘HighBP’, ‘HighChol’, ‘MentHlth’, ‘Education’, ‘Income’, ‘NoDocbcCost’, ‘AnyHealthcare’, ‘Smoker’, ‘HeartDiseaseorAttack’, ‘Fruits’, ’Diabetes_binary’.

Table 6 summarises the accuracy of PCF and PTree methods, after the change in variable order. As the table shows, reordering the variables led to an increase in the model accuracy for PCF as well as for PTree. However, AUC-ROC seems to have decreased in both. The results of these reordering experiments highlight the complex interplay between data-driven structure learning and expert-informed adjustments. In two of the three datasets, the original variable order derived from the CBN produced slightly better performance, suggesting that the data-driven approach may capture subtle dependencies not easily articulated through domain expertise alone. However, in the third dataset, the clinician-informed reordering led to performance gains, illustrating the potential value of expert insight when aligned with the data. These findings underscore the flexibility of the PCF framework while also pointing to the need for careful evaluation when integrating domain knowledge, especially in models sensitive to causal ordering.

5.4.2. *Counterfactual statements for specific datasets*

In addition to reordering variables, PCF can generate counterfactual statements by altering the values of specific variables within a patient's observed history. This addresses retrospective "what if" questions, for example, how ICU length of stay (LOS) might have changed had a fever not occurred, while keeping the rest of the clinical context fixed. Unlike forward-looking interventions, which model hypothetical changes without reference to prior states, counterfactuals are anchored in the factual record. This allows PCF to provide patient-level insights that closely mirror real clinical scenarios.

Following Pearl's framework, PCF implements counterfactual reasoning through a "twin network" representation: one copy of the model encodes the factual scenario, while a parallel copy encodes the counterfactual. This construction enables direct comparison between observed outcomes and their hypothetical alternatives, isolating the marginal contribution of the variable change under consideration.

Formally, PCF computes counterfactual probabilities of the form

$$P(Y_C \mid X = x),$$

where Y_C denotes the outcome under the counterfactual assumption that variable C takes a different value, and $X = x$ represents the factual context. This notation follows Genewein et al. [6], where Y_C represents the subjunctive outcome under assumption C. It is equivalent to the more widely adopted notation $Y_{C=c}$ used in Pearl's framework. Simulations are restricted to clinically plausible changes, such as increasing heart rate from bradycardic to normal or restoring normothermia in febrile patients.

Building on this formalisation, PCF generates counterfactual probability trees by (1) conditioning on the observed evidence and (2) applying a structural intervention using the do() operator, which severs upstream dependencies and reinitialises downstream variables. Implementation follows the procedure of Genewein et al. [6] via the function $\mathrm{CF}(n, m, \delta)$, where n and m are the roots of the reference and factual trees, respectively, and δ is the min-cut set for the intervention. Recursive alignment with

$$\mathrm{ZIP}(A, B) = \{(a_n, b_n)\}_{n=1}^{N}$$

re-evaluates subpaths, allowing PCF to isolate the marginal effect of a single variable change.

To ensure tractability and interpretability, PCF employs a principle of feature sparsity. Counterfactuals are limited to a small set of clinically relevant variables, avoiding wholesale alterations across the dataset. Sparsity therefore plays a dual role, it reduces computational complexity and constrains the analysis to scenarios that are realistic and actionable. This perspective is consistent with the emerging field of counterfactual explainability [91], which emphasises the importance of limiting interventions to those that are both theoretically sound and practically implementable.

Guided by this principle, we selected features for each dataset based on their established presence in the literature and their recognised clinical relevance. Incorporating such well-supported variables enhances the robustness and interpretability of our analyses, while ensuring that the resulting counterfactuals are meaningful in real-world decision-making contexts. Fig. 8 provides a visual illustration of this process.

As an illustration, consider a patient with hypotension and fever, predicted to have a 72% probability of prolonged ICU stay. A counterfactual adjustment, setting temperature to a normal range do(temperature = 0), reduced this estimate to 51%, suggesting that earlier fever control might have shortened the patient's stay.

MIMIC-IV:. This section explores the factors influencing the LOS in the ICU using counterfactual analysis. Specifically, we examine the conditional probability of a patient requiring an extended ICU stay ($los = 1$, i.e., more than 4 days). By employing counterfactual explanations, we investigate hypothetical scenarios where certain vital signs or laboratory values are altered. This allows us to assess the impact of these changes on the probability of an extended ICU stay. Features were chosen based on their potential to yield valuable insights into the determinants of prolonged ICU stays.

(a) **Heart Rate:** We analysed the effect of heart rate by comparing the baseline probability $P(los = 1 \mid heart\_rate = 0)$ for patients with low heart rate (0) to the counterfactual scenario where their heart rate is normal (1). The counterfactual probability $P(los^* = 1 \mid heart\_rate = 0), \quad los^* = los_{heart\_rate \leftarrow 1}$[1] suggests a decrease in the likelihood of extended ICU stay when the heart rate is normal.

(b) **Saturation:** Similarly, we examined the effect of oxygen saturation by comparing the baseline probability $P(los = 1 \mid saturation = 3)$ for patients with very low oxygen saturation (3) to the counterfactual scenario with normal oxygen saturation (0). The counterfactual probability $P(los^* = 1 \mid saturation = 3), \quad los^* = los_{saturation \leftarrow 0}$ indicates that there is only a marginal difference in the probability of an extended ICU stay when the patient's saturation level is normal.

(c) **Glucose and Urea Nitrogen:** We further analysed the role of glucose and Urea Nitrogen levels. Interestingly, for both high glucose $P(los = 1 \mid Glucose = 2)$ and high Urea Nitrogen $P(los = 1 \mid UreaNitrogen = 2)$, the counterfactual scenarios with normal levels (glucose = 1 and Urea Nitrogen = 1, respectively) showed slightly increased probabilities of extended ICU stay $P(los^* = 1 \mid Glucose = 2), \quad los^* = los_{Glucose \leftarrow 1}$ and $P(los^* = 1 \mid UreaNitrogen = 2), \quad los^* = los_{UreaNitrogen \leftarrow 1}$. Thus, the probability of extended ICU stay may be increased even if the patient had normal glucose and urea nitrogen levels.

(d) **Temperature:** Finally, we explored the influence of body temperature. The baseline probability for extended ICU stay with high fever $P(los = 1 \mid temperature = 2)$ was compared to the counterfactual scenario with normal temperature (0). This resulted in a minor decrease in the probability $P(los^* = 1 \mid temperature = 2), \quad los^* = los_{temperature \leftarrow 0}$, which suggests a slight decrease in the likelihood of extended ICU stay, if the patient had normal body temperature.

Fig. 9 illustrates the probability of a patient remaining in the ICU for more than four days ($los = 1$) under various counterfactual scenarios involving five different variables: Heart Rate, Saturation, Glucose, Urea, and Temperature. Each line in the plot represents one of these variables, with the x-axis displaying the factual and counterfactual scenarios and the y-axis showing the probability values. This figure provides insights into how alterations in these variables influence the likelihood of an extended ICU stay.

Framingham data:. This data offers a wealth of information on cardiovascular risk factors. To leverage counterfactual explanations effectively, we strategically select the features with high explanatory potential for predicting CHD.

(a) **BMI:** We began our analysis by examining the impact of Body Mass Index (BMI) on the likelihood of developing CHD. Starting with a BMI of 2 (High), indicative of CHD, we delved into

[1] Here, los^* denotes a copy of the outcome variable los in the counterfactual world created by the intervention. The expression $los^* = los_{heart\_rate \leftarrow 1}$ follows the probability-tree convention of Genewein et al. [6], and corresponds to Pearl's standard notation $Y_{X=1}$, where Y is the outcome (los) and X is the intervened variable ($heart\_rate$).

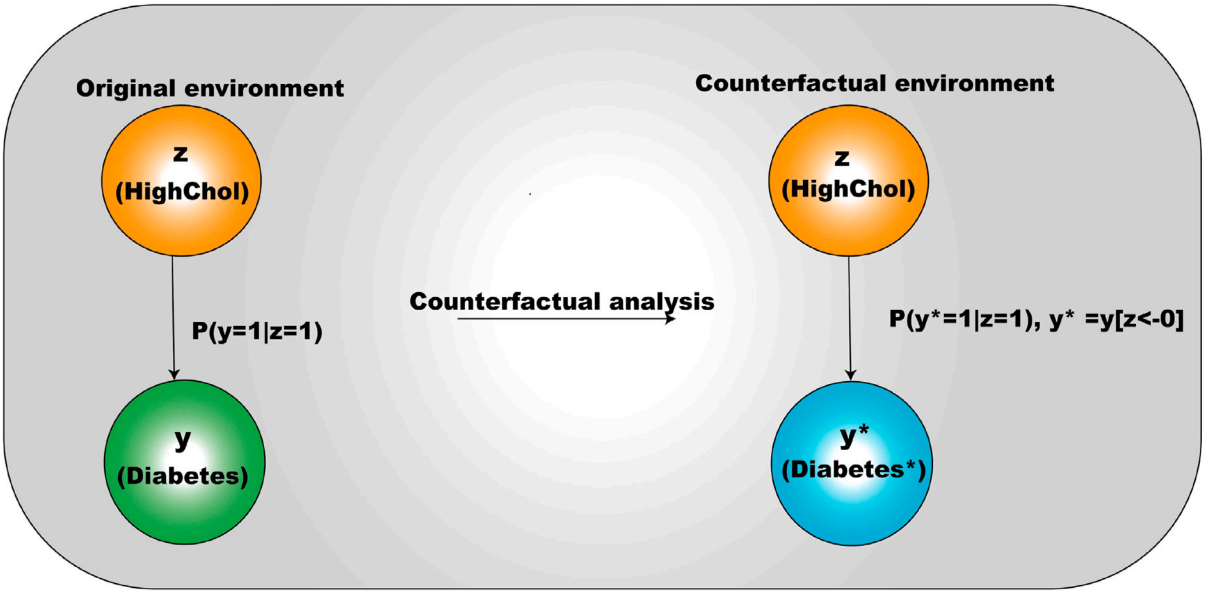

Fig. 8. Counterfactual analysis using PCF. The left panel represents the observed environment, where the probability of diabetes is conditioned on the presence of high cholesterol: $P(y = 1 \mid z = 1)$. The right panel depicts a counterfactual world in which we estimate the outcome if the high cholesterol condition were absent, represented as $P(y^* = 1 \mid z = 1)$, where $y^* = y_{z\leftarrow 0}$. Here, z denotes the causal variable HighChol, y is the observed outcome Diabetes, and y^* is the counterfactual outcome under the intervention.

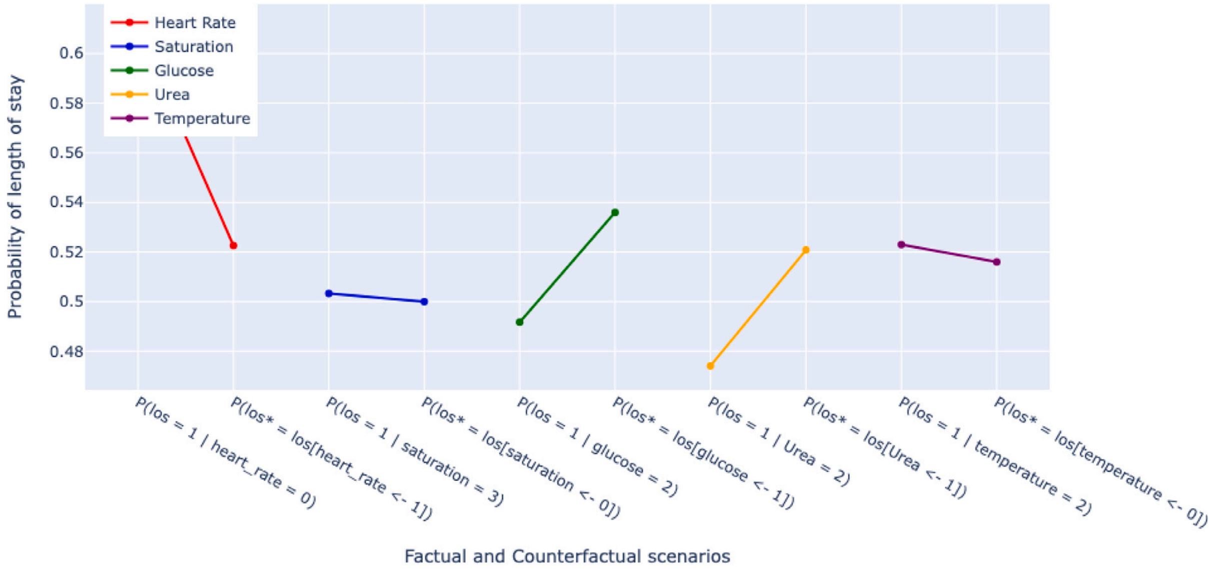

Fig. 9. Line plots showing the probability of ICU stay exceeding 4 days under factual and counterfactual scenarios for various health variables.

counterfactual scenarios to explore the probability of CHD had the patient possessed a normal BMI (0). The baseline probability $P(CHD = 1 \mid BMI = 2)$, represents the likelihood of CHD under the existing BMI condition. Interestingly, the counterfactual probability $P(CHD^* = 1 \mid BMI = 2)$, $CHD^* = CHD_{BMI\leftarrow 0}$, reveals that even with a normal BMI, the risk of CHD might still remain relatively high.

(b) **Systolic Blood Pressure (sysBP) and Diastolic Blood Pressure (diaBP):** We investigated the influence of blood pressure measurements (systolic pressure, or sysBP, and diastolic pressure, or diaBP) of patients on CHD prevalence. Individuals with high blood pressure (represented by a score of 3 for both sysBP and diaBP) were found to have a higher chance of having CHD $P(CHD = 1 \mid sysBP = 3)$, $P(CHD = 1 \mid diaBP = 3)$. We then considered a hypothetical scenario: what if these patients with high blood pressure had normal values instead (sysBP = 1 and diaBP = 1)? The corresponding probabilities, $P(CHD^* = 1 \mid sysBP = 3)$, $CHD^* = CHD_{sysBP\leftarrow 1}$ and $P(CHD^* = 1 \mid diaBP = 3)$, $CHD^* = CHD_{diaBP\leftarrow 1}$, show how lowering blood pressure could potentially decrease the risk of CHD.

(c) **Total Cholesterol (totChol):** We explored the potential influence of total cholesterol (totChol) on the development of CHD using counterfactual analysis. In the factual scenario, we assessed patients based on their actual totChol (totChol = 3). The baseline probability was determined as $P(CHD = 1 \mid totChol = 3)$. In the counterfactual scenario, we posed the question: what if these

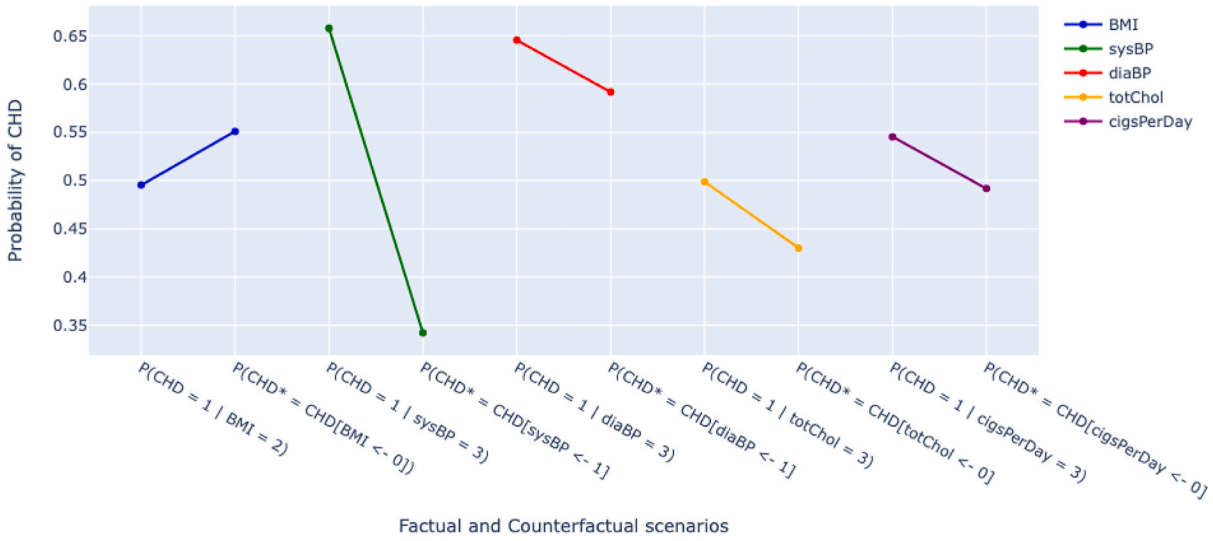

Fig. 10. Probability distribution of TenYearCHD for factual and counterfactual scenarios for various variables.

patients with high totChol had normal values (totChol = 0)? The resulting probability $P(CHD^* = 1 \mid totChol = 3)$, $CHD^* = CHD_{totChol \leftarrow 0}$ suggests that maintaining normal total cholesterol levels might be beneficial for reducing CHD risk.

(d) **Cigarettes per day (cigsPerDay):** Finally, we explored the influence of smoking. In the factual scenario, we examined patients based on their actual cigarette consumption (cigsPerDay = 3). The baseline probability was determined as $P(CHD = 1 \mid cigsPerDay = 3)$. The counterfactual scenario asks, "what if" these patients who smoke heavily cigsPerDay = 3(≥ 11 cigarettes/day) had never smoked (cigsPerDay = 0)? The resulting probability $P(CHD^* = 1 \mid cigsPerDay = 3)$, $CHD^* = CHD_{cigsPerDay \leftarrow 0}$ suggests that quitting smoking could be beneficial for reducing CHD risk.

Fig. 10 illustrates the line plots generated to explore the distribution of risk factors for CHD using counterfactual analysis. It can be observed that the counterfactual scenarios pertaining to sysBP, diaBP, totChol and cigsPerDay potentially change the probability of $CHD = 1$.

Diabetes data:. This section explores the application of counterfactual explanations within the diabetes dataset. By strategically selecting features, we focus on identifying modifiable risk factors.

(a) **HighBP:** We examined the relationship between HighBP and the prevalence of diabetes. In the actual scenario, patients were evaluated based on their recorded blood pressure levels (HighBP = 1). The baseline probability was calculated as $P(Diabetes\_binary = 1 \mid HighBP = 1)$. However, in the hypothetical scenario, we posed the question: what if these patients with high blood pressure had normal blood pressure (HighBP = 0)? The resulting probability $P(Diabetes\_binary^* = 1 \mid HighBP = 1)$, $Diabetes\_binary^* = Diabetes\_binary_{HighBP \leftarrow 0}$ suggests a potential benefit of maintaining normal blood pressure to reduce the risk of diabetes. This is in accordance to the literature which states that Blood pressure control is just as important as glycemic control [92].

(b) **HighChol:** We examined the influence of high cholesterol (HighChol) on the prevalence of Diabetes. In the factual scenario, patients were evaluated based on their actual cholesterol measurements (HighChol = 1). The baseline probability was computed as $P(Diabetes\_binary = 1 \mid HighChol = 1)$. However, in the counterfactual scenario, we considered: what if these patients with high cholesterol had normal levels (HighChol = 0)? The resulting probability $P(Diabetes\_binary^* = 1 \mid HighChol = 1)$, $Diabetes\_binary^* = Diabetes\_binary_{HighChol \leftarrow 0}$ highlights the potential advantage of normalising cholesterol levels in mitigating the risk of diabetes.

(c) **BMI:** We explored the impact of BMI on diabetes prevalence using counterfactual analysis. In the factual scenario, we assessed patients based on their actual BMI (BMI = 2). The baseline probability was $P(Diabetes\_binary = 1 \mid BMI = 2)$ The counterfactual scenario investigated the hypothetical scenario where patients with high BMI (BMI = 2) had a normal BMI (BMI = 0). We aimed to determine the impact of this hypothetical change on diabetes risk. The resulting probability $P(Diabetes\_binary^* = 1 \mid BMI = 2)$, $Diabetes\_binary^* = Diabetes\_binary_{BMI \leftarrow 0}$ suggests a potential benefit of maintaining a healthy weight (represented by normal BMI) in reducing diabetes risk.

(d) **GenHealth:** This study investigated the link between a patient's overall health (GenHealth) and their risk of developing diabetes using counterfactual analysis. The indicative premise is that the patient has poor health (GenHealth =3), and the subjunctive (counterfactual) premise is if the patient has excellent health (GenHealth =1) in an alternate reality. The baseline probability was determined as $P(Diabetes\_binary = 1 \mid GenHealth = 3)$ The probability resulting from the subjunctive premise, $P(Diabetes\_binary^* = 1 \mid GenHealth = 3)$, $Diabetes_binary^* = Diabetes_binary_{GenHealth \leftarrow 1}$illustrates the potential benefit of maintaining the overall well-being so as to reduce the risk of diabetes.

Fig. 11 illustrates the change in $P(Diabetes = 1)$ as a result of the counterfactual statements described.

Standard predictive models typically highlight associations between risk factors and outcomes, but they do not offer insight into how modifying those factors would alter individual patient trajectories. In contrast, PCF enables counterfactual analysis that estimates the expected change in outcome probability under specific hypothetical interventions. For instance, in the Framingham dataset, adjusting systolic blood pressure from high to normal reduces the estimated probability of CHD, while in the MIMIC-IV cohort, normalising heart rate results in a lower predicted probability of extended ICU stay. These estimates move beyond associative insights by quantifying potential benefits of intervention at the patient level. Such information is particularly valuable in clinical settings where decisions must balance risk, feasibility, and expected benefit, offering a pragmatic framework to assess how targeted changes in modifiable variables might influence outcomes.

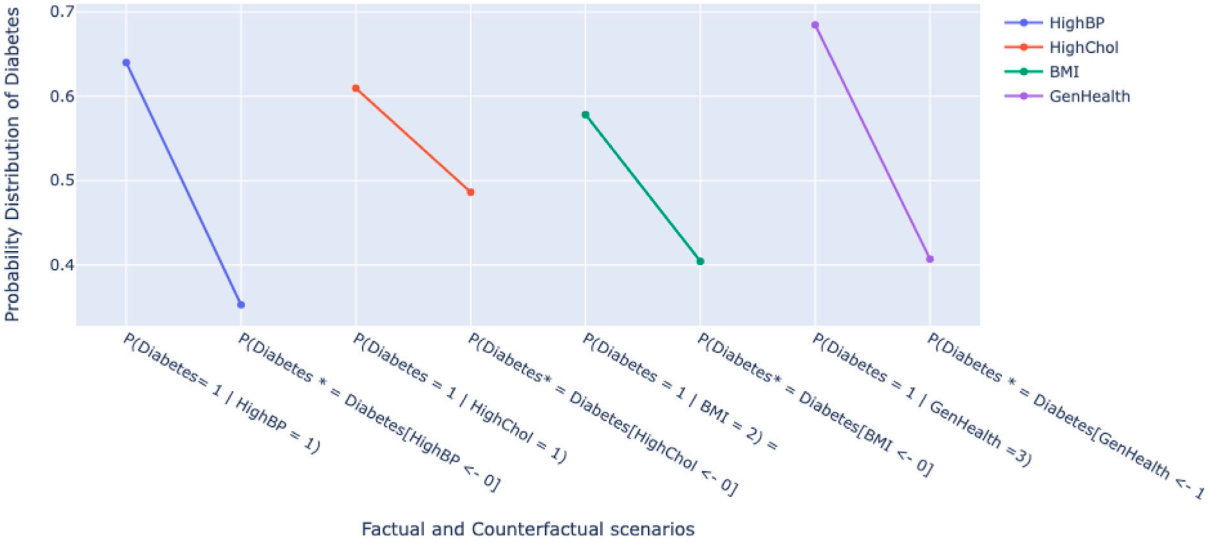

Fig. 11. Probability Distribution of Diabetes for factual and counterfactual scenarios for various variables.

6. Conclusion

This study introduces the Probabilistic Causal Fusion (PCF) framework, which combines Causal Bayesian Networks (CBNs) and Probability Trees (PTrees) to enhance healthcare decision-making. This innovative approach harnesses the causal structure learned by the CBN to establish the foundational framework of the PTree. This synergy yields a three-fold benefit: (1) it captures the inherent causal relationships within the data, leading to a more robust understanding of the factors influencing outcomes, (2) it facilitates the incorporation of domain knowledge through counterfactual analysis, empowering clinicians to integrate their expertise into the model, and (3) it facilitates the creation of a centralised repository of causal knowledge across institutions. This fosters collaboration, knowledge exchange, and continuous improvement in healthcare delivery.

Rigorous validation using three real-world medical datasets, MIMIC-IV, Framingham Heart Study, and BRFSS, demonstrates that the proposed methodology achieves prediction performance on par with established models. Importantly, these datasets span diverse clinical domains, critical care, cardiovascular health, and chronic disease, highlighting the framework's generalisability across settings. However, PCF's true strength lies in its ability to surpass mere prediction and empower clinicians. Unlike traditional machine learning methods, this framework facilitates the exploration of hypothetical interventions and counterfactual scenarios through counterfactual analysis. While PCF achieves predictive performance that is comparable to standard models such as decision trees and ensemble methods, it does not consistently exceed them in accuracy. This outcome highlights a trade-off inherent in the design of PCF: the framework emphasises causal interpretability and support for interventional analysis, which may come at the expense of slight reductions in predictive performance.

This enhanced functionality translates into a more comprehensive toolkit for healthcare professionals. Unlike conventional models that focus solely on prediction, PCF enables prediction, interventional reasoning, and counterfactual analysis within a single framework. This multi-faceted capability offers a holistic approach to clinical decision support, allowing clinicians to not only anticipate outcomes but also explore the potential effects of modifiable risk factors and hypothetical treatment strategies. This deeper understanding of variable interactions and intervention effects significantly improves clinical decision-making, ultimately leading to optimised patient care.

A key strength of this approach is its dual applicability at the individual and population levels. Clinicians can leverage this framework to gain insights into broader population trends while simultaneously exploring personalised treatment options for specific patients through counterfactual analysis. This versatility empowers healthcare professionals to tailor their decision-making to the unique circumstances of each patient while simultaneously informing clinical best practices for the entire population.

In addition to its predictive and causal capabilities, PCF supports multi-level interpretability through the integration of sensitivity analysis and SHAP. Sensitivity analysis provides macro-level insights by highlighting how changes in causal parameters affect outcomes across the CBN structure. In contrast, SHAP offers micro-level explanations by attributing individual predictions to specific input features. This dual approach enables clinicians to understand both the broader causal mechanisms at play and the factors driving patient-specific outcomes, offering a more complete rationale for clinical decision-making.In summary, the PCF framework offers a combination of (i) robust validation across diverse clinical datasets, (ii) integrated prediction, intervention, and counterfactual capabilities, and (iii) interpretability at both macro and micro levels. These features collectively distinguish PCF as a practical and transparent alternative to standard predictive models in clinical machine learning.

Our study suggests that this approach holds significant promise for evidence-based clinical decision-making. However, further exploration is needed to address several limitations:

1. Computational and scalability constraints: Optimising computational efficiency, especially for large datasets, is crucial for broader applicability. While PTrees offer clear interpretability, they are susceptible to the "curse of dimensionality," as the number of variables and branching paths increases. To address this, future work could explore the use of Chain Event Graphs (CEGs), which have been shown to represent context-specific and asymmetric relationships more compactly. Notably, CEGs have been applied to causal inference problems with success, offering a potential direction to enhance the scalability and expressiveness of the PCF framework [93]. The computational complexity of the PCF framework also presents a practical limitation. The ensemble-based structure learning procedures and SHAP value computations, particularly when using KernelExplainer, can be resource-intensive, especially when applied to high-dimensional datasets such as MIMIC-IV and BRFSS. These demands may limit the framework's scalability or its applicability in real-time clinical settings. Future research

may explore more computationally efficient alternatives or approximations to maintain interpretability while improving performance in large-scale environments.

2. Assumptions and causal identifiability: As with all causal inference methods based on observational data, the validity of PCF's causal claims depends on the assumption that all relevant confounders are observed and included in the dataset. This assumption underpins both the CBN structure and the resulting interventional and counterfactual analyses. While model averaging enhances robustness to algorithmic variability, it does not eliminate the risk of unmeasured confounding. Moreover, although ensemble-based structure learning improves the stability of the inferred causal graph, no algorithm can guarantee full recovery of the true data-generating process, especially in the presence of limited data or complex dependencies. This limitation highlights the value of integrating expert domain knowledge alongside data-driven discovery, particularly in clinical settings where the validity of causal insights is critical. PCF does not claim to surpass expert-informed approaches but is designed to flexibly integrate both data-driven discovery and domain expertise. Acknowledging these challenges, future extensions of the framework could explore validation using experimental or interventional data to further reinforce the credibility of the inferred causal relationships.

3. Data-sharing and governance considerations: The proposed centralised causal knowledge repository depends on the ability to share pre-trained PCF models and causal structures across institutions. In practice, such sharing may be constrained by privacy regulations, interoperability issues, and institutional governance policies. Ensuring secure, standardised, and ethically governed mechanisms for model exchange will be essential for successful deployment.

4. Need for prospective and external validation: While this study focused on specific medical domains, future research should investigate the framework's generalisability to a wider range of healthcare settings. Incorporating clinical expertise in selecting variables for counterfactual and interventional analysis is essential. Clinicians' insights can refine the methodology and enhance its practical utility by ensuring the system uses the most relevant and useful data.

Addressing these limitations and broadening the scope of applications, including the potential use of genomic and multi-omics data, will further demonstrate the framework's potential to advance healthcare. Building on this approach can lead to the development of more effective and transparent tools that enhance patient care. Such tools have the potential to support evidence-based clinical decision-making and contribute to a more efficient and impactful healthcare system overall.

CRediT authorship contribution statement

Sheresh Zahoor: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Pietro Liò:** Writing – review & editing, Validation, Supervision, Resources, Investigation, Conceptualization. **Gaël Dias:** Writing – review & editing, Validation, Supervision, Conceptualization. **Mohammed Hasanuzzaman:** Writing – review & editing, Validation, Supervision, Conceptualization.

Code availability

All code used to implement the PCF framework, including data preprocessing, CBN construction, PTree ensemble generation, and evaluation pipelines, will be made publicly available in a GitHub repository upon publication to support reproducibility and further research.

Ethics statement

This study used only publicly available, de-identified datasets (MIMIC-IV, Framingham Heart Study, and BRFSS). All data were collected with appropriate ethical approvals and informed consent by the original data custodians. No new data collection or patient contact was involved. Therefore, no additional ethical approval was required for this study.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has been jointly supported by Research Ireland [Grant number 18/CRT/6222] and the European Union's Horizon 2020 research and innovation programme [Grant agreement number 101017385]. Additionally, for the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Appendix

See Table 7.

Table 7
Summary of hyperparameters and implementation choices for pcf and baseline models.

Component	Parameter	Value/Description
CBN Structure Learning		
	Algorithms used	Hill Climbing (HC), TABU, SaiyanH, MAHC, Greedy Equivalence Search (GES)
	Scoring function	BIC (Bayesys default); algorithm-specific where applicable
	Target-aware search	Enabled (outcome-directed search in Bayesys)
	Max parents per node	3
	Edge frequency threshold	Included if present in at least 2 of 5 algorithms (≥ one-third frequency, as per Bayesys default)
PTree Ensemble (PCF)		
	Number of trees	100 PTrees (bootstrapped mini-batches)
	Batch size	50 samples per tree
	Variable ordering	Derived from model-averaged CBN topological sort (partial ordering)
	Max depth	Not explicitly fixed; controlled via pruning
	Pruning threshold θ	Dataset-specific; selected via pruning-curve analysis, range explored $\theta \in [0.0, 0.2]$
Decision Threshold τ		
	Threshold selection	Clinically chosen trade-off between sensitivity and specificity ($\tau = 0.454$)
Baseline Models		
	Logistic Regression (LR)	Solver = `liblinear`; default L_2 regularisation
	Decision Tree (DT)	Max depth = 4
	Random Forest (RF)	100 trees; max depth = 15; max leaf nodes = 150; min samples split = 200; max features = `sqrt`; class_weight = `balanced`
	Gradient Boosting (GB)	Standard implementation; no additional hyperparameter tuning beyond defaults
	XGBoost (XGB)	Objective = `binary:logistic`; max depth = 10; learning rate = 1.0; $\alpha = 10$; 100 estimators
	AdaBoost	SAMME algorithm; 100 estimators
	SVM	LinearSVC with polynomial (degree = 2, interaction-only) and RBF random features ($\gamma = 0.01$); $C = 1$; with StandardScaler
	KNN	$k = 3$; Euclidean distance

References

[1] Pearl J. Causality: Models, reasoning and inference. Cambridge University Press; 2009, URL http://bayes.cs.ucla.edu/BOOK-2K/neuberg-review.pdf.

[2] Sanchez P, Voisey JP, Xia T, Watson HI, O'Neil AQ, Tsaftaris SA. Causal machine learning for healthcare and precision medicine. R Soc Open Sci 2022;9(8):220638. http://dx.doi.org/10.1098/rsos.220638, URL https://royalsocietypublishing.org/doi/10.1098/rsos.220638.

[3] Kopitar L, Kocbek P, Cilar L, Sheikh A, Stiglic G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Sci Rep 2020;10(1):11981.

[4] Feuerriegel S, Frauen D, Melnychuk V, Schweisthal J, Hess K, Curth A, Bauer S, Kilbertus N, Kohane IS, van der Schaar M. Causal machine learning for predicting treatment outcomes. Nature Med 2024;30(4):958–68.

[5] Prosperi M, Guo Y, Sperrin M, Koopman JS, Min JS, He X, Rich S, Wang M, Buchan IE, Bian J. Causal inference and counterfactual prediction in machine learning for actionable healthcare. Nat Mach Intell 2020;2(7):369–75. http://dx.doi.org/10.1038/s42256-020-0197-y, URL https://www.nature.com/articles/s42256-020-0197-y.

[6] Genewein T, McGrath T, Delétang G, Mikulik V, Martic M, Legg S, Ortega PA. Algorithms for causal reasoning in probability trees. 2020, http://dx.doi.org/10.48550/arXiv.2010.12237, arXiv preprint arXiv:2010.12237.

[7] Ambags EL, Capitoli G, Imperio VL, Provenzano M, Nobile MS, Liò P. Assisting clinical practice with fuzzy probabilistic decision trees. 2023, http://dx.doi.org/10.48550/arXiv.2304.07788, arXiv:2304.07788.

[8] Leonelli M, Varando G. Structural learning of simple staged trees. Data Min Knowl Discov 2024;38(3):1520–44.

[9] Kitson NK, Constantinou AC. Eliminating variable order instability in greedy score-based structure learning. In: International conference on probabilistic graphical models. 2024, p. 147–63.

[10] Chen L, Ji P, Ma Y, Rong Y, Ren J. Custom machine learning algorithm for large-scale disease screening-taking heart disease data as an example. Artif Intell Med 2023;146:102688.

[11] Usman TM, Saheed YK, Nsang A, Ajibesin A, Rakshit S. A systematic literature review of machine learning based risk prediction models for diabetic retinopathy progression. Artif Intell Med 2023;143:102617.

[12] Mennickent D, Rodríguez A, Farías-Jofré M, Araya J, Guzmán-Gutiérrez E. Machine learning-based models for gestational diabetes mellitus prediction before 24–28 weeks of pregnancy: A review. Artif Intell Med 2022;132:102378.

[13] Ahsan MM, Siddique Z. Machine learning-based heart disease diagnosis: A systematic literature review. Artif Intell Med 2022;128:102289.

[14] Ma H, Li D, Zhao J, Li W, Fu J, Li C. HR-BGCN: Predicting readmission for heart failure from electronic health records. Artif Intell Med 2024;150:102829.

[15] Yao L, Chu Z, Li S, Li Y, Gao J, Zhang A. A survey on causal inference. ACM Trans Knowl Discov from Data (TKDD) 2021;15(5):1–46.

[16] Blumberg CJ. Causal inference for statistics, social, and biomedical sciences: An introduction. Wiley Online Library; 2016.

[17] Belthangady C, Giampanis S, Jankovic I, Stedden W, Alves P, Chong S, Knott C, Norgeot B. Causal deep learning reveals the comparative effectiveness of antihyperglycemic treatments in poorly controlled diabetes. Nat Commun 2022;13(1):6921.

[18] Richens JG, Lee CM, Johri S. Improving the accuracy of medical diagnosis with causal machine learning. Nat Commun 2020;11(1):3923.

[19] Holland PW. Statistics and causal inference. J Amer Statist Assoc 1986;81(396):945–60.

[20] Rajendran K, Jayabalan M, Thiruchelvam V. Predicting breast cancer via supervised machine learning methods on class imbalanced data. Int J Adv Comput Sci Appl 2020;11(8).

[21] Shahmirzalou P, Khaledi MJ, Khayamzadeh M, Rasekhi A. Survival analysis of recurrent breast cancer patients using mix Bayesian network. Heliyon 2023;9(10).

[22] Jang B-S, Chun S-J, Choi HS, Chang JH, Shin KH, et al. Estimating the risk and benefit of radiation therapy in (y) pN1 stage breast cancer patients: A Bayesian network model incorporating expert knowledge (KROG 22–13). Comput Methods Programs Biomed 2024;245:108049.

[23] Ordovás JM, Rios-Insua D, Santos-Lozano A, Lucia A, Torres A, Kosgodagan A, Camacho JM. A Bayesian network model for predicting cardiovascular risk. Comput Methods Programs Biomed 2023;231:107405. http://dx.doi.org/10.1016/j.cmpb.2023.107405.

[24] Nan T, Zheng S, Qiao S, Quan H, Gao X, Niu J, Zheng B, Guo C, Zhang Y, Wang X, et al. Deep learning quantifies pathologists' visual patterns for whole slide image diagnosis. Nat Commun 2025;16(1):5493.

[25] Nan T, Ding Y, Quan H, Li D, Li L, Zhao G, Cui X. Establishing causal relationship between whole slide image predictions and diagnostic evidence subregions in deep learning. 2024, arXiv preprint arXiv:2407.17157.

[26] Dahabreh IJ, Bibbins-Domingo K. Causal inference about the effects of interventions from observational studies in medical journals. JAMA 2024.

[27] Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. J Amer Statist Assoc 2018;113(523):1228–42.

[28] Chipman HA, George EI, McCulloch RE. BART: Bayesian additive regression trees. Ann Appl Stat 2010;4(1):266–98. http://dx.doi.org/10.1214/09-AOAS285.

[29] Hill JL. Bayesian nonparametric modeling for causal inference. J Comput Graph Statist 2011;20(1):217–40.

[30] Shalit U, Johansson FD, Sontag D. Estimating individual treatment effect: generalization bounds and algorithms. In: International conference on machine learning. PMLR; 2017, p. 3076–85.

[31] Louizos C, Shalit U, Mooij JM, Sontag D, Zemel R, Welling M. Causal effect inference with deep latent-variable models. Adv Neural Inf Process Syst 2017;30.

[32] Spirtes P, Glymour CN, Scheines R, Heckerman D. Causation, prediction, and search. MIT Press; 2000, URL https://link.springer.com/book/10.1007/978-1-4612-2748-9.

[33] Heckerman D, Geiger D, Chickering DM. Learning Bayesian networks: The combination of knowledge and statistical data. Mach Learn 1995;20:197–243. http://dx.doi.org/10.1023/A:1022623210503.

[34] Bouckaert RR. Properties of Bayesian belief network learning algorithms. In: Proceedings of the tenth international conference on uncertainty in artificial intelligence. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1994, p. 102–9. http://dx.doi.org/10.1016/B978-1-55860-332-5.50018-3.

[35] Bouckaert R. Bayesian belief networks: from construction to inference [Ph.D. thesis], Utrecht, Netherlands: University of Utrecht; 1995, URL https://core.ac.uk/download/pdf/39700264.pdf.

[36] Constantinou AC. Learning Bayesian networks that enable full propagation of evidence. IEEE Access 2020;8:124845–56. http://dx.doi.org/10.1109/ACCESS.2020.3006472, URL https://ieeexplore.ieee.org/document/9136714.

[37] Constantinou AC, Liu Y, Kitson NK, Chobtham K, Guo Z. Effective and efficient structure learning with pruning and model averaging strategies. Int J Approx Reason 2022;151(C):292–321. http://dx.doi.org/10.1016/j.ijar.2022.09.016.

[38] Chickering DM, Meek C. Finding optimal bayesian networks. In: Proceedings of the eighteenth conference on uncertainty in artificial intelligence. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 2002, p. 94–102.

[39] Constantinou A. The Bayesys user manual. 2019, http://bayesianai.eecs.qmul.ac.uk/bayesys/, [Online]. [Accessed 21 April 2023].

[40] Constantinou AC, Guo Z, Kitson NK. The impact of prior knowledge on causal structure learning. Knowl Inf Syst 2023;65(8):3385–434. http://dx.doi.org/10.1007/s10115-023-01858-x, URL https://link.springer.com/article/10.1007/s10115-023-01858-x.

[41] Constantinou A, Kitson NK, Liu Y, Chobtham K, Amirkhizi AH, Nanavati PA, Mbuvha R, Petrungaro B. Open problems in causal structure learning: A case study of COVID-19 in the UK. Expert Syst Appl 2023;234:121069.

[42] Petrungaro B, Kitson NK, Constantinou AC. Investigating potential causes of sepsis with Bayesian network structure learning. Appl Intell 2025;55(6):496.

[43] Zahoor S, Constantinou AC, Curtis TM, Hasanuzzaman M. Investigating the validity of structure learning algorithms in identifying risk factors for intervention in patients with diabetes. Knowl-Based Syst 2025;114382.

[44] Kahn AB. Topological sorting of large networks. Commun ACM 1962;5(11):558–62. http://dx.doi.org/10.1145/368996.369025.

[45] Tarjan RE. Depth-first search and linear graph algorithms. SIAM J Comput 1972;1(2):146–60. http://dx.doi.org/10.1137/0201010.

[46] Kjærulff U, Van Der Gaag LC. Making sensitivity analysis computationally efficient. 2013, arXiv preprint arXiv:1301.3868.

[47] BayesFusion. GeNIe modeler USER MANUAL. 2023, https://support.bayesfusion.com/docs/GeNIe.pdf, [Online]. [Accessed 21 April 2023].

[48] Biecek P, Burzykowski T. Explanatory model analysis: explore, explain, and examine predictive models. Chapman and Hall/CRC; 2021.

[49] Marshall JC, Bosco L, Adhikari NK, Connolly B, Diaz JV, Dorman T, Fowler RA, Meyfroidt G, Nakagawa S, Pelosi P, et al. What is an intensive care unit? A report of the task force of the world federation of societies of intensive and critical care medicine. J Crit Care 2017;37:270–6. http://dx.doi.org/10.1016/j.jcrc.2016.07.015.

[50] Weil MH, Tang W. From intensive care to critical care medicine: a historical perspective. Am J Respir Crit Care Med 2011;183(11):1451–3. http://dx.doi.org/10.1164/rccm.201008-1341OE.

[51] Robinson GH, Davis LE, Leifer RP. Prediction of hospital length of stay. Health Serv Res 1966;1(3):287.

[52] Stone K, Zwiggelaar R, Jones P, Mac Parthaláin N. A systematic review of the prediction of hospital length of stay: Towards a unified framework. PLoS Digit Health 2022;1(4):e0000017. http://dx.doi.org/10.1371/journal.pdig.0000017, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9931263/.

[53] Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV (version 2.2). 2023, http://dx.doi.org/10.13026/6mm1-ek67, URL https://physionet.org/content/mimiciv/2.2/.

[54] Hempel L, Sadeghi S, Kirsten T. Prediction of intensive care unit length of stay in the MIMIC-IV dataset. Appl Sci 2023;13(12). http://dx.doi.org/10.3390/app13126930, URL https://www.mdpi.com/2076-3417/13/12/6930.

[55] WHO. Cardiovascular diseases (CVDs). 2019, https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds), [Online]. [Accessed 24 April 2024].

[56] Mahmood SS, Levy D, Vasan RS, Wang TJ. The framingham heart study and the epidemiology of cardiovascular disease: a historical perspective. Lancet 2014;383(9921):999–1008. http://dx.doi.org/10.1016/S0140-6736(13)61752-3, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4159698/.

[57] Ali MK, Galaviz KI, Weber MB, Narayan KV. The global burden of diabetes. Textb Diabetes 2017;65–83. http://dx.doi.org/10.1002/9781118924853.ch5.

[58] An X, Zhang Y, Sun W, Kang X, Ji H, Sun Y, Jiang L, Zhao X, Gao Q, Lian F, et al. Early effective intervention can significantly reduce all-cause mortality in prediabetic patients: a systematic review and meta-analysis based on high-quality clinical studies. Front Endocrinol 2024;15:1294819. http://dx.doi.org/10.3389/fendo.2024.1294819, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10941028/.

[59] CDC - Behavioral Risk Factor Surveillance System (BRFSS). 2015, http://www.cdc.gov/brfss/index.html, [Online]. [Accessed 21 March 2022].

[60] LaValley MP. Logistic regression. Circulation 2008;117(18):2395–9. http://dx.doi.org/10.1161/CIRCULATIONAHA.106.682658, URL https://www.ahajournals.org/doi/full/10.1161/circulationaha.106.682658.

[61] Suthaharan S, Suthaharan S. Decision tree learning. Mach Learn Model Algorithms Big Data Classif: Think Examples Eff Learn 2016;237–69. http://dx.doi.org/10.1007/978-1-4899-7641-3_10.

[62] Rigatti SJ. Random forest. J Insur Med 2017;47(1):31–9. http://dx.doi.org/10.17849/insm-47-01-31-39.1, URL https://pubmed.ncbi.nlm.nih.gov/28836909/.

[63] Vapnik VN. The nature of statistical learning theory. Springer-Verlag; 1995.

[64] Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inform Theory 1967;13(1):21–7. http://dx.doi.org/10.1109/TIT.1967.1053964, URL https://ieeexplore.ieee.org/document/1053964.

[65] Mucherino A, Papajorgji PJ, Pardalos PM. k-nearest neighbor classification. In: Data mining in agriculture. Springer New York; 2009, p. 83–106. http://dx.doi.org/10.1007/978-0-387-88615-2_4.

[66] Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput System Sci 1997;55(1):119–39. http://dx.doi.org/10.1006/jcss.1997.1504, URL https://www.sciencedirect.com/science/article/pii/S002200009791504X.

[67] Friedman J. Greedy function approximation: A gradient boosting machine. Ann Statist 2000;29. http://dx.doi.org/10.1214/aos/1013203451.

[68] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery; 2016, p. 785–94. http://dx.doi.org/10.1145/2939672.2939785.

[69] Freund Y, Schapire RE, et al. Experiments with a new boosting algorithm. In: ICML, vol. 96, 1996, p. 148–56, https://dl.acm.org/doi/10.5555/3091696.3091715.

[70] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artificial Intelligence Res 2002;16:321–57. http://dx.doi.org/10.1613/jair.953.

[71] He H, Bai Y, Garcia EA, Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE international joint conference on neural networks. IEEE; 2008, p. 1322–8. http://dx.doi.org/10.1109/IJCNN.2008.4633969, URL https://ieeexplore.ieee.org/document/4633969.

[72] Luo C, Duan Z, Xia Z, Li Q, Wang B, Zheng T, Wang D, Han D. Minimum heart rate and mortality after cardiac surgery: retrospective analysis of the multi-parameter intelligent monitoring in intensive care (MIMIC-III) database. Sci Rep 2023;13. http://dx.doi.org/10.1038/s41598-023-29703-9, URL https://www.nature.com/articles/s41598-023-29703-9.

[73] Weiner I, Mitch W, Sands J. Urea and ammonia metabolism and the control of renal nitrogen excretion. Clin J Am Soc Nephrol : CJASN 2014;10:1444–58. http://dx.doi.org/10.2215/CJN.10311013, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4527031/.

[74] Tonelli M, Sacks F, Arnold M, Moye L, Davis B, Pfeffer M. Relation between red blood cell distribution width and cardiovascular event rate in people with coronary disease. Circulation 2008;117(2):163–8. http://dx.doi.org/10.1161/CIRCULATIONAHA.107.727545, URL https://pubmed.ncbi.nlm.nih.gov/18172029/.

[75] De Rosa S, Greco M, Rauseo M, Annetta MG. The good, the bad, and the serum creatinine: exploring the effect of muscle mass and nutrition. Blood Purif 2023;52(9–10):775–85.

[76] Puskarich MA, Nandi U, Long BG, Jones AE. Association between persistent tachycardia and tachypnea and in-hospital mortality among non-hypotensive emergency department patients admitted to the hospital. Clin Exp Emerg Med 2017;4:2–9. http://dx.doi.org/10.15441/ceem.16.144, URL https://api.semanticscholar.org/CorpusID:14065571.

[77] Cunha BA, Shea KW. Fever in the intensive care unit. Infect Dis Clin 1996;10(1):185–209.

[78] Flint AC, Conell C, Ren X, Banki NM, Chan SL, Rao VA, Melles RB, Bhatt DL. Effect of systolic and diastolic blood pressure on cardiovascular outcomes. N Engl J Med 2019;381(3):243–51. http://dx.doi.org/10.1056/NEJMoa1803180, URL https://www.nejm.org/doi/full/10.1056/NEJMoa1803180.

[79] Poznyak A, Litvinova L, Poggio P, Sukhorukov V, Orekhov A. Effect of glucose levels on cardiovascular risk. Cells 2022;11:3034. http://dx.doi.org/10.3390/cells11193034, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9562876/.

[80] Peters SA, Singhateh Y, Mackay D, Huxley RR, Woodward M. Total cholesterol as a risk factor for coronary heart disease and stroke in women compared with men: A systematic review and meta-analysis. Atherosclerosis 2016;248:123–31.

[81] Dharod A, Soliman EZ, Dawood F, Chen H, Shea S, Nazarian S, Bertoni AG, Investigators M, et al. Association of asymptomatic bradycardia with incident cardiovascular disease and mortality: the multi-ethnic study of atherosclerosis (MESA). JAMA Intern Med 2016;176(2):219–27.

[82] Gallucci G, Tartarone A, Lerose R, Lalinga AV, Capobianco AM. Cardiovascular risk of smoking and benefits of smoking cessation. J Thorac Dis 2020;12(7):3866. http://dx.doi.org/10.21037/jtd.2020.02.47, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7399440/.

[83] Tillmann T, Vaucher J, Okbay A, Pikhart H, Peasey A, Kubinova R, Pajak A, Malyutina S, Hartwig F, Fischer K, Veronesi G, Palmer T, Bowden J, Smith G, Bobak M, Holmes M. Education and coronary heart disease: Mendelian randomisation study. BMJ 2017;358:j3542. http://dx.doi.org/10.1136/bmj.j3542, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5594424/.

[84] Held C, Hadziosmanovic N, Aylward P, Hagström E, Hochman J, Stewart R, White H, Wallentin L. Body mass index and association with cardiovascular outcomes in patients with stable coronary heart disease – A STABILITY substudy. J Am Hear Assoc 2022;11. http://dx.doi.org/10.1161/JAHA.121.023667.

[85] Wei GS, Coady SA, Goff Jr. DC, Brancati FL, Levy D, Selvin E, Vasan RS, Fox CS. Blood pressure and the risk of developing diabetes in african americans and whites: ARIC, CARDIA, and the framingham heart study. Diabetes Care 2011;34(4):873–9. http://dx.doi.org/10.2337/dc10-1786, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3064044/.

[86] Rhee E-J, Han K, Ko S-H, Ko K-S, Lee W-Y. Increased risk for diabetes development in subjects with large variation in total cholesterol levels in 2,827,950 Koreans: A nationwide population-based study. PLoS One 2017;12:e0176615. http://dx.doi.org/10.1371/journal.pone.0176615, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5436642/.

[87] European Society of Cardiology. Body mass index is a more powerful risk factor for diabetes than genetics. ScienceDaily 2020. www.sciencedaily.com/releases/2020/08/200831090129.htm.

[88] DiPietro L, Buchner D, Marquez D, Pate R, Pescatello L, Whitt-Glover M. New scientific basis for the 2018 U.S. physical activity guidelines. J Sport Health Sci 2019;8:197–200. http://dx.doi.org/10.1016/j.jshs.2019.03.007, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6525104/.

[89] Sil K, Das BK, Pal S, Mandal L. A study on impact of education on diabetic control and complications. Natl J Med Res 2020;10(01):26–9, URL https://njmr.in/index.php/file/article/view/48.

[90] Diabetes UK. Cardiovascular disease and diabetes. 2024, https://www.diabetes.org.uk/guide-to-diabetes/complications/cardiovascular_disease, [Online]. [Accessed 15 April 2024].

[91] Verma S, Boonsanong V, Hoang M, Hines KE, Dickerson JP, Shah C. Counterfactual explanations and algorithmic recourses for machine learning: A review. 2022, http://dx.doi.org/10.48550/arXiv.2010.10596, arXiv:2010.10596.

[92] Mushlin SB, Greene HL. Decision making in medicine: an algorithmic approach. Elsevier Health Sciences; 2009, http://dx.doi.org/10.1016/S0377-1237(02)80153-8.

[93] Thwaites P, Smith JQ, Riccomagno E. Causal analysis with chain event graphs. Artificial Intelligence 2010;174(12–13):889–909.