


Università degli Studi di Napoli Federico II

Ph.D. Program in

Information Technology and Electrical Engineering

XXXVIII Cycle

Thesis for the Degree of Doctor of Philosophy

Causally-Driven Clinical Decision Support: Interventions, Counterfactuals, and What-If Scenarios from Real-World Data

by

Patrizia Quaranta

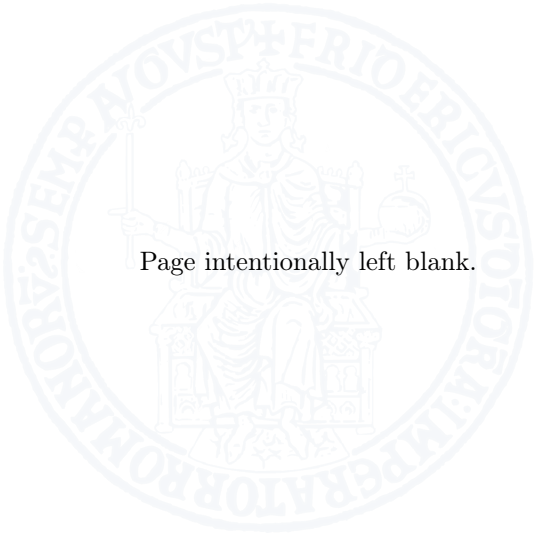
Advisor: Prof. Roberto Pietrantuono

Co-advisor: Dot. Giuseppe Cascone

Scuola Politecnica e delle Scienze di Base

Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione

Alla mia famiglia



Page intentionally left blank.

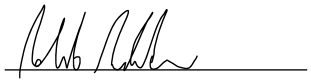
Causally-Driven Clinical Decision Support: Interventions, Counterfactuals, and What-If Scenarios from Real-World Data

Ph.D. Thesis presented
for the fulfillment of the Degree of Doctor of Philosophy
in Information Technology and Electrical Engineering
by

Patrizia Quaranta

December 2025

Approved as to style and content by



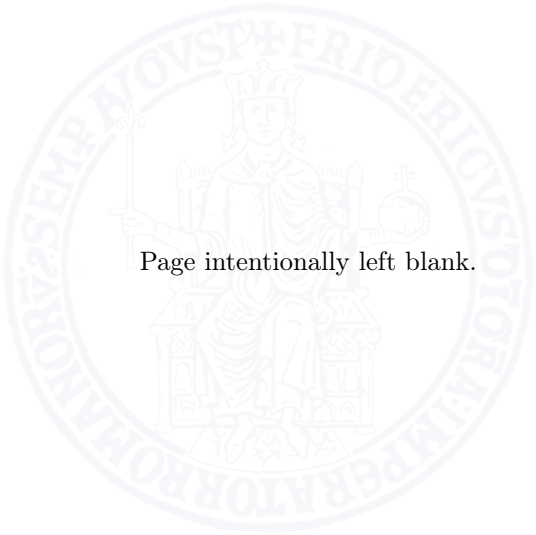
Prof. Roberto Pietrantuono, Advisor

Dot. Giuseppe Cascone, Co-advisor

Università degli Studi di Napoli Federico II

Ph.D. Program in Information Technology and Electrical Engineering
XXXVIII cycle - Chairman: Prof. Stefano Russo

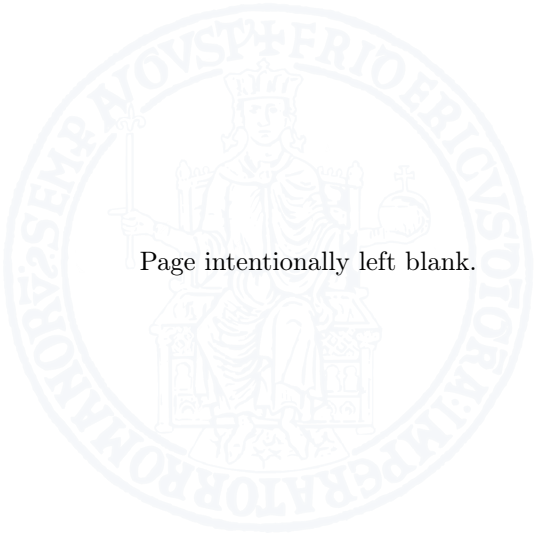
http://itee.dieti.unina.it



Page intentionally left blank.

Finanziato dall'Unione europea
NextGenerationEU

La borsa di dottorato è stata cofinanziata con risorse del Piano Nazionale di Ripresa e Resilienza, D.M. 352/2022 Missione 4 – Componente 2 – Investimento 3.3 “Dottorati innovativi”
CUP: E66G22000400009

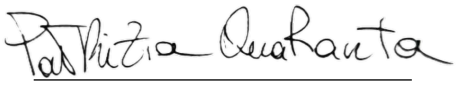


Page intentionally left blank.

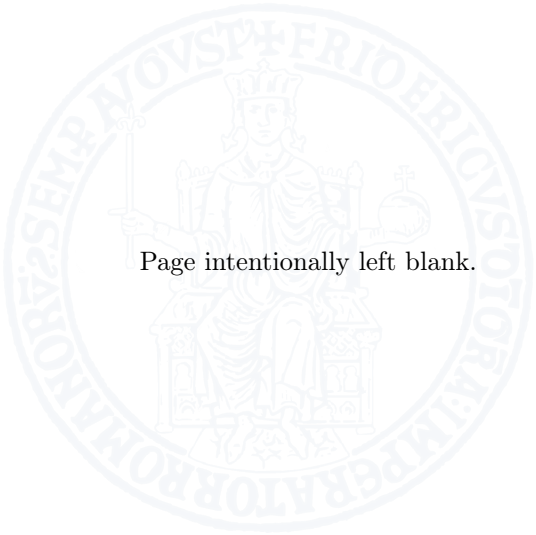
Candidate's declaration

I hereby declare that this thesis submitted to obtain the academic degree of Philosophiæ Doctor (Ph.D.) in Information Technology and Electrical Engineering is my own unaided work, that I have not used other than the sources indicated, and that all direct and indirect sources are acknowledged as references.
Parts of this dissertation have been published in international journals and/or conference articles (see list of the author's publications at the end of the thesis).

Napoli, December 5, 2025



Patrizia Quaranta



Page intentionally left blank.

Abstract

Healthcare research increasingly relies on data-driven methodologies to support clinical decision-making while avoiding expensive, hard-to-implement and sometimes unethical, controlled trials. However, traditional Machine Learning (ML) models are primarily based on statistical correlations and often fail to provide an understanding of the underlying causal relationships. In a context where interventions directly affect human well-being, purely correlational approaches may be insufficient.

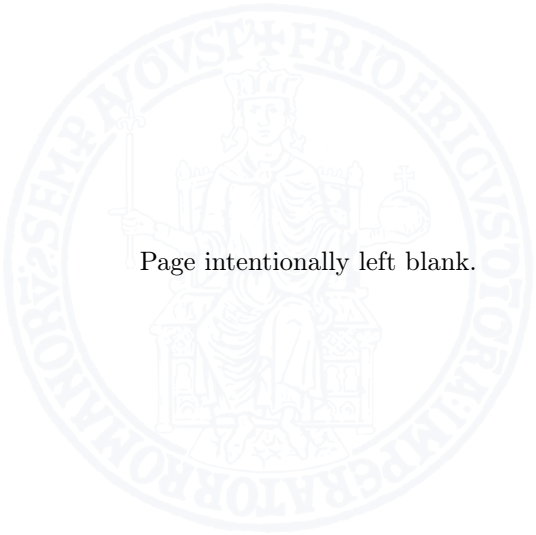
Causal Reasoning (CR) enables the estimation of intervention effects and counterfactual scenarios from observational data, offering a more interpretable and action-oriented perspective. Nevertheless, its application to real-world clinical data remains challenging due to data heterogeneity, incompleteness, and the lack of prior causal knowledge.

This thesis addresses these challenges by proposing a framework for causal discovery and inference from observational clinical data. The framework integrates data pre-processing, feature selection through ML techniques, causal structure discovery, validation using Large Language Models (LLMs), and statistical verification procedures to assess robustness. The resulting causal model supports both interventional and counterfactual queries, allowing the simulation of clinical protocols or therapeutic modifications before their real-world implementation.

The methodology was validated through two case studies. The first focuses on diabetes mellitus, using the PIMA Indians Diabetes dataset to evaluate the impact of lifestyle-related interventions. The second applies the framework to occupational health in the maritime sector, using real-world data collected from two shipping companies. In both cases, the inferred causal relationships align with clinical evidence and support what-if scenario analysis.

Overall, the results highlight the potential of causal reasoning as a bridge between predictive modeling and clinical interpretability, providing a realistic and interpretable methodology to support safer and more informed medical decision-making.

Keywords: Causal Reasoning, Causal Discovery, Intervention, Counterfactual, Diabetes, Occupational Health



Page intentionally left blank.

Sintesi in lingua italiana

La ricerca in ambito sanitario si affida sempre più a metodologie basate sui dati per supportare il processo decisionale clinico e migliorare gli esiti dei pazienti. Tuttavia, i modelli di apprendimento automatico tradizionali si limitano a individuare correlazioni statistiche, senza fornire una comprensione delle relazioni causali sottostanti. In un contesto in cui gli interventi incidono direttamente sul benessere delle persone, le sole correlazionali potrebbero risultare insufficienti.

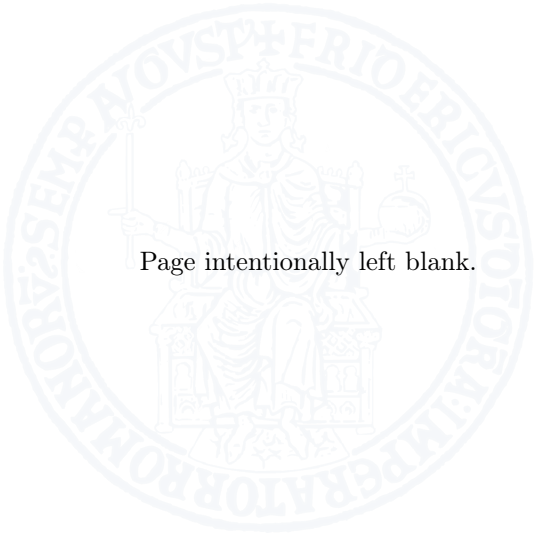
Il ragionamento causale consente di stimare gli effetti di interventi e scenari controfattuali a partire da dati osservazionali, offrendo un approccio più interpretabile e orientato all'azione. Tuttavia, la sua applicazione a dati clinici reali è ostacolata da problemi di eterogeneità, incompletezza e assenza di conoscenza causale a priori.

Questa tesi affronta tali sfide proponendo un framework per la scoperta e l'inferenza causale da dati clinici osservazionali. Il framework integra fasi di preprocessing, selezione delle variabili tramite tecniche di machine learning, scoperta della struttura causale, validazione tramite modelli linguistici di grandi dimensioni LLM e procedure di verifica statistica per valutarne la robustezza. Il modello causale risultante consente di eseguire query intervenzionali e controfattuali, simulando gli effetti di protocolli clinici o modifiche terapeutiche prima della loro applicazione reale.

La metodologia è stata validata su due casi di studio. Il primo riguarda il diabete mellito, utilizzando il dataset PIMA Indians Diabetes per valutare l'impatto di interventi sullo stile di vita. Il secondo applica il framework all'ambito della medicina del lavoro marittima, utilizzando dati reali raccolti da due compagnie di navigazione. In entrambi i casi, le relazioni causali inferite risultano coerenti con le evidenze cliniche e supportano l'analisi di scenari what-if.

Nel complesso, i risultati mostrano il potenziale del ragionamento causale come ponte tra modellazione predittiva e interpretabilità clinica, proponendo una metodologia realistica e interpretabile per supportare decisioni mediche più sicure e consapevoli.

Parole chiave: Ragionamento Causale, Scoperta Causale, Intervento Causale, Analisi di scenari alternativi, Diabete, Medicina del Lavoro.



Page intentionally left blank.

Acknowledgements

I would like to express my deepest gratitude to Prof. Roberto Pietrantuono, who has continuously supported and encouraged me throughout my Ph.D. study, even during the most challenging moments. I am sincerely thankful for his patience, motivation, and the depth of knowledge with which he has guided me during these years.

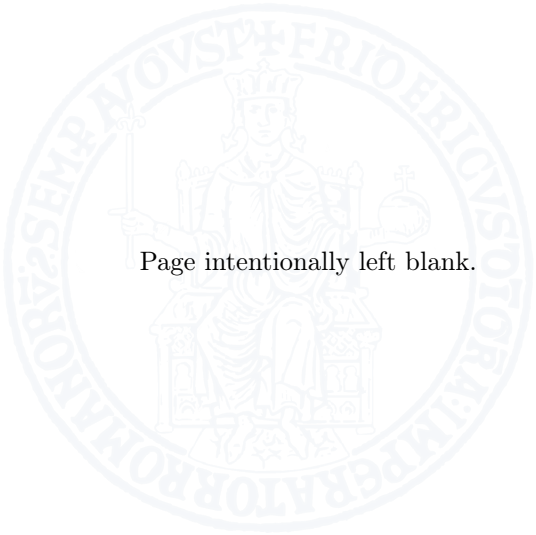
I would also like to thank Professor Barbara Moutchouri for her valuable support and guidance during my research period abroad at the University of Thessaly, Medical School, Laboratory of Hygiene and Epidemiology, Larisa, Greece.

My sincere thanks go to CML Vesuvio S.R.L. and all its members, whose support allowed me to learn a great deal and gain new perspectives in a field that was previously unfamiliar to me.

I am also deeply grateful to all my fellow Ph.D. colleagues, who shared this journey with me and helped make it easier and more enjoyable.

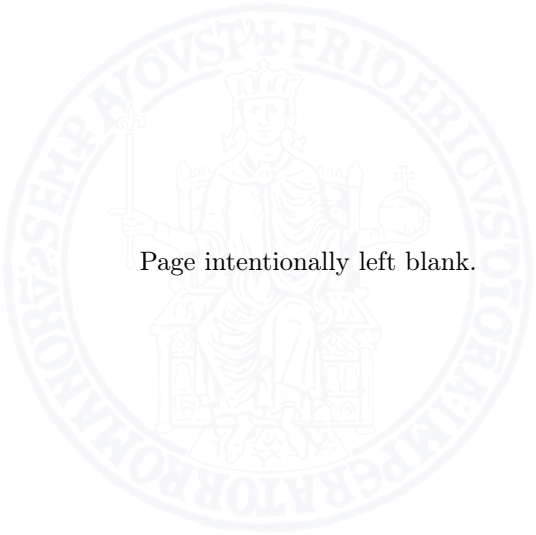
Finally, I would like to thank the DEpendable and Secure Software Engineering and Real-Time Systems (DESSERT) research group for making the lab a place of sharing — mostly of anxiety and despair, but also of knowledge.

The research presented in this dissertation has been supported by the Italian Ministry of University and Research (MUR) under program PNRR DM. 352 "Mis.: I.3.3 Dottorati innovativi".



Page intentionally left blank.

Contents



Page intentionally left blank.

List of Acronyms

The following acronyms are used throughout the thesis.

ATE	Average Treatement Effect
BMI	Body Mass Indicator
CATE	Causal Average Treatment Effect
CD	Causal Discovery
CI	Causal Inference
CM	Causal Model
CR	Causal Reasoning
DAG	Direct Acyclic Graph
DO	Deck Officers
DT	Decision Tree
ED	Engineering Department
EP	Engineering Positions
FFW	Fitness for work

FI	Features Importance
GB	Gradient Boost
LiNGAM	Linear Non-Gaussian Acyclic Model
LLM	Large Language Model
LMC	Local Markov Condition
LR	Logistic Regression
ML	Machine Learning
MLP	Multi-Layer Perceptron
OR	Odds Ratio
PCC	Professional Certification and Credentials
RCT	Randomized Controlled trial
RF	Random Forest
RQ	Research Question
RR	Relative Risk
SJ	Shipyard Jobs
SCM	Structural Causal model
SVM	Support Vector Machine
TPa	Parental Triples

List of Figures

List of Tables

Chapter 1

Introduction

To find out what happens when you change something, it is necessary to change it.

George E. P. Box

Motivation The modern era has transformed the volume and availability of data across every domain and in no field is this transformation more critical than in medicine and public health. Historically, the fundamental challenge has always been twofold: to understand the complex mechanisms that govern human health, and to determine how to intervene effectively to prevent disease and improve outcomes. In recent years, the widespread adoption of Electronic Health Records (EHRs) and other digital health systems has revolutionized data collection in healthcare, enabling the large-scale accumulation of observational clinical data. These data offer unprecedented opportunities to explore relationships between patient characteristics, treatments, and outcomes in real-world settings.

While data-driven models have been increasingly central to medical research and healthcare decision-making [50], most Machine Learning (ML) approaches remain inherently correlation-based. They identify statistical association that may predict outcomes but do not explain *why* those outcomes occur. In a field where interventions directly affect human well-being, correlations relying solely on correlational evidence may not always be sufficient, to support safe and effective decision-making.

At the same time, the theory of Causal Inference (CI) has undergone a conceptual leap with the introduction of causal graphical models and the do-calculus, developed by Judea Pearl [80]. These tools have made it possible to formally represent and analyze cause–effect mechanisms, providing a unified language for reasoning about interventions and counterfactual scenarios.

Although these classical and modern methods have been instrumental in formalizing causal reasoning and remain highly effective in controlled or well-defined populations, their direct application to real-world clinical data can be challenging. Frameworks such as structural equation modeling (SEM) or potential outcome analysis depend on strong assumptions, which are easier to ensure in randomized or demographically homogeneous settings. While Randomized Controlled trial (RCT), though methodologically rigorous, are often expensive and sometimes ethically unfeasible, continue to represent the gold standard for establishing causal effects. However, implementation can be resource-intensive and not always feasible at scale, motivating the complementary use of observational data for causal analysis. Indeed, observational healthcare data are often high-dimensional, heterogeneous, and incomplete, making some of these assumptions difficult to verify in practice. This does not undermine the validity of traditional causal inference, but rather highlights the need for complementary approaches that can accommodate the complexity of real-world data while maintaining causal interpretability.

Problem Statement Despite the theoretical progress of CI, its application to real-world healthcare data remains limited by practical constraints. Most clinical dataset are **observational**, **structured**, and **non-randomized**, where the underlying causal structures are unknown and potential confounders are only partially observed. Classical causal inference frameworks depend on predefined models or prior expert knowledge, while existing causal discovery algorithms often struggle with high dimensionality, noise, and limited interpretability. As a result, there is a persistent gap between the theoretical foundations of causal inference and its practical implementation on large-scale observational data. This gap restricts the capacity to move from correlation-based prediction to causal understanding — a crucial step toward evidence-based and interpretable

decision-making in data-driven healthcare. Addressing this challenge requires a framework capable of inferring causal structures directly from observational data, validating them through both domain knowledge and statistical robustness, and subsequently performing causal inference and effect estimation within the same pipeline.

Beyond methodological innovation, such a framework would enable the simulation and evaluation of clinical protocols in a causal context — both interventional and counterfactual — before their application to real patients. This capability would provide a powerful tool for anticipating the potential effects of medical decisions, guiding safer and more efficient clinical experimentation.

Contribution This thesis contributes to the ongoing effort to make causal inference more applicable to real-world, observational healthcare data. It presents a practical framework that combines causal discovery, domain-based validation, and causal inference within a coherent analytical process. The main contributions can be summarized as follows:

- A data-driven causal discovery approach capable of extracting causal structures from tabular observational data, without relying on predefined causal models or prior expert assumptions.
- A hybrid validation strategy that combines domain-level plausibility checking, supported by **LLMs!**s (**LLMs!**s), with statistical falsification and robustness tests to ensure consistency between inferred structures and empirical data.
- A unified process for causal inference and validation, allowing for the estimation and testing of treatment effects directly on the inferred causal model.
- An application-oriented perspective, showing how causal reasoning can support the exploration and preliminary evaluation of clinical protocols through interventional and counterfactual analysis.

Rather than aiming to solve all challenges of causal inference on observational data, this work focuses on developing a feasible and interpretable methodology that leverages modern computational tools to enhance the robustness and usability of causal analysis in healthcare research.

The rest of thesis is structured as follows.

Chapter 2 introduces the background concepts underlying Causal Reasoning (CR), the statistical validation techniques adopted in this work, and an overview of Large Language Models (LLMs) used as domain expert. It also presents the related work on the use of causal reasoning — in the sense formalized by Pearl's framework [80] — within the healthcare domain.

Chapter 3 describes the methodology proposed in this thesis, detailing each component of the framework for causal discovery, validation, and inference.

Chapter 4 presents an application of the proposed methodology to a diabetes-related problem, using the well-known PIMA Indians Diabetes dataset.

Chapter 5 extends the analysis to a real-world case study, where the framework is applied to data collected from two maritime companies during occupational health examinations.

Chapter 6 summarizes the conclusion.

Chapter 2

Background and Related Work

Causation is not merely a relation between events; it is the cement of the universe.

David Hume

Since the earliest efforts to interpret observations, humans have sought to identify the mechanisms that generate events and to imagine how outcomes would change under different conditions. This intrinsic tendency to reason in causal terms anticipates the formal frameworks that now underpin modern data analysis, modeling, and causal inference. CR provides the theoretical foundation for understanding how and why different factors within a system interact and influence each other. In contrast to purely statistical approaches, which describe associations among observed quantities, causal analysis aims to uncover the underlying mechanism that generates those associations and to predict the effects of interventions or hypothetical changes.

Over the past decades, the study of causality has evolved from a philosophical question into a formal scientific discipline, supported by well-defined mathematical frameworks and computational methods that find application across numerous scientific domains- including the natural and social sciences[71, 46], medicine [68, 92], epidemiology [39], and economics [44]. These developments have made it possible to address causal questions even in complex, high-dimensional settings, where controlled experimentation is impractical, and data are primarily observational.

The chapter is organized as follows: Section 2.1 introduces the foundations of causal reasoning, clarifying its distinction from statistical correlation and presenting its main components — the theoretical frameworks that formalize causal relations, the approaches for causal discovery from data, and the principles of causal inference, including interventional and counterfactual reasoning. Section 2.2 describes experimental and observational study designs, outlining their methodological differences, respective strengths and limitations, and their roles in establishing causal effects. Section 2.3 discusses model validation and robustness in causal inference, covering methods for falsification, refutation, and the validation of Causal Average Treatment Effect (CATE) estimates. Section 2.4 provides an overview of LLM-based causal reasoning, reviewing recent developments and the current state of the art in the integration of LLMs with causal analysis.

2.1 Causal Reasoning

CR differs fundamentally from purely statistical reasoning. Statistical analysis captures **correlations** between factors, revealing that some factors tend to co-occur or change together. However, correlation alone cannot determine whether one factor truly causes another, whether the observed relationship is spurious, or whether both depend on an unobserved confounder. Causal reasoning, instead, aims to uncover directional, generative relationships that explain how changes in one part of a system produce changes elsewhere [80, 84]. The levels of causal understanding are clearly illustrated by Pearl and Mackenzie through the theory of **Ladder of Causality**[79], which organizes causal reasoning into three conceptual levels:

- **Seeing (Correlation)**: Describes statistical dependence between factors, expressed as $P(Y \mid X)$. It relies solely on what can be observed, and underlies most traditional data analysis and ML methods.

- **Doing (Intervention)**: Concerns the effect of actively changing a factor, expressed as $P(Y \mid do(X = x))$. It evaluates the impact of

actively setting one factor to observe the effect on another- answering questions such as "What happens if we set X equal to x?"

- **Imagining (Counterfactual)**: Involves reasoning about hypothetical scenarios, answering questions such as "What would have happened if X had been different?" Counterfactual reasoning supports explanation, accountability, and individualized decision-making.

Each level subsumes the previous one: knowing how interventions affect outcomes implies knowledge of correlations, but the reverse is not true. Importantly, moving up the ladder requires additional assumptions about the underlying data-generating process, such as causal structure, independence relations, or the presence of confounding. These assumptions are what distinguish causal reasoning from purely empirical or predictive modeling. Causal reasoning links data analysis with causal understanding, allowing correlations to be interpreted in terms of cause and effect.

2.1.1 Causal Framework

Causal frameworks provide mathematical or graphical representations of the relationships among factors within a system or population [75]. They incorporate prior or domain-specific knowledge - assumptions about the data-generating process and plausible causal mechanisms - which distinguishes them from purely associational or correlational approaches [51]. This section presents the main causal frameworks applied in causal reasoning and inference [32].

Causal Direct Acyclic Graph (DAG) (CausalDAG) : represents a system of causal relationships through nodes and directed edges. Nodes correspond to variables - including the treatment, the outcome, and other observed and unobserved variables- while directed edges denote causal influence. The acyclic structure ensures that no variable can be its own cause, maintaining a unidirectional causal flow. Formally, a CausalDAG $G = (\boldsymbol{X}, \mathscr{E})$ is a directed acyclic graph that describes causal effects among variables, where $\boldsymbol{X}$ is the node set and $\mathscr{E}$ the edge set. In a CausalDAG, a directed edge $X \to Y$ indicates that Y has a causal effect on Y. The CausalDAG encodes assumptions about conditional independencies among variables and implies a joint probability distribution that is consistent with

the graph structure. This consistency, known as the Markov property, establishes the connection between the graphical representation and the underlying probabilistic model [80, 82, 84].

Bayesian Network (BN) extends DAGs by incorporating probabilistic dependencies. A BN encodes conditional independencies among variables and defines a probability distribution consistent with the graph structure [75]. A BN is formally defined as: $B = (G, P)$ where P factorizes over the graph G and is expressed as a product of conditional probability distributions associated with G's nodes $P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i | Pa_{X_i}^{\mathcal{G}})$ where X_i are the nodes of G and $Pa_{X_i}^{G}$ are their parents, namely the nodes with directed edges pointing to X_i. While BNs encode probabilistic relationships, they do not inherently represent causality. This factorization allows for a compact representation of complex probabilistic dependencies and facilitates reasoning under uncertainty.

Structural Causal model (SCM) extends CausalDAGs by associating each node with a structural equation that formally defines how the variable is generated from its causes. This formulation connects the graphical representation of causality with the underlying functional mechanisms that govern the data-generating process. A **Functional Causal Model (FCM)** assumes that each variable X_i is a deterministic function of its direct causes and an exogenous noise term, expressed as $X_i = f_i(Pa(X_i, U_i))$ where $Pa(X_i)$ denotes the set of parent variables of X_i in the graph and U_i represents independent noise capturing unobserved influences. This formulation makes it clear that variability in the data is due to external disturbances, not to randomness in the causal mechanism [54]. An SCM formalizes a collection of such functional relationships within a CausalDAG $\mathcal{G} = (\boldsymbol{X}, \mathscr{E})$ where the set of edges $\mathscr{E}$ encodes the causal dependencies among variables. These structural assignments define the random variables X_i as functions of their parents $Pa(X_i)$ and of independent noise variables U_i. In this framework, X_i are *endogenous* - determined by the structural equations and their parent variables- while U_i are considered *exogenous*, representing independent external influences not explained within the model. The SCM therefore integrates both the graphical structure and the functional assignments, offering a complete representation of the sys-

tem's causal semantics. SCMs generalizes traditional Structural Equation Models (SEMs), allowing for nonlinear and nonparametric relationships between variables. They constitute the formal foundation of modern causal inference and enable reasoning about interventions and counterfactual scenarios in complex systems [79, 80].

Potential Outcome (PO) framework, also known as the Neyman–Rubin Causal Model (CM), provides a conceptual foundation for reasoning about cause and effect through counterfactual comparison [90, 34, 80, 75]. It focuses on comparing the outcome that actually occurred with the outcome that would have been observed under an alternative scenario. In this formulation, causality is defined as the difference between potential outcomes corresponding to different treatment conditions. For each individual u, two potential outcomes are defined: $Y_{T=1}(u)$ and $Y_{T=0}(u)$, representing the outcome if the individual receives respectively the treatment $T = 1$ or does not receive the treatment $T = 0$. In this sense, causal reasoning is inherently counterfactual: it involves comparing the observed outcome with the hypothetical outcome that would have been realized under an alternative intervention. Formally, a PO is defined as $Y_{T=t}(u)$ of variable Y is the value that Y would take for individual u if the treatment T assumed the value t. Only one of these two outcomes can be observed for each unit- the treated or the untreated- while the other remains counterfactual. This condition, referred to as the fundamental problem of causal inference, prevents the direct observation of causal effects at the individual level. The causal effect for a single unit is expressed through the **Individual Treatment Effect** (ITE), defined as: $ITE = Y_{T=1}(u) - Y_{T=0}(u)$. Because both potential outcomes for the same unit cannot be observed simultaneously, causal inference relies on estimating expected effects across a population. The expected value of individual effects, at the population level, corresponds to the Average Treatment Effect (ATE), where $ATE = E[Y_{T=1} - Y_{T=0}]$. Depending on the context, related quantities can also be considered, such as the Average **Treatment Effect on the Treated** (ATT) or CATE[75]. The ATT represents the average effect among treated units, while the CATE captures heterogeneous causal effects within a population and is formally defined in a later section. Through its counterfactual reasoning paradigm, the PO framework provides a rigorous and intuitive founda-

tion for causal analysis, serving as the theoretical basis for numerous estimation methods, including randomized experiments, matching, weighting, and other counterfactual-based techniques.

Interventional Framework introduced by Pearl [80] provides a formal language to represent and reason about external manipulations within a causal system. Within this formalism, causality is expressed through the operator $do(\cdot)$, representing the effect of external manipulations on the data-generating process. An intervention on a variable X, represented as $do(X = x)$ simulates the effect of an external manipulation that fixes X to the value x while leaving all other structural relations in the system unchanged. Graphically, this corresponds to removing all incoming edges into X - that is, the edges from its parent nodes $P_a(X)$- thereby isolating X from its direct causes. This graphical intervention effectively blocks any spurious associations transmitted through these parent nodes, which often act as confounders of the relationship between X and Y. The resulting post-intervention distribution $P(Y|do(X = x))$, quantifies the effect of the intervention on the outcome variable Y. It differs from the observational conditional probability $P(Y|X = x)$, which reflects statistical associations that may be confounded by common causes of X and Y [79].

2.1.2 Causal Discovery

Causal Discovery (CD) aims to identify the set of causal dependencies among variables and to represent them through a graphical representation - typically a CausalDAG- that makes the underlying cause–effect structure explicit [75]. It provides the methodological basis for inferring how variables influence one another when the underlying causal structure is not known a priori. The main challenge lies in reconstructing causal relations from observational data, where interventions are not available, and multiple models may be statistically indistinguishable. To achieve identifiability, CD methods rely on assumptions about the data-generating process—such as conditional independence, functional relations, or distributional asymmetries—that constrain the set of plausible causal graphs. CD algorithms operate under the assumption that causal relationships can be derived from statistical dependencies. To derive causal structure from data in a principled way, these algorithms rely on a set of foundational assumptions

that connect probabilistic patterns in the data to causal relations in the underlying model, among which, the *Causal Markov*, *Faithfulness*, and *Sufficiency* assumptions [33, 75, 114]. The *Causal Markov* condition states that each variable X in a DAG G is probabilistically independent of its non-descendants given its parents. This property ensures that a causal graph implies a set of conditional independence relations among variables, which can be formally characterized through the concept of *d-separation*. The variables X and Y are *d-separated* given a conditioning set Z if, in the graph, all paths between them are blocked by Z variable. When this occurs, X and Y are conditionally independent in any probability distribution that is Markov to G[80, 84, 103]. The *Faithfulness* condition complements the Markov property by assuming that the independencies present in the data correspond precisely to those implied by *d-separation* in the causal graph [34]. This implies that no additional or accidental independencies arise from specific parameterizations of the underlying probability distribution. Markov and Faithfulness together define an equivalence relation over causal graphs where all graphs within the same class encode the same set of (conditional) independence relationships. Finally, the *Sufficiency* condition requires that, for any pair of observed variables, all their common causes are also included among the observed variables. This assumption ensures that no unmeasured confounders induce spurious dependencies that could distort the inferred causal structure. Building on these theoretical foundations, different approaches to CD have been developed, each relying on distinct assumptions and methodological principles. From a methodological perspective, causal discovery approaches can be classified into five main categories: constraint-based, score-based, continuous optimization-based, functional causal model-based, and hybrid strategies. These methodological categories, as illustrated in Figure 2.1, are adapted from the taxonomy proposed by [114].

Constraint-based strategies infer a causal structure through statistical tests of (conditional) independence between variables[103]. Based on causal Markov and faithfulness assumptions, these methods identify edges and directions by evaluating which conditional independencies are supported by the data. Classical examples include the *PC* (Peter-Clark) algorithm [103], its extensions such as *FCI* and *RFCI* for handling latent

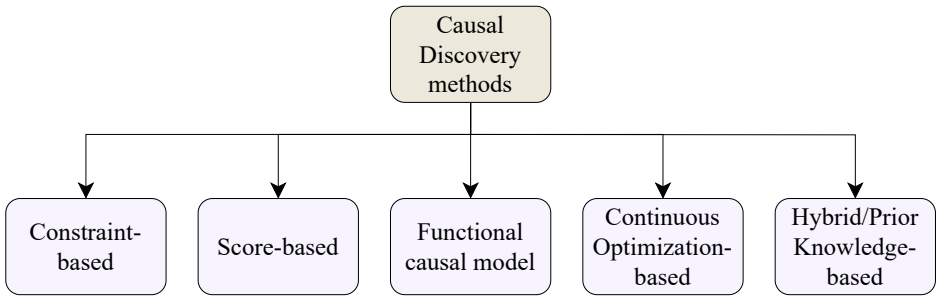


Figure 2.1. Schematic representation of causal discovery methods, adapted from [114]

confounders [20], and scalable variants like *FastPC* [57].

Score-based strategies formulate CD as an optimization problem. They assign a numerical score—typically based on likelihood, penalized likelihood, or information criteria—to each candidate graph and search for the one that best balances model fit and complexity [67]. These methods map candidate DAGs to evaluatable numerical scores that quantify how well each structure explains the observed data, enabling direct comparison among alternative causal graphs[114]. In practice, algorithms optimize a scoring function defined over the space of candidate DAG, often exploiting adjustment criteria such as the *Bayesian Information Criterion*, which approximates the posterior probability of a model given the data. They relax the *faithfulness* assumption, allowing for the estimation of causal structures even when statistical dependencies deviate from perfect d-separation relations. Representative algorithms include *GES* (Greedy Equivalence Search)[18], *FGES* [87], *GIES* for interventional data [38]. Despite their theoretical soundness, score-based methods are computationally demanding, since the number of possible DAGs grows super-exponentially with the number of variables.

Functional Causal Model-based strategies aim to identify the true causal structure by determining the causal direction of edges among variables, distinguishing the correct model from all graphs within a Markov equivalence class. These algorithms exploit assumptions about the func-

tional form of the causal mechanisms that generate the data. By leveraging properties such as *nonlinearity*, *non-Gaussianity*, or *additive noise*, they can infer causal direction even when statistical independencies are insufficient to ensure identifiability. Representative examples include Linear Non-Gaussian Acyclic Model (LiNGAM) [99], which assumes linear relations with non-Gaussian disturbances. Among the main advantages of FCM-based methods are their ability to achieve identifiability within Markov equivalence classes, the explicit representation of data-generating mechanisms through structural functions, and the natural extension to interventional and counterfactual reasoning. Their flexibility and reduced reliance on strict statistical assumptions make them particularly suitable for modeling complex or nonlinear systems [84].

Continuous optimization-based strategies reformulate CD as a differentiable optimization problem under acyclicity constraints. These methods combine the benefits of *score-based* methods, with *gradient-based* optimization enforcing the DAG constraint via smooth regularization [114]. These methods are ML-based and the most representative is *NOTEARS*, which introduced a differentiable acyclicity function [121], and *DAG-GNN* [117].

Hybrid/ Prior knowledge-based strategies combine ideas from constraint-based and score-based paradigms to improve robustness and computational efficiency. Methods such as *MMHC [107]* use independence tests to restrict the search space and scoring functions to refine edge orientations, merging the interpretability of statistical tests with the optimization efficiency of continuous search. Recent frameworks, such as *Joint Causal Inference* (JCI) [70], generalize this hybrid perspective by integrating heterogeneous sources of information—observational, interventional, and contextual—into a unified causal discovery framework. JCI enables reasoning across multiple environments or experimental conditions and allows for the simultaneous identification of causal relations and context-specific mechanisms within a single coherent model. Together, these approaches provide complementary strategies for uncovering causal structure under different data conditions and modeling assumptions. The resulting causal graphs constitute a fundamental step for subsequent tasks such as intervention

modeling, effect estimation, and counterfactual analysis within SCMs.

2.1.3 Causal Inference

Once a causal model is established, it can be used to perform CI — that is, to formulate and evaluate queries about the consequences of potential changes within a system. In essence, CI aims to estimate the impact that altering a variable has on an outcome of interest[80, 75]. This purpose differs from that of conventional ML, which focuses on uncovering statistical patterns in the data. ML answers question such as: *What outcome Y do we expect if the input X happens to be equal to x?*, which corresponds to computing the conditional probability $P(Y|(X = x))$. Such questions capture associational relationships that describe how variables co-vary in the observed data. CI, instead, targets causal relationships, and it addresses more explanatory queries such as: *"What outcome Y do we expect if we actively set X to x?"* - an **interventional** queries expressed as $P(Y|do(X = x))$; and *"What would have happened to Y if we had set X to x?"* — a counterfactual question, concerning hypothetical scenarios different from the observed reality. Whereas machine learning predicts outcomes based solely on what has been seen, causal inference allows reasoning about what would happen—or would have happened—under interventions. This capability enables not only explanation but also the design of actions and policies aimed at influencing outcomes [80, 79]. These causal questions motivate a range of methodological strategies that translate theoretical concepts— such as interventions and counterfactuals— into practical tools for effect estimation [34, 46, 80, 84]. Reaching this objective typically involves two main steps: *identification* and *estimation*. The identification step determines whether a causal effect can be expressed in terms of statistically observable quantities[79]. This process involves distinguishing genuine causal effects from spurious associations and relies on the assumptions encoded in the causal model. When confounding or unobserved variables prevent direct identification, additional assumptions or model refinements may be required to establish valid causal relationships. In graphical causal models, this often involves applying identification strategies such as the back-door, front-door, or mediator criteria, which provide conditions under which causal effects can be expressed in terms of observable quantities [79, 80]. Once a causal effect has been identified, different

methodological strategies can be employed to *estimate* it from data. Depending on the available information and the nature of the data, several methodological methods have been developed to perform this estimation, ranging from randomized experiments to matching, statistical adjustment, and machine learning–based approaches. These approaches differ according to the assumptions made about the presence of unobserved confounders and the complexity of the data. An overview of the main causal inference methodologies, adapted from [34], is illustrated in Figure 2.2. Broadly,

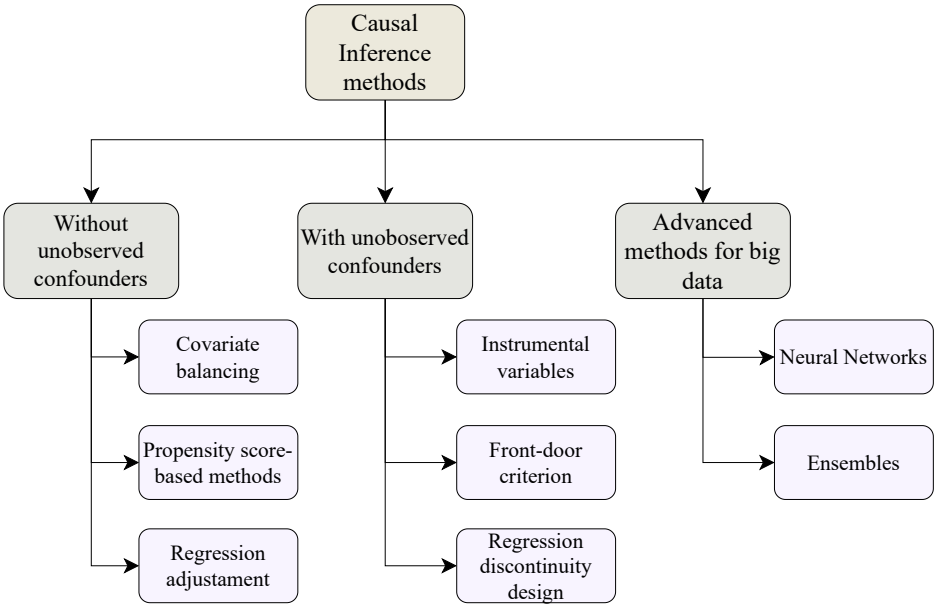


Figure 2.2. Schematic representation of causal inference methods, adapted from[34]

these approaches can be categorized into three main groups: methods that assume the absence of unobserved confounders, methods that explicitly handle unobserved confounding, and advanced methods designed for high-dimensional or large-scale data. In the ideal setting where all relevant confounders are observed, causal effects can be estimated through methods that rely on the conditional independence (or unconfoundedness) assumption. This condition implies that, given a suitable set of covariates, treatment assignment is independent of potential outcomes. Within this

framework, several classical methods can be applied, including regression adjustment, covariate balancing, and propensity score–based techniques such as matching or inverse probability weighting [14, 109, 89] These approaches aim to approximate the conditions of a randomized experiment by controlling or balancing observed covariates, thus allowing unbiased estimation of treatment effects such as the Average Treatment Effect (ATE) or CATE. In many real-world contexts, however, the assumption of no unobserved confounders may not hold. However, in many real-world scenarios, some confounders remain unobserved, and conditional conditional independence assumption is violated. When hidden confounding is present, identification must rely on alternative strategies that exploit additional information or structural properties of the system. *Instrumental variable* (IV) methods, for instance, make use of variables that affect the treatment but have no direct influence on the outcome except through it. Other strategies, such as the *front-door criterion* [80], leverage observed mediators to block indirect confounding paths, while regression discontinuity designs (RDD) approximate randomization by exploiting threshold-based treatment assignments [45]. These techniques extend causal inference to more complex or imperfect data-generating settings, though at the cost of stronger structural assumptions or domain-specific constraints. Finally, the increasing availability of large-scale and high-dimensional data has fostered the development of advanced methods that integrate causal inference with machine learning. Neural network–based architectures, ensemble learners, and representation learning models can capture complex nonlinear relationships while maintaining causal interpretability [34]. Such data-driven approaches enable the estimation of heterogeneous and individualized effects, expanding causal inference beyond traditional parametric frameworks. Recent generative approaches, such as the *Causal Effect Variational Autoencoder* (CEVAE) [63] and the *Do-sampler* [11], extend this integration by explicitly modeling latent confounders and approximating interventional or counterfactual distributions through deep generative mechanisms. While *CEVAE* leverages deep latent-variable models to infer causal effects in the presence of unobserved confounders, the *Do-Sampler* enables practical causal inference within graphical causal models by simulating interventional distributions, thus supporting a wide range of causal queries beyond effect estimation, including attribution, root-cause analy-

sis, and distribution-shift diagnosis. Although conceptually distinct, both methods share the goal of operationalizing causal reasoning in complex or partially observed systems, bridging the gap between theoretical causal models and data-driven estimation. These methods represent a promising direction for estimating causal effects in complex, high-dimensional settings where traditional assumptions or parametric forms are insufficient. Despite their methodological differences, all these approaches pursue a common aim — to make causal reasoning operational, allowing the estimation and interpretation of cause–effect relations consistent with the underlying data-generating process.

2.2 Experimental and Observational Study Designs

Understanding the distinction between experimental and observational study designs is essential for causal inference, as it defines the degree of control over treatment assignment and the strategies required to mitigate bias in real-world data. In CI research, the traditional benchmark is the RCT, which represented the gold standard for establishing cause-effect relationships because it minimizes confounding. Indeed, in RCTs, participants are randomly assigned to treatment or control groups, thereby minimizing confounding and ensuring that observed differences in outcomes can be more confidently attributed to the intervention itself. However, in ML and many real-world applications, the available data is come from observational sources, often in tabular form. This difference introduces specific methodological challenges: while RCTs ensure direct control over treatment assignment, observational data require preprocessing procedure and analytical strategies designed to reduce bias and make their analysis more reliable. Observational studies can take different forms depending on the temporal structure and the way data are collected. The three main approaches are: cross-sectional, which provides a snapshot of the population at a single point in time; cohort, which follows groups of individuals over time based on shared characteristics; and case-control, which starts from the outcome to trace back potential exposures [53],[24],[73]. Each design has strengths and limitations that must be carefully considered in building the methodological framework. Table 2.1 summarizes the main differences between the types of study considered in this work.

Table 2.1. Comparison of Study Designs

	Observational Study		**Experimental study**
	Cross-Sectional	**Cohort**	**RCTs**
Definition	Data on exposures and outcomes are collected at a single point in time.	A group (cohort) sharing a characteristic is followed over time to study outcomes.	Participants are randomly assigned to treatment and control groups to test causal effects.
Exposure assignment	Measured at the same time as outcome, not manipulated.	Based on membership in the cohort (exposed vs. non-exposed).	Controlled and randomized by the researcher.
Time dimension	Snapshot: single time point.	Longitudinal: prospective or retrospective.	Prospective: follow-up after intervention.
Typical purpose	Estimate prevalence, explore correlations, generate hypotheses.	Estimate incidence and relative risk; study temporal relationships.	Establish causal effects with high internal validity.
Advantages	Quick, cheap, good for prevalence.	Good for rare exposures; clear temporal sequence.	Minimizes confounding; strong causal evidence.
Limitations	Cannot establish temporal order; confounding risk.	Time-consuming, expensive; attrition bias.	Costly, ethical constraints.

2.3 Model Validation and Robustness in Causal Inference

Causal reasoning aims to identify and estimate causal relationships between variables. However, deriving a causal model is the one step in the broader process of causal analysis. Once a model has been proposed, it becomes essential to evaluate if its implications are consistent with the

observed data and if the estimated causal effects are robust to possible unobserved factors [81]. This stage is known as model validation. Model validation in causal reasoning seeks to evaluate both the *plausibility* and the *stability* of the inferred causal structure. *Plausibility* reflects how well the causal assumptions of a model are supported by empirical evidence-for instance, whether the conditional independencies predicted by its structure are observed in the data. *Robustness* captures the sensitivity of causal estimates to changes in model specification, data, or unobserved confounding. Together, *plausibility* and *robustness* ensure that the causal conclusions are both theoretically justified and empirically supported. Several strategies have been developed to address this goal, the major ones being **falsification tests**, **refutation tests**, and **CATE validation** focuses on evaluating the validity of heterogeneity in estimated causal effect.

2.3.1 Falsification Test

The **falsification test** aims to answer the following question: *"Can the proposed causal model survive empirical testing? Is it empirically plausible?"*. Falsification reflects the idea that CMs gain credibility by exposing their assumptions to potential empirical refutation [85]. Being *falsifiable* means that a model can be tested with data—its assumptions can, in principle, be proven wrong. A falsified model, instead, is one whose predictions or implied relations are contradicted by the evidence. In causal inference, this principle translates into verifying the internal consistency between the assumed causal structure and the data, assessing whether the conditional independencies implied by a model align with those observed empirically [85, 103, 80]. A CM that is *falsifiable* yet not *falsified* can therefore be considered both empirically coherent and informative, whereas one that cannot be falsified lacks explanatory power [85]. Several approaches can be used to test if a CM is *falsifiable* or *falsified*. The most common involves testing conditional independence constraints (CIs) implied by a given causal graph [103, 80, 84]. Other approaches include **placebo or negative control tests**, which check for spurious causal effects where none should exist [60, 82], and **invariance-based methods**, which examine whether the model's implied relationships remain stable across different environments or subpopulations [83]. In summary, falsification serves as the first step in assessing the empirical adequacy of a

CM. By testing whether the model's assumptions and implied relationships hold in the data, researchers can identify inconsistencies early and refine or reject models that fail to align with empirical evidence. Subsequent validation steps, such as refutation and robustness analysis, build on this foundation to evaluate the stability and credibility of the inferred causal effects.

2.3.2 Refutation Test

After assessing the empirical plausibility of a CM through falsification, the next step is to evaluate the robustness of the causal conclusions it produces. It aims to answer the following question: *"Do our causal conclusions remain stable when the model or data are perturbed?"*. Refutation test is designed for this purpose: it assesses whether the estimated causal effects remain stable under plausible variations in modeling assumptions or data structure[80]. The central idea is that reliable causal estimates should not depend excessively on any single assumption, sample, or modeling choice. If small perturbations in the data or analytical setup lead to large changes in the estimated effects, the results may reflect model fragility rather than genuine causal relationships [46, 19, 80]. Several Types of refutation tests are commonly used:

- **Placebo**: introduce a fictitious treatment or outcome that, by design, should have no causal effect. Detecting a significant association in this setting indicates potential confounding or model misspecification [46, 97].
- **Subset or Sample-Split**: recompute causal estimates across different subsamples or population subgroups. Consistency of results across these subsets supports the robustness of the causal conclusions [46, 97].
- **Confounder Sensitivity**: evaluate how the estimated effects change when potential confounders are added, removed, or replaced, helping to identify the influence of omitted variable bias [46, 97].
- **Random Noise**: introduce controlled random perturbations in the data or variable assignments. If the inferred effects disappear or vary

widely under randomization, the original findings may lack robustness [97].

Refutation tests provide an empirical assessment of the robustness of causal conclusions. By perturbing the data or relaxing key assumptions, they help determine whether the estimated causal effects reflect genuine causal relationships or artifacts of model specification. Having established the plausibility and stability of the CM, the next step is to validate the reliability of the estimated treatment effects — particularly their conditional variations across individuals or subgroups.

2.3.3 Validation of CATE Estimates

"Do the heterogeneous treatment effects we estimate reflect true causal differences?" This is the central aim of **CATE validation**, which seeks to determine if the observed heterogeneity in estimated effects reflects true causal heterogeneity rather than statistical noise or model bias[113]. Building on the potential outcome [90]- previously introduced - the CATE is defined as $CATE(x) = \mathbb{E}[Y(1) - Y(0)|X = x]$ capturing the expected causal effect for subpopulations defined by covariates X. While the Average Treatement Effect (ATE) provides a single population-level summary, the CATE expresses this effect conditional on observed covariates, thereby allowing for the study of treatment effect heterogeneity[112, 113]. Understanding how treatment effects vary across individuals or groups has become a central theme in modern causal inference and supports personalized decision-making and targeted interventions [116]. However, estimating heterogeneous effects introduces additional sources of uncertainty. The challenge lies in ensuring that the detected heterogeneity reflects true causal variation rather than spurious patterns resulting from overfitting, limited overlap, or violations of causal assumptions. For this reason, validation of CATE estimates focuses on evaluating their reliability, stability, and interpretability—typically through benchmarking, policy risk evaluation, or other model-agnostic performance metrics. These approaches assess whether estimated CATEs align with empirical evidence and theoretical expectations, providing a foundation for credible individualized causal inference. Several approaches have been proposed to evaluate the reliability of CATE estimates, differing in their assumptions and objectives. **Bench-**

marking methods compare estimated treatment effects against reference models, simulated ground truths, or experimental data when available [40]. **Policy evaluation metrics** assess whether the estimated heterogeneity leads to effective and consistent treatment assignment decisions, often using measures such as policy risk or value functions [113]. **Score-based and calibration metrics** quantify how closely estimated CATEs align with true or expected causal effects [95]. Validating CATE estimates is essential for distinguishing meaningful causal heterogeneity from artifacts of model specification or sampling noise. It provides the basis for credible individualized causal inference and serves as a foundation for methods that exploit heterogeneous effects for decision-making.

2.4 LLM-based Causal Reasoning: State of the Art

The intersection of causal reasoning and LLMs has recently emerged as a promising research direction that investigates the use of LLMs as *virtual domain experts* to support the discovery and validation of causal approaches. In this contest, Kiciman et al. conduct a "behavioral" study evaluating the causal reasoning abilities of LLMs [52]. Their work shows that LLMs can infer cause-effect directions from variable names with very high accuracy. Beyond type-level relations, LLMs also demonstrates competence in token-level causal reasoning, including counterfactual judgments and the identification of necessary and sufficient causes. Authors also show that LLMs can generalize across different domains, including medical applications (e. g. diagnostic reasoning) which makes them particularly relevant in contexts where domain expertise is crucial. Despite these strengths, the authors highlight some limitations, including sensitivity to prompt design, occasional inconsistencies in reasoning, and difficulties in socially nuanced judgments. Kiciman's work implies that LLM can be used as *virtual domain expert*. From this perspective, Vashishtha et al.[110] proposed leveraging LLMs in causal discovery by employing them to estimate a causal topological order through triplet-based prompting. This order is then used to guide classical discovery algorithms thereby improving the robustness of causal effect estimation, particularly in small-sample settings. In a later work, Vashishtha et al. [111] argue that, instead of asking an expert (hu-

man or LLM) to output a full causal graph, it is more stable to request a causal topological order- e. g. a partial ordering of variables consistent with causal relationships. They show that even a perfect expert with pairwise prompts can lead to errors in the inferred graph, while the causal order remains correct. They propose a triplet prompting strategy: for each pair of variables, an auxiliary variable is introduced and the LLM is asked to decide directions within the triple so as to avoid cycles; outcomes are aggregated by voting. Empirical results on multiple real-world graphs show that this method yields more accurate orders than pairwise prompting for both LLMs and human annotators. Moreover, they demonstrate how the expert-provided order can be employed to reduce errors in downstream tasks of causal discovery and effect estimation. On other aspects, Jiralerspong et al. [48] focused their attention. Their work addressed the inefficiency of pairwise prompting by introducing a breadth-first search (BFS) strategy, which reduces the number of LLM queries from $O(n^2)$ to $O(n)$. The method iteratively identifies root nodes and expands successors while enforcing acyclicity and can incorporate observational statistics such as correlations into the prompts. This approach achieved state-of-the-art performance on benchmark datasets, particularly for larger graphs where classical algorithms (like PC) become impractical. Nevertheless, it remains sensitive to variable naming and heavily dependent on the LLM' internal knowledge. Finally, two recent and complementary contributions extend the role of LLMs in causal discovery. Liu et al. [61] propose *COAT*, a framework that leverages LLMs to extract candidate causal factors directly from unstructured data, annotate raw inputs accordingly, and iteratively refine these factors using feedback from causal discovery. This strategy addresses the upstream challenge of variable construction and demonstrates strong results in different domains, including healthcare. Constantinou et al. [22] explore a hybrid strategy in which GPT-4 does not directly construct a full causal graph, but instead provides causal hints based solely on variable labels. These hints are incorporated as constraints into classical ML algorithms, improving alignment with expert graphs while highlighting the trade-off between guidance and model flexibility.

These works broaden the scope of LLMs in causal reasoning, showing that they LLMs can act both as mediators between unstructured data and causal representations, and as external guides that enhance traditional

discovery methods. They therefore suggest that LLMs may be used as *virtual domain experts*, provided that the models they help generate are validated through statistical methods.

2.5 Related work

Several recent studies have demonstrated that causal inference methods can be effectively applied to observational clinical data. For instance, Charpignon et al. emulate a randomized controlled trial using large-scale electronic health records (EHRs) to estimate the causal effect of metformin on dementia, employing propensity weighting and competing-risk survival models [17]. This approach represents a significant methodological advance over traditional observational studies, combining the rigor of experimental design with the scale and heterogeneity of real-world data. Although the authors do not explicitly construct or learn a causal graph, their analysis is grounded in a well-defined causal model under the potential outcomes framework: they specify the treatment–outcome relationship of interest and estimate a treatment-assignment model (propensity score) to adjust for observed confounding.

Similarly, Xie et al. adopt a target-trial emulation framework to estimate the causal effect of different second-line anti-hyperglycaemic drug classes on kidney outcomes among individuals with type 2 diabetes [115]. They construct a treatment-assignment model using both pre-specified clinical variables and algorithmically selected features through a high-dimensional propensity score approach, and use the resulting weights in survival models to estimate causal effects. Although no explicit causal graph is defined, the study remains grounded in the potential outcomes framework and employs weighting to approximate random treatment assignment and achieve covariate balance across groups—thus implementing a genuine causal inference pipeline rather than a simple predictive association.

Moving beyond causal effect estimation, Shen et al. (2021) propose a causal structure discovery (CSD) method specifically tailored to EHR data, demonstrating its application in the context of type 2 diabetes mellitus [98]. Their algorithm modifies a score-based search method (Fast GES) by incorporating EHR-specific adaptations—distinguishing incident from

pre-existing conditions, compensating for noisy timestamps, and leveraging longitudinal ordering to orient causal edges. Although primarily methodological, this study underscores both the feasibility and the challenges of causal discovery in real-world clinical data. Particularly in the healthcare domain, the ability to infer causal structure from EHR can support clinicians in understanding disease mechanisms and designing interventions.

Building on the idea of integrating domain knowledge into causal discovery, Hasan et al. (2022) introduce KCRL, a knowledge-based causal discovery framework that combines prior domain information with data-driven causal structure learning [37]. Their approach leverages prior causal constraints to guide the search space of possible graphs, improving both identifiability and interpretability in high-dimensional observational data. Such hybrid methods stand at the intersection of causal discovery and causal inference: rather than merely estimating treatment effects, they aim to uncover the underlying causal mechanisms that structure the data, providing a richer basis for downstream causal reasoning and intervention design. Particularly in the healthcare domain, this approach can guide practitioners in making informed clinical decisions, as the inferred causal structures reveal potential mechanistic relationships between clinical variables and support evidence-based interventions.

Finally, Naik et al. (2024) apply a hybrid causal structure learning framework to healthcare data that integrates EHRs and genomic information for patients with non-small cell lung cancer [74]. Their method leverages LLMs as surrogates for expert knowledge to determine the directionality of causal edges, using zero-shot prompting to generate DAGs that are subsequently validated through Bayesian Dirichlet equivalent scoring. The resulting LLM-informed DAGs outperform traditional causal discovery algorithms such as NOTEARS and PC in both data fit and clinical interpretability, demonstrating how language-based reasoning and prior knowledge can enhance causal discovery and support informed clinical decision-making.

In line with these previous studies, this thesis focuses on inferring causal structures from observational clinical data in the context of occupational maritime medicine and diabetes. Sharing methodological aspects with the approaches of Shen [98], and Hasan [37], our work applies causal discovery algorithms to identify the underlying causal model without assuming prior

knowledge. Consistent with Naik [74], we exploit the capabilities of LLMs to assist in assessing and validating potential causal relations. The LLM-suggested causal structures are then statistically verified through falsification tests to ensure methodological soundness and empirical robustness. Furthermore, the inferred causal model is leveraged to simulate potential interventions and evaluate the effects of new clinical protocols before their application to real patients. This enables a safe, data-driven assessment of hypothetical scenarios, providing insights into the expected impact of specific treatments or preventive strategies. To assess the robustness of these simulated interventions, we perform refutation tests and compute the CATE, quantifying the heterogeneity of causal effects across different patient subgroups. Our framework is explained in the next chapter.

Chapter 3

Methodology

This chapter presents the methodological framework that represents the core contribution of this dissertation. The framework is designed with the aim of constructing a causal model directly from observational tabular data. Unlike traditional approaches, where domain experts play a central role in validating the plausibility of the inferred model, this work introduces one layer of domain validation. In particular, LLMs are used as an alternative to human experts, providing a way to assess the plausibility of the causal graph discovered in the absence of specific domain knowledge. By combining data-driven CD and CI, LLM-based domain validation and statistical validation, the framework deals with the dual challenge of inferring CMs from observational data and evaluating the effects of potential treatments before they are actually administered in the real world. Figure 3.1 shows the framework, and each step is discussed in the remainder of the chapter.

3.1 Data Collection and Preprocessing

In the data collection phase, it is crucial to define the type of study being conducted, as the research design directly affects the validity of causal conclusions. In this work, an observational study design was adopted, selected for its cost-effectiveness, and to overcome the ethical limits associated with RCTs [116]. The downside of this choice is that, in such studies, the data preprocessing phase becomes critical for the rest of the

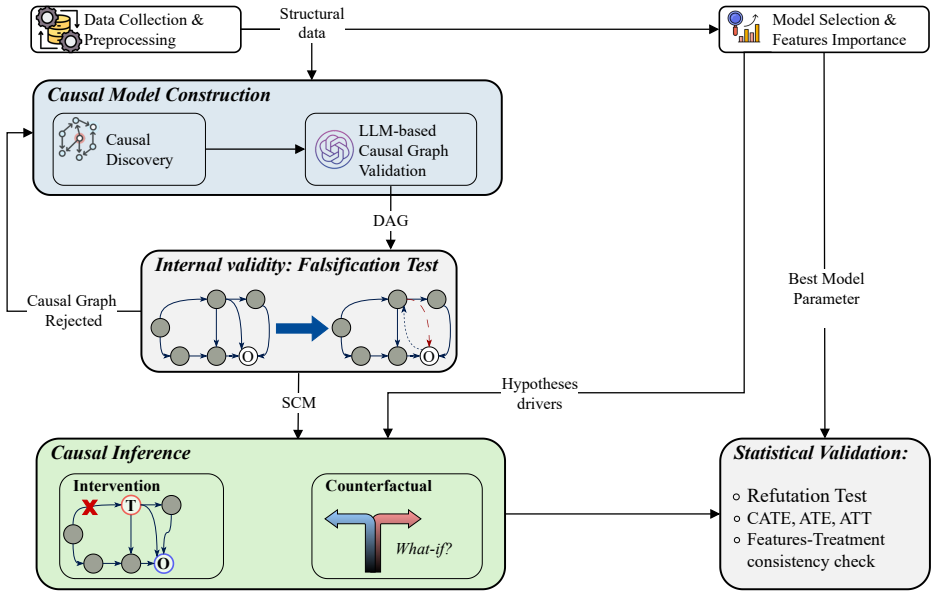


Figure 3.1. Diagram of the proposed methodology.

analysis.

The preprocessing stage is an essential step when working with observational data, as such datasets are often affected by issues such as missing values, outliers, or redundant features. The goal of this step is not only to clean the data but to ensure that the final dataset is suitable for causal analysis. In this work, preprocessing is conceived as an integral part of the methodological framework rather than a preliminary technical step. The specific procedures applied depend on the characteristics of the dataset, and their implementation will be described in the following chapters. In the remainder of this section, we briefly describe the main issues that we addressed. Missing values are a frequent problem in observational data and, if ignored, can introduce systematic bias or reduce statistical power; several reviews emphasize the need for adequate imputation strategies in such contexts [66, 62]. Nevertheless, the absence of a measurement may itself carry information [122], for example, a test might not have been performed on a patient precisely because it was deemed clinically unnecessary. This perspective highlights that missingness should not always be treated

as noise to be imputed, but also as a potential signal that can provide meaningful insights. Inconsistencies in recorded variables, such as heterogeneous coding schemes, unit mismatches, or logically implausible values, are also frequent in large-scale datasets and can lead to spurious associations if not harmonized [108]. Structural biases represent another critical issue: demographic imbalances, selection bias, or the overrepresentation of certain subpopulations may distort the estimation of causal effects if left uncorrected, a problem widely recognized in both epidemiology and machine learning [62]. Finally, semantic or syntactic errors are often introduced during the data collection process, especially when integrating heterogeneous sources; these include mislabeling, ambiguous variable definitions, or formatting mistakes.

3.2 Model Selection and Features Importance

In this work, model selection and the evaluation of Features Importance (FI) are intermediate steps to validate and support subsequent analyses. Model selection was carried out with the dual aim of ensuring predictive accuracy and interpretability, using representative algorithms from distinct methodological families, namely tree-based models, linear probabilistic models, and margin-based classifiers. FI was evaluated through a combination of model specific and model agnostic approaches. In three-based models, FI was derived from impurity reduction, assigning higher scores to features that, when used for node splits, consistently decrease heterogeneity in the child nodes [64]. In Logistic Regression (LR), standardized coefficients were used, with larger absolute values interpreted as stronger associations with the outcome [41]. For linear Support Vector Machines (SVMs), importance was inferred from the magnitude of the weight vector defining the separating hyperplane, again after standardizing the predictors [36]. In the case of non-linear kernels, where direct coefficient interpretation is not feasible, a permutation procedure was adopted: predictive performance was measured before and after randomly permuting the values of each feature, with the resulting accuracy drop serving as an indicator of its relevance [29]. This combined strategy ensured comparability of importance rankings across models while mitigating the limitations inherent in each individual method. This combined strategy ensured com-

parability of FI across models while mitigating the limitations inherent in each individual method. FI is therefore interpreted as an indicative measure of association and predictive contribution, rather than as causal evidence, given its sensitivity to scaling, feature cardinality, and correlations among predictors.

3.3 Causal Model Construction

The input of this stage is the cleaned dataset obtained from the Data collection and preprocessed stage. This dataset provides the empirical ground for applying CD algorithms, which infer potential causal relationships directly from the data. The resulting structure is formalized as a DAG, which represents the CM. The model is then evaluated through model validation with an LLM and a falsification test.

3.3.1 Causal Discovery

CD can be addressed through different algorithms, as outlined in Chapter 2, each providing distinct strategies to infer causal structures. In this work, we focus on two methods: the **LiNGAM**[99], and **PC** algorithm [103]. LiNGAM assumes that a set of observed variables $x = (x_1, ..., x_m)^T$ is generated through a linear, recursive process. Each variable is modeled as $x_i = \sum_{k(j)<K(i)} b_{ij}\, x_j + e_i + c_i$ where b_{ij} are causal coefficients, c_i are constants and e_i are noise terms. The causal order, given by $k(i)$, is assumed to be acyclic, meaning that no feedback loops or cycles are allowed. The distinctive feature of LiNGAM is that the noises e_i are assumed to be independent and non-Gaussian. This assumption allows the identification of not only dependencies among variables but also their causal direction. Through non-Gaussianity, it is possible to discriminate between alternative causal structures allowing to recognize the causal directions $(X \to Y \neq Y \to X)$ [100]. LiNGAM can therefore reconstruct a complete DAG from observational data. In this work we employ DirectLiNGAM [101], an extension of LiNGAM, that maintains the same assumptions. DirectLiNGAM introduces a deterministic algorithm for causal discovery, avoiding the use of Independent Component Analysis. The insight is to order variables directly according to their causal position by iteratively

identifying those without any predecessors (the exogenous variables). Its key property is that an exogenous variable should be independent of the residuals obtained when the other variables are regressed on it. The algorithm evaluates each candidate in turn by regressing the remaining variables on it and assessing the independence between the candidate and the residuals. If the candidate is truly exogenous, the independence condition will hold. Once such a variable is identified, it is removed from the system, and the procedure is repeated on the reduced set of variables. Through this iterative process, DirectLiNGAM constructs a causal ordering of all variables, after which the causal coefficients can be consistently estimated using standard linear regression. Compared to the original LiNGAM, DirectLiNGAM provides a deterministic solution that does not rely on random initialization, and is often more stable and accurate when working with finite samples.

The PC algorithm is a constraint-based approach. The algorithm begins with a fully connected undirected graph, where an edge exists between every pair of variables. This setup reflects the absence of initial assumptions in their relationships. The algorithm progressively removes edges by testing conditional independencies of the form $X \perp\!\!\!\perp Y \mid Z$ where X and Y are two variables and Z is a conditioning set. If the null hypothesis is not rejected, the edge between X and Y is removed. When the skeleton of the graph has been identified, orientation rules are applied to assign direction to the edges, ensuring that no directed cycles are formed, according to the DAG nature. The final output is a Completed Partially Directed Acyclic Graph (CPDAG), which represents a Markov equivalence class of DAGs consistent with the data. This means that the PC algorithm does not always return a unique causal direction for every relationship, but only those that are uniquely determined by the independence structure. The PC algorithm is particularly effective in detecting the overall structure of causal relations, provided that the sample size is sufficient to ensure reliable independence tests. Like LiNGAM, the PC algorithm relies on the assumption of no unobserved confounders. If latent variables exist and induce spurious dependencies, conditional independence tests become unreliable, and the resulting graph may be misleading.

3.3.2 Validation of the DAG with LLM

The input to this stage is a DAG derived from a CD algorithm. Prior to conducting CI —whether through the specification of treatments or the exploration of counterfactual scenarios— it is necessary to establish the validity of the inferred causal model. This validation requires both domain-expert assessment and statistical evaluation. The former is particularly critical, as the inferred relationships — especially the causal directions-may fail to capture the true underlying mechanisms among the variables. These errors can be derived from biases inherent in the data, as well as from latent confounders or other unobserved factors that compromise the identifiability of the causal structure. As reported in Chapter 2, several authors have examined the use of LLMs as *virtual domain expert*. In this work, we adopt **GPT-4** as a domain expert, in order to perform the domain-level validation of the causal graph. To this end, we designed a generic structured prompt to guide GPT-4 in the validation of causal graphs. The prompt was designed as a general-purpose validation framework for causal graphs expressed in DOT format. Its structure is modular, with each component targeting a specific aspect of validation, namely structural soundness of the DAG and plausibility with respect to domain knowledge. While the prompt is domain-agnostic in design, in this study, the framework is applied to two different datasets: one in the field of traditional medicine and one in the field of Occupational Health. The prompt is structured in the following steps:

1. **Causal order validation**: The model extracts a topological order from the DOT graph. This order is then compared against domain knowledge (e.g. the age cannot be an effect). The goal of this step is to verify whether variables are arranged in a plausible causal sequence [110]

2. **Local edge validation**: For each directed edge $X \to Y$, the LLM evaluates whether the effect is:

    - *Direct*: biologically or theoretically plausible
    - *Mediated*: transmitted through one or more intermediate variables
    - *Implausible*: unsupported by domain knowledge

3. **Cycle check:** Since a causal graph must be a DAG, the LLM verifies acyclicity. Any detected cycles must be flagged, and corrections suggested. The purpose is to enforce the DAG property and avoid logical inconsistencies [80, 103].

4. **Adjustment set validation**: The model identifies the minimal valid adjustment set (backdoor adjustment set) for estimating a causal effect of interest.

Before executing the prompt, it is essential to reduce the risk of hallucinations or unfounded statements by the LLM [120, 106]. For this reason, the prompt explicitly constrains the model to rely only on the information available in the DOT graph and in the provided domain description. If information is missing or unclear, the model is instructed to reply with "I don't know" instead of guessing. In addition, every answer must be accompanied by a confidence score and, where possible, presented in a structured format such as bullet points or tables. These constraints were introduced to enhance the reliability, reproducibility, and interpretability of the validation performed by GPT-4. After executing the prompt, the output is a causal graph that has been validated both with respect to domain expertise and to the structural constraints of a DAG. On this validated graph, statistical tests will be performed to further validate its correctness. Figure 3.2 presents the prompt in its domain-free version.

You are an expert in causal inference and graphical models. I will provide you with a causal graph in DOT format.
Important instructions:

- Only use the information from the DOT graph and the provided domain description.
- If something is unclear or missing, explicitly say *"I don't know"* instead of guessing.
- For every answer, give a confidence score (0–100%).
- Use structured outputs (bullet points or tables where possible).

Your task is to audit the DOT graph step by step, following this checklist:

1. **Global order validation**
    - Extract a topological order of the variables from the DOT graph.
    - Is this order plausible given the domain [insert domain description]?

- Explicitly report any variable that appears isolated (i.e., has no incoming edges) and discuss whether this is plausible in the given domain.
- Point out up to 3 variable pairs that may be incorrectly ordered.
- Provide an overall confidence score.

2. **Local edge validation**
   - For each edge $X \rightarrow Y$ in the DOT graph:
   - Is the effect plausibly direct, mediated, or implausible?
   - Answer in the format: [direct / mediated / implausible] + short explanation + confidence.
3. **Cycle check**
   - Identify if any cycles exist in the DOT graph.
   - If cycles exist, are they plausible in a causal model?
   - Suggest which edge(s) should be removed or reversed to restore acyclicity, with justification.
4. **Adjustment set validation**
   - Suppose we want to estimate the causal effect of T (treatment) on Y (outcome).
   - What is the minimal valid adjustment set to block all backdoor paths?
   - Is this set plausible in the domain [insert domain description]?
   - If not, suggest corrections (variables to remove/add) with reasoning.

Now here is the DOT graph to audit.

Figure 3.2. Prompt for Causal Domain Validation of a DAG

3.3.3 Internal Validity: Falsification test

To evaluate the plausibility of the obtained causal assumptions, we employ the falsification test implemented in *DoWhy* [97], through the function *falsify_graph*. While the theoretical foundations of falsifiability have been introduced in 2, this section focuses on the practical implementation of the falsification procedure based on the permutation-based approach proposed by Euling et al. [26]. In *DoWhy*, this test is built on the observation that a candidate DAG G implies a set of testable conditional independencies. Two central families of such constraints are:

- Local Markov Condition (LMC): For each node X_i, the graph implies that $X_i \perp\!\!\!\perp ND_G(i) \backslash Pa_G(i) | Pa_G(i)$, meaning that X_i is independent of all its non-descendants that are not its parents, conditional on its parents. A violation of this condition indicates that the graph fails to capture a dependency present in the data [80, 88].

- Parental Triples (TPa): A parental triple is of the form $(i, j|Z)$ where $j \in ND_G(i) \backslash Pa_G(i)$ and $Z = Pa_G(i)$ corresponding to the conditional independence $X_i \perp\!\!\!\perp X_j | Pa_G(i)$. These triples represent local, testable implications derived directly from the parental structure of the DAG. Violations of the TPa constraints indicate that the graph fails to reproduce the conditional independencies implied by its local Markov property, thus reducing its plausibility as a representation of the data-generating process [26].

The falsification test evaluates whether the conditional independencies implied by the LMC and TPa constraints are consistent with the observed data. For each constraint, conditional independence tests are performed and the total number of violations, denoted $v(G)$ is computed. To assess whether the observed number of violations is acceptable or indicates poor model fit, the procedure follows the permutation-based framework introduced in [26]. The key idea is to compare $v(G)$ with a reference distribution obtained by randomly permuting the node labels of G, thereby preserving its topology but destroying its causal semantics. For each permuted graph $\sigma(G)$, the same independence tests are performed, and the corresponding number of violations $v(\sigma(G))$ is computed. The empirical *p-value* is then calculated as

$$p(G) = \frac{1}{T} \sum_{t}^{T} 1v(\sigma_t(G)) \leq v(G) \tag{3.1}$$

where T is the number of random permutations and $1\{\cdot\}$ is the indicator function. This value measures the proportion of permuted graphs that fit the data at least as well as the candidate graph. A small $p(G)$ suggests that the candidate DAG reproduces the observed conditional independencies significantly better than expected under random structure and is therefore *not falsified*. Two complementary *p-value* are computed in practice: p_{LMC} that quantifies the relative fit of the candidate graph with respect to the

Condition	Interpretation
$p_{TPa} \leq \alpha$	The graph is falsifiable: its conditional independencies can be empirically tested.
$p_{TPa} \leq \alpha$ and $p_{LMC} > \alpha$	The graph is falsified: independencies inconsistent with the data.
$p_{TPa} \leq \alpha$ and $p_{LMC} \leq \alpha$	The graph is not falsified: no contradictions with the data.
$p_{TPa} > \alpha$	The graph is not falsifiable: its structure cannot be empirically tested.

Table 3.2. Decision criteria for the permutation-based falsification test, based on [26].

Local Markov constraints, and p_{TPa} that assesses the *falsifiability* of the model by evaluating how many node permutations yield Markov-equivalent graphs. Following [26], the decision criteria are reported in Table 3.2. The distinction between falsifiability and falsification is essential: a graph can only be rejected if it is empirically testable. The procedure, implemented in the *falsify_graph()* function of DoWhy [97], performs the conditional independence tests implied by the LMC and TPa constraints, generates the permutation baseline. Figure 3.3 illustrates the overall testing workflow.

3.4 Causal Inference

After validating the causal graph, this phase focuses on causal inference, which represents the analytical core of the proposed methodology. The objective of this phase is to estimate the causal effect of a treatment variable on an outcome, moving beyond statistical associations and allowing for the evaluation of hypothetical changes in the treatment. The analysis follows the formalism of SCMs and combines symbolic causal reasoning, implemented through DoWhy, with simulation-based inference provided by Graphical Causal Models (GCM)[97, 11]. By fitting the data to the model, the causal DAG is instantiated as a corresponding SCM, where each structural equation captures the quantitative dependencies implied

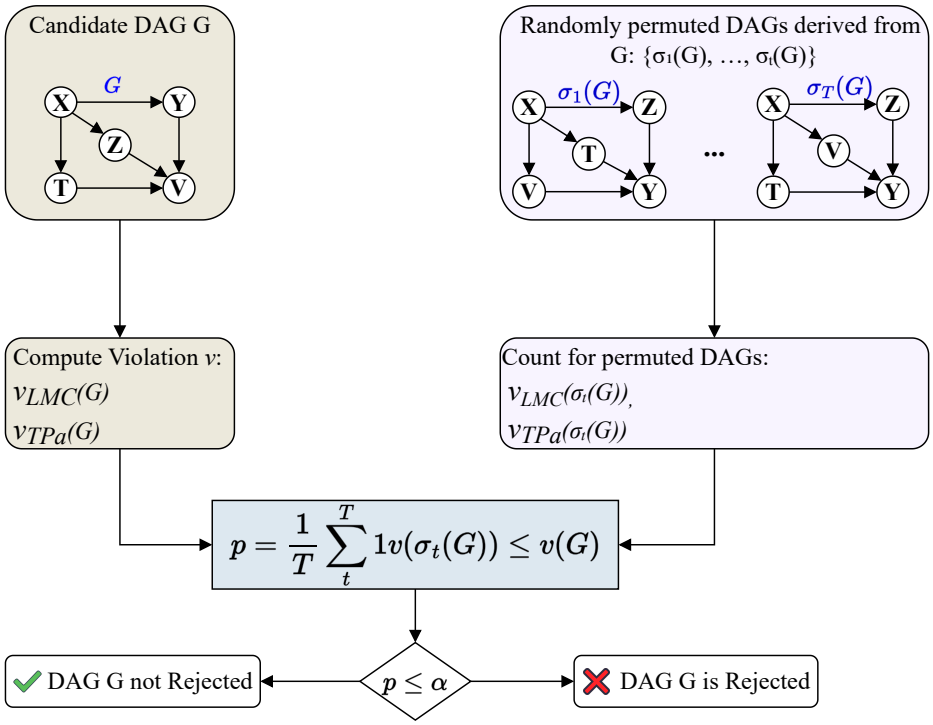


Figure 3.3. Schematic workflow of the permutation-based falsification test implemented in **DoWhy**

by the graph. The SCM's structural equations and exogenous noise terms induce a joint probability distribution over the variables and, under interventions, define the interventional distributions $P(\cdot)$ and $P(\cdot|do(\cdot))$. This probabilistic foundation provides a model on which to perform causal inference; accordingly, the first step concerns the identification of the causal estimand. This step establishes whether the causal effect of interest, typically expressed as $P(Y|do(X = x))$, can be expressed in terms of observable quantities derived from the data. The causal inference process begins with the identification of the causal effect. At this stage, the goal is to determine whether the causal relationship between treatment and outcome, typically expressed as $P(Y|do(X = x))$, can be derived from the joint distribution of the observed variables encoded in the causal graph. DoWhy performs

this step by interpreting the structure of the graph and applying the logical rules of causal reasoning formalized in the do-calculus. It examines the causal pathways linking X and Y, identifies potential confounders, and verifies whether the effect can be expressed using observable quantities—thus defining a *causal estimand*. The estimand represents a formal expression of the causal effect under the model's assumptions, specifying the exact conditioning or adjustment required to isolate the effect of the treatment from spurious dependencies. Through this procedure, DoWhy translates the theoretical causal question into a mathematically estimable form, ensuring that subsequent inference is both identifiable and consistent with the causal structure previously validated. Once the estimand is defined, the next step is its empirical estimation based on the available data. DoWhy ensures that the estimation procedure directly corresponds to the identified causal estimand and provides confidence intervals to quantify the associated uncertainty. Beyond the estimation phase, causal inference extends to *interventional* and *counterfactual* analyses.

Interventional analysis evaluates how the distribution of the outcome would change if the treatment variable were externally fixed to a specific value. In practical terms, this addresses questions such as "What would the expected outcome Y be if the treatment X were set to a specific value x?" In this setting, all incoming causal paths into X are conceptually removed, and the resulting effect is propagated throughout the graph to assess the consequences on Y. This framework provides a principled way to evaluate *what-if* scenarios under controlled hypothetical manipulations of the treatment. n the context of interventional analysis, the propagation of the intervention throughout the causal graph relies on the specification of causal mechanisms for each variable. Within the GCM framework, each node in the graph is associated with a structural function that defines how its value depends on its parent variables, typically expressed as $V_i = f_i(Pa(V_i), U_i)$, where $Pa(V_i)$ are the parent nodes and U_i denotes independent noise. These mechanisms can take different functional forms—such as **Additive Noise Models** (ANMs) or **Post-Nonlinear Models** (PNL)—depending on the type of data and the assumed causal relationships. When an intervention is applied, the mechanism governing X is replaced by the constant assignment $X = x$, effectively removing the influence of its parents. The structural equations of the remaining nodes then propagate this change

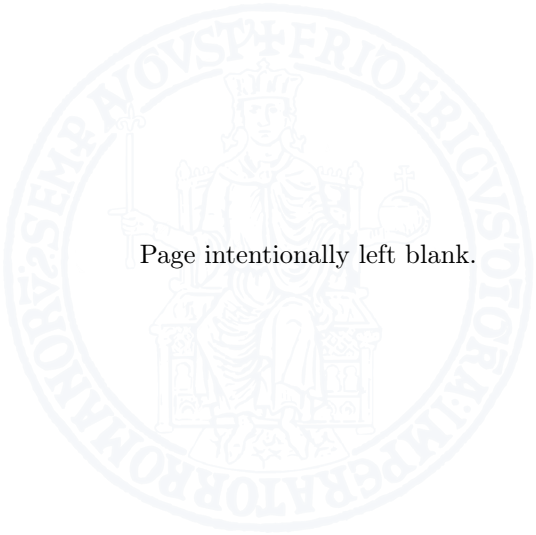
across the graph through their respective causal mechanisms, producing a new distribution over the outcome variables. This generative formulation allows the interventional distribution to be estimated coherently with the underlying causal structure.
Counterfactual analysis enables reasoning about hypothetical alternatives for individual observations. It answers questions of the form "What would the outcome have been for this specific observation if the treatment had been different?" Following the abduction–action–prediction paradigm, counterfactual reasoning proceeds by first inferring the latent factors consistent with the observed data (abduction), then applying the intervention (action), and finally generating the corresponding hypothetical outcomes (prediction) [80]. This approach allows for the estimation of individual treatment effects and provides a finer-grained interpretation of causal relationships, complementing the broader inference obtained from population-level estimates.
Through the integration of symbolic identification, empirical estimation, and structural simulation, the causal inference framework adopted in this study provides a coherent and interpretable approach for quantifying causal effects. The combination of DoWhy and GCM enables analytical rigor and empirical flexibility, ensuring that causal effects are both theoretically grounded in the causal graph and statistically supported through consistent estimation and confidence intervals.

3.5 Statistical Validation

After estimating the causal effects, the next phase focuses on validating the robustness and reliability of the obtained results. In causal inference, statistical validation is essential to ensure that the estimated effects are not artifacts of model specification, data noise, or unobserved confounding, but rather reflect genuine causal relationships consistent with the underlying structure of the model. To address this, the analysis first employs the refutation framework provided by DoWhy. This framework performs a series of robustness tests that intentionally modify certain aspects of the model or the data to verify whether the estimated causal effect remains stable. In this work, four refutation strategies were applied. The Dummy Outcome Refuter replaces the true outcome with a random variable to confirm that

no effect is detected when the causal relationship is spurious. The Random Common Cause Refuter adds a simulated confounder to test the sensitivity of the estimate to irrelevant variables. The Placebo Treatment Refuter permutes the treatment variable to check that the effect disappears when the causal link is broken. Finally, the Data Subset Refuter re-estimates the effect on random subsets of the data to assess its stability under sampling variation. Consistent results across these tests indicate that the estimated causal effect is robust and not dependent on specific data configurations or modeling assumptions. In addition to robustness checks, statistical validation also involved the evaluation of heterogeneous treatment effects. This analysis was conducted using the Causal Forest estimator implemented in the EconML library [25]. Causal forests are ensemble-based models designed to estimate the CATE across different subgroups of the population [7, 113]. By combining sample splitting with effect-oriented tree growth, the model provides individual-level causal effect estimates while controlling for overfitting. This approach allows the exploration of how the treatment effect varies among individuals or covariate profiles, offering a more detailed view of the underlying causal mechanism. Validation tests were then applied to assess the reliability and interpretability of the estimated CATEs, focusing on metrics such as calibration, slope consistency, and uplift-based performance. These tests evaluate whether the model correctly captures genuine heterogeneity rather than random variation in the data. Together, the refutation tests from DoWhy and the validation of CATE estimates using the Causal Forest provide complementary evidence that the inferred causal effects are both statistically sound and robust to changes in model assumptions. The CATE estimation was performed using the hyperparameters of the best-performing model identified during the feature importance phase. Furthermore, the CATE results were cross-checked against the feature importance ranking to assess the internal consistency of the findings. Variables showing strong predictive relevance in the feature importance analysis also exhibited meaningful heterogeneity in the CATE estimates, supporting the coherence of the causal interpretation. In summary, the methodological framework presented in this work integrates causal discovery, causal inference, and statistical validation into a coherent analytical pipeline. After identifying and validating the causal graph, the inference phase quantifies both population-level and

individual-level effects through interventional and counterfactual analyses. The subsequent robustness and validation procedures ensure that the estimated causal relationships are consistent with the data and stable under variations in model assumptions. This integrated approach combines theoretical rigor with empirical reliability, providing a solid foundation for the interpretation of causal mechanisms and their practical implications in the context of the study.



Page intentionally left blank.

Chapter 4

Causal Reasoning for Diabetes Prediction and Management

Gestational diabetes mellitus (GDM) is a metabolic disorder characterized by glucose intolerance that develops during pregnancy in women without a prior history of diabetes [6]. This condition typically emerges during the second or third trimester as a result of placental hormones that impair insulin sensitivity [96], [49]. Although GDM often resolves after childbirth, it significantly increases the risk of future gestational diabetes and preeclampsia in the mother, as well as preterm birth and respiratory distress in the newborn. Moreover, it represents a major risk factor for the later development of type 2 diabetes in both mother and child [2], [86]. Early diagnosis and proper management—through dietary interventions, physical activity, and, when necessary, pharmacological treatment—are crucial to ensuring a healthy pregnancy and reducing adverse outcomes. A deeper understanding of the causal mechanisms underlying diabetes and its associated risk factors is essential to improve both predictive modeling and clinical decision-making. In this chapter, the analysis framework introduced in Chapter 3, and showed in Figure 3.1, is applied to the study of gestational diabetes, using the **Pima Indians Diabetes** dataset, a well-known clinical dataset in diabetes research [77]. This dataset provides a meaningful context for exploring the relationships among clinical and physiological variables and assessing their influence on disease onset and progression. The objectives of this study are formulated through the

following Research Question (RQ):

- **RQ1**: What are the causal relationships between the selected features and diabetes diagnosis?
- **RQ2**: What is the effect of specific interventions on the Glucose variable?
- **RQ3**: What are the counterfactual outcomes with respect to diabetes diagnosis?

By applying the causal framework to gestational diabetes, this chapter seeks to demonstrate how data-driven causal reasoning can enhance clinical insight and support more informed decision-making in healthcare contexts.

4.1 Related work

This section reviews the main studies related to diabetes prediction and analysis, with particular attention to those using the Pima Indians Diabetes Dataset, a widely used resource in diabetes research. The PIMA dataset was widely used in the literature because it represents a real-world scenario of a high-risk population, and it includes physiological and lifestyle factors crucial to diabetes onset. Utilizing the PIMA dataset facilitates the exploration of feature patterns and relationships, as well as evaluating the performance of various machine learning algorithms for predicting diabetes. Numerous studies have relied on this dataset to investigate predictive modeling techniques and identify key factors associated with diabetes onset.

In his work, Mousa [72] conducts a comparative analysis of three machine learning algorithms—Long Short-Term Memory (LSTM), Random Forest (RF), and Convolutional Neural Network (CNN)—for detecting diabetes using the PIMA dataset. In Lakhwani's research [56], a two-layer Artificial Neural Network (ANN) is applied to forecast diabetes onset, achieving a classification rate of 55%. According to Lakhwani, the classification strength is influenced by the weights assigned in the first layer's edges. Furthermore, this approach suggests that ANN can be realistically applied for predicting diabetes, although further measurable testing is needed. The authors in [78] propose a machine learning framework for

diabetes prediction and diagnosis, employing techniques such as Spearman correlation and polynomial regression to manage missing data and select features. They observe a prediction accuracy of 97.931% using their developed twice-growth deep neural network (2GDNN).
In [15], authors use the PIMA dataset to train three machine learning models —Naive Bayes classifier, Random Forest classifier, and J48 decision tree classifier— to create an e-diagnosis system for diabetes detection.
While prior research focused on improving diabetes prediction through machine learning, Liang et al. [59] utilize BNs for a deeper evaluation of diabetes risk. BNs facilitate the examination of variable relationships and enable risk inference by predicting disease occurrences from known variables. Liang's study emphasizes the utility of BNs in improving diabetes risk assessment.
This work aims to advance this through applying causal inference to more thoroughly comprehend the cause-effect relationships in diabetes.

4.2 Dataset and Preprocessing

The **Pima Indians Diabetes** dataset originates from the *National Institutes of Diabetes and Digestive and Kidney Disease*. It focuses on a group of Pima Indian women. The dataset has 9 features and comprises 768 entries, each representing a woman between the ages of 21 and 81. The features are:

- **Pregnancies** (P): Number of times pregnant.
- **Glucose** (G): Plasma glucose concentration over 2 hours in an oral glucose tolerance test.
- **Blood Pressure** (BP): Diastolic blood pressure (mm Hg).
- **Skin Thickness** (ST): Triceps skinfold Thickness (mm).
- **Insulin** (I): 2-Hour serum insulin (mu U/ml).
- **Body Mass Indicator (BMI)**.
- **Diabetes Pedigree Function** (DPF): Diabetes Pedigree function, a genetic score of diabetes.

- **Age**: Age in years.
- **Outcome**: Binary classification indicating the presence or absence of diabetes.

The dataset has no missing entries, yet certain records exhibit biologically improbable values for features like *BMI*, *Glucose*, *Blood Pressure*, *Insulin*, and *Skin Thickness*. Rows containing at least one zero in any of these features are removed. This step is the sole pre-processing action performed. While managing outliers with domain expertise could enhance the model's performance, we opted not to handle outliers to prevent unintentional removal of significant values and avoid bias. This resulted in a final dataset whose demographic and clinical distributions are detailed in Table 4.1. The age of the participants is mostly between 20 and 29 years (63.27%), followed by the age groups 30-39 years (18.88%), 40-49 years (10.71%), and over 50 years (7.14%). BMI values are categorized into six groups, with a higher prevalence in the Obese I category (30.61%), followed by Obese II (22.19%), Overweight (21.68%), and Obese III (14.03%). A smaller percentage of individuals fall into the Healthweight category (11.22%) and Underweight (0.26%). The number of pregnancies is divided into four quantile-based intervals: most participants have between 0 to 1 pregnancy (38.01%), followed by intervals of 2 to 5 pregnancies (23.72%), 5 to 17 pregnancies (21.94%), and 1 to 2 pregnancies (16.33%). Clinical parameters are categorized into four quartile intervals each: glucose levels (ranging from 56 to 198 mg/dL), blood pressure (from 24 to 110 mmHg), skin thickness (from 7 to 63 mm), insulin (from 14 to 846 µU/mL), and diabetes pedigree function (from 0.084 to 2.42). Each interval exhibits a nearly uniform distribution around 25%. The outcome variable indicates the presence or absence of diabetes, with 66.84% of participants classified as healthy and 33.16% classified as diabetic.

Table 4.1. Features values distributions

Feature	Range	Frequency	Percentage %
Age	[20, 29]	248	63.27
	[30, 39]	74	18.88
	[40, 49]	42	10.71
	50+	25	7.14
BMI	Underweight	1	0.26
	Healthweight	44	11.22
	Overweight	85	21.68
	Obese I	120	30.61
	Obese II	87	22.19
	Obese III	55	14.03
Pregnancies	[0, 1.0]	149	38.01
	(1.0, 2.0]	64	16.33
	(2.0, 5.0]	93	23.74
	(5.0, 17.0]	86	21.94
Glucose	(55.9, 99.0]	102	26.02
	(99.0, 119.0]	95	24.23
	(119.0, 143.0]	98	25.00
	(143.0, 198.0]	97	24.74
Blood Pressure	(23.9, 62.0]	102	26.02
	(62.0, 70.0]	99	25.26
	(70.0, 78.0]	96	24.49
	(78.0, 110]	95	24.23
Skin Thickness	(6.9, 21.0]	102	26.02
	(21.0, 29.0]	100	25.51
	(29.0, 37.0]	102	26.02
	(37.0, 63.0]	88	22.45
Insulin	(13.9, 76.75]	98	25
	(76.75, 125.5]	98	25
	(125.5, 190.0]	99	25.26
	(190.0, 846.0]	97	24.74
Diabetes Pedigree Functions	(0.084, 0.27]	98	25
	(0.27, 0.45]	98	25
	(0.45, 0.687]	100	25.51
	(0.687, 2.42]	96	24.49
Outcome	Healthy / 0	262	66.84
	Diabetic / 1	130	33.16

4.3 Machine Learning and Features Importance

The main goal of this step is to identify the most relevant features associated with diabetes prediction. In predictive modeling, identifying the effects of the features on the target variable, is a crucial phase for improving both model performance and interpretability. Performing a correlation analysis provides an initial overview of the relationships between the features and the outcome. However, correlation captures only pairwise associations and may not reflect the actual predictive contribution of each feature within a model. To improve understanding of the impact of features, after identifying the best model, feature importance techniques were used to evaluate how much each variable influences the model's predictions. This combined assessment of correlation and importance supports the design of the subsequent causal analysis, including treatment selection and estimation of treatment effects.

To this end, five supervised learning algorithms were evaluated: LR, Random Forest (RF), Multi-Layer Perceptron (MLP), Gradient Boost (GB), and SVM. The dataset was split into training and test sets using a 70/30 ratio, with stratification on the outcome variable to preserve its distribution across both subsets. Prior to model training, all features were normalized to the range [-1, 1] to ensure comparability across variables and to facilitate convergence, particularly for models sensitive to feature scaling. For each algorithm, optimal hyperparameters were identified through a grid search procedure with cross-validation, aimed at maximizing predictive performance and ensuring fair model comparison. Model performance was assessed using accuracy, F1 score, and the area under the ROC curve (AUC). The performance comparison among the evaluated models is presented in Table 4.4. As reported in Table 4.4, the RF achieved the best overall performance, with an accuracy of 0.81, an F1 score of 0.70, and an AUC of 0.83. The GB model also performed well, reaching comparable results (accuracy: 0.79, F1 score: 0.69, AUC: 0.80). Both Logistic Regression and SVM obtained similar but slightly lower scores, while the MLP recorded the lowest performance across all metrics.

Based on these results, the Random Forest was selected as the most effective model and was subsequently used for the feature importance analysis. This analysis revealed that Glucose accounted for the highest proportion

of the model's predictive power (27.2%), followed by Insulin (15.5%), Age (14.9%), and BMI (11.6%). Other variables such as DPF and Skin Thickness contributed to a lesser extent, while Pregnancies and Blood Pressure showed minimal influence on the model's output. Accordingly with the framework 3 the parameters of the optimal model are employed in the CATE estimation phase to evaluate the treatment effects. In this case, the treatment variable selected is not necessarily the most statistically significant one, but rather the most significant among those that can be assessed and modified without the assistance of a domain expert. The chosen variable is *BMI*, which, in any case, is one of the key parameters used in the management of diabetes.

Model	Accuracy	F1 Score	ROC
Gradient Boosting	0.79	0.69	0.80
Logistic Regression	0.77	0.64	0.81
Multi Layer Perceptron	0.72	0.58	0.79
Random Forest	**0.81**	**0.70**	**0.83**
Support Vector Machine	0.77	0.64	0.78

Table 4.4. Performance comparison of the evaluated models

4.4 RQ1: Causal Structure of Diabetes Mechanisms

To derive the causal model, we employed the DirectLiNGAM algorithm [99], which assumes a linear and non-Gaussian data-generating process. Limited domain knowledge was incorporated by specifying *Age* as an exogenous variable. The resulting causal graph is presented in Fig. 4.1. While DirectLinGAM captures the pairwise relations and quantify them with structural equations, it did not manage correctly identify the direction for the *Outcome* variable. The graph indicates that *Outcome* is influenced solely by *Age* and serves as a causal determinant for the other variables. This wrong identification can derive from the fact that some causal

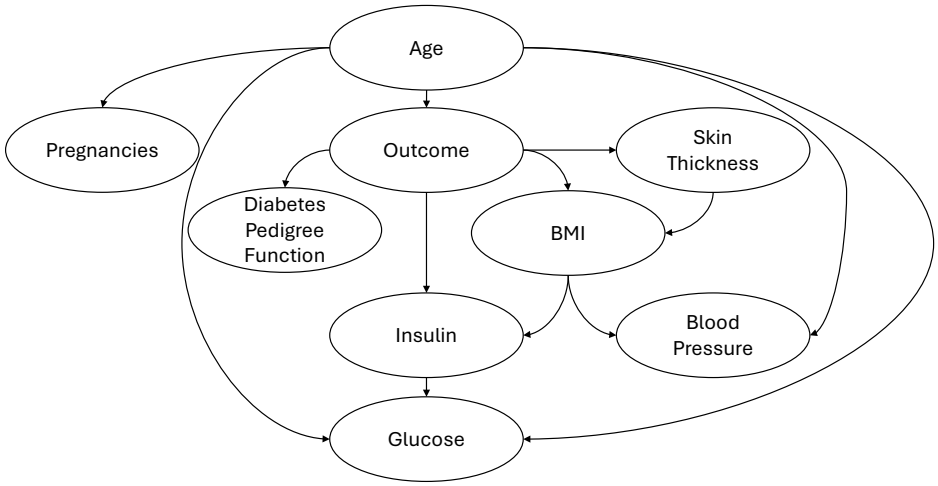


Figure 4.1. Causal graph inferred from PIMA dataset using the LiNGAM algorithm.

direction can appear misleading when interpreted temporally [103]. For instance, the edge Outcome → Insulin might seem inverted if read chronologically: one would typically expect insulin dysfunction to contribute to risk of developing diabetes and not the reverse. However, DirectLiNGAM infers causal directions based on statistical independence patterns, without incorporating biological or temporal knowledge. Futhermore, upon closer inspection of the variables, we noted that the relationship between *Outcome* and *Glucose* (that is not linear due to threshold [21]) are not captured from DirectLiNGAM. Consequently, we decided to derive the causal graph using the PC algorithm, which is better accommodates nonlinear dependencies [119]. Fig 4.2 shows the causal graph obtained with the PC algorithm.

This highlights the importance of combining data-driven causal discovery with domain expertise. To incorporate domain knowledge, and following the framework introduced in Chapter 3, we used a prompt-based validation to assess the causal graph obtained with the PC algorithm. To support the validation process, the causal graph derived from the PC algorithm was evaluated using the prompt showed in Figure 4.3. The prompt specifies a practical application of the causal audit procedure. It requires analyz-

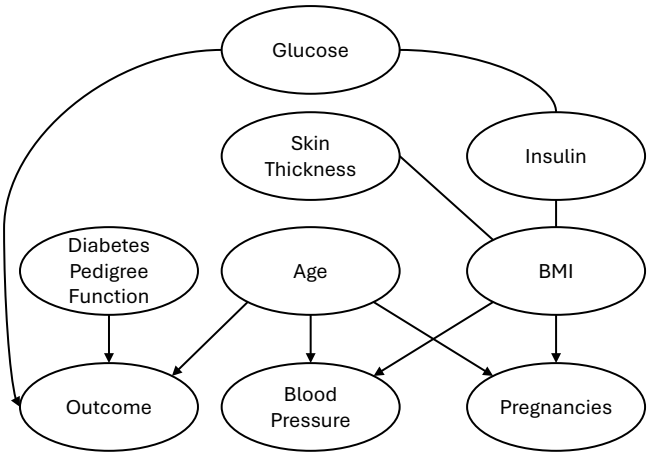


Figure 4.2. Causal graph inferred from the dataset using the PC algorithm.

ing the graph by checking variable order, causal link plausibility, potential cycles, and the validity of the adjustment set. Each evaluation must be justified, include a confidence score, and be presented in a structured format.

You are an expert in causal inference and graphical models. I will provide you with a causal graph in DOT format.
You are an expert in causal inference and medical epidemiology.
Important instructions:

- Only use the information from the DOT graph and the provided domain description.
- If something is unclear or missing, explicitly say "I don't know" instead of guessing.
- For every answer, give a confidence score (0–100%).
- Use structured outputs (bullet points or tables where possible).

Domain description: This graph comes from the medical domain, specifically diabetes research. The variables follow the PIMA Indians Diabetes dataset, including:

- Pregnancies (number of times pregnant)
- Glucose (plasma glucose concentration)
- BloodPressure (diastolic blood pressure)

- SkinThickness (triceps skinfold thickness)
- Insulin (2hour serum insulin)
- BMI (body mass index)
- DiabetesPedigreeFunction (genetic risk score for diabetes)
- Age (in years)
- Outcome (diagnosis of type 2 diabetes: 0 = no, 1 = yes).

Use medical knowledge about diabetes risk factors and epidemiology when judging causal plausibility. Do not invent new variables; only use the ones provided.
Your task is to audit the DOT graph step by step following this checklist:

1. **Global order validation**
   - Extract a topological order of the variables from the DOT graph.
   - Is this order plausible given the domain?
   - Explicitly report any variable that appears isolated (i.e., has no incoming edges) and discuss whether this is plausible in the given domain.
   - Point out up to 3 variable pairs that may be incorrectly ordered.
   - Provide an overall confidence score.
2. **Local edge validation**
   - For each edge X $\rightarrow$ Y in the DOT graph:
   - Is the effect plausibly direct, more likely mediated, or implausible?
   - Answer in the format: [direct / mediated / implausible] + short explanation + confidence.
3. **Cycle check**
   - Identify if any cycles exist in the DOT graph.
   - If cycles exist, are they plausible in a causal model?
   - Suggest which edge(s) should be removed or reversed to restore acyclicity, with justification.
4. **Adjustment set validation**
   - Suppose we want to estimate the causal effect of BMI (treatment) on Outcome (diabetes diagnosis).
   - What is the minimal valid adjustment set to block all backdoor paths?
   - Is this set plausible in the medical domain?
   - If not, suggest corrections (variables to remove/add) with reasoning.

Figure 4.3. Prompt for Causal Domain Validation of PIMA DAG

Below is the response produced by the GPT-4 model after applying the described prompt, which outlines step by step the causal graph evaluation according to consistency, plausibility, and inferential accuracy criteria. The answer is reported in the Figure 4.4. GPT-4 applies the causal audit procedure to the graph, assessing its structure and consistency with causal inference principles. After reconstructing the topological order of variables, the model identifies biologically inconsistent relationships, such as the reversed connection between *Glucose* and *Insulin* and the questionable direction *Pregnancies* → *BMI*. In its local edge evaluation, it distinguishes between direct, mediated, and implausible effects, providing physiological reasoning and confidence levels. During the cycle check, GPT-4 detects bidirectional links incompatible with an acyclic causal model, particularly between Age and *Pregnancies*, *Pregnancies* and *BMI*, and *Glucose* and *Insulin*. Consequently, it suggests removing the direction *BMI* ← *Pregnancies* (keeping *BMI* → *Pregnancies* if appropriate), reversing *Glucose* → Insulin to Insulin → *Glucose*, and maintaining Age → *Pregnancies* as a one-way relationship. To estimate the causal effect of *BMI* on *Outcome*, the model recommends adjusting for Age, and possibly for *Pregnancies*, while avoiding adjustment for mediators such as *Glucose* or *Insulin*. These structural adjustments are illustrated in Figure 4.5.

Alright, let's carefully walk through the audit of this DOT graph step by step. I'll stick to the checklist you gave me and only use what's in the graph plus the diabetes domain knowledge.

1. **Global order validation:** Extracted topological order (from edges):
   - Age → Pregnancies, BloodPressure, Outcome
   - Pregnancies → Insulin, BMI
   - BMI → BloodPressure, SkinThickness, Insulin, Outcome
   - Glucose → Insulin, Outcome
   - Insulin → Outcome
   - DiabetesPedigreeFunction → Outcome

   So a plausible order is: Age → Pregnancies → BMI → (BloodPressure, SkinThickness, Insulin) → Outcome with Glucose entering before Insulin/Outcome, and DPF only pointing to Outcome.

2. **Plausibility check:**
   - Age before Pregnancies: plausible (older women more likely to have had more pregnancies).
   - Pregnancies → BMI: questionable (pregnancies can affect BMI long-term, but usually BMI is an antecedent risk factor). BMI → BloodPressure: plausible (obesity increases hypertension risk).
   - Glucose → Insulin: biologically this is reversed in physiology (insulin regulates glucose, not the other way).

   Overall ordering is partly plausible but has some inverted edges. 3 variable pairs that may be incorrectly ordered:
   - Glucose → Insulin (should be Insulin → Glucose; insulin deficiency leads to high glucose).
   - Pregnancies → BMI (BMI is mostly determined by lifestyle/genes; pregnancies can contribute, but BMI usually precedes diabetes risk, not the other way around).
   - Age → Outcome directly is okay, but Age → BloodPressure → Outcome is probably a more natural mediator path. Confidence: 70% (some plausible, but key physiological relations look inverted).
3. **Local edge validation:**
   For each edge:
   - Pregnancies → Insulin: [implausible] physiology doesn't support this, pregnancies don't directly change 2h serum insulin in general; effect mediated via BMI and age. (Confidence 90
   - Pregnancies → BMI: [mediated] pregnancies can contribute to long-term BMI changes, but lifestyle/genetics are stronger drivers. (Confidence 70
   - Age → Pregnancies: [direct] obvious chronological constraint. (Confidence 100
   - Glucose → Insulin: [implausible] direction is reversed; insulin regulates glucose. (Confidence 95
   - Glucose → Outcome: [direct] hyperglycemia directly diagnostic for diabetes. (Confidence 95
   - BMI → BloodPressure: [direct] obesity causes hypertension. (Confidence 95
   - Age → BloodPressure: [direct] aging increases BP. (Confidence 95
   - BMI → SkinThickness: [direct] triceps skinfold is a proxy for adiposity, so BMI → SkinThickness makes sense. (Confidence 90

- BMI → Insulin: [mediated] obesity drives insulin resistance, raising insulin levels. (Confidence 90
- Insulin → Outcome: [mediated] hyperinsulinemia is a compensatory marker; diabetes outcome mainly through glucose dysregulation. (Confidence 80
- BMI → Outcome: [direct] obesity is a causal risk factor. (Confidence 95
- DiabetesPedigreeFunction → Outcome: [direct] genetic risk. (Confidence 90
- Age → Outcome: [direct] age increases diabetes risk. (Confidence 90

4. **Cycle check:**
   The graph has cycles because edges are marked 'dir=both' (bidirectional). Examples:
   - Pregnancies ↔ BMI (0 → 5 with dir=both).
   - Age ↔ Pregnancies.
   - Glucose ↔ Insulin.

   Cycles are not plausible in a causal DAG (by definition). Edges to remove or reverse:
   Remove BMI ← Pregnancies (keep BMI → Pregnancies if you argue obesity increases likelihood of gestational diabetes/pregnancies). Reverse Glucose → Insulin to Insulin → Glucose. Keep Age → Pregnancies one-way (not symmetric).
5. **Adjustment set validation**
   Treatment: BMI
   Outcome: Diabetes diagnosis
   Backdoor paths:
   - BMI ← Age → Outcome
   - BMI ← Pregnancies ← Age → Outcome
   - BMI ← Age → BloodPressure → Outcome (if BP were linked to Outcome, though not explicit here)
   - BMI ← (unblocked confounders like DPF don't affect BMI, only Outcome).

   Minimal valid adjustment set (to block backdoors):
   - Age
   - Pregnancies (possibly redundant if Age already controls Pregnancies)

So a minimal set = Age. Adding Pregnancies would be extra cautious.

Plausibility: High — Age is the main confounder linking BMI and diabetes. Correction suggestion: ensure Glucose and Insulin are not adjusted for, since they are mediators on the causal path.

Summary of issues:

- Biggest misspecification: Glucose ↔ Insulin direction (should be Insulin → Glucose).
- Cycles from 'dir=both' need pruning (Age ↔ Pregnancies, Pregnancies ↔ BMI).
- Pregnancies → BMI is questionable; likely reversed or mediated by Age.
- Minimal adjustment set for BMI–Outcome = Age (possibly + Pregnancies).

Do you want me to redraw a corrected DOT graph with these fixes so you have a clean DAG to work with?

Figure 4.4. Prompt for Causal Domain Validation of PIMA DAG

To further support the validity of the causal relationships identified through intervention queries and counterfactual analysis, we conducted a falsification test using the *falsify_graph* function of the DoWhy library [26]. This procedure evaluates if the proposed DAG is informative and consistent with the observed data, by testing its LMCs against permuted alternatives. Although the test does not provide definitive proof of correctness, it offers evidence on the robustness of the causal structure by highlighting potential violations and suggesting model refinements. Furthermore, to evaluate the impact of domain-expert refinement on the causal structure we compared the falsification reports of the pre-expert (Figure 4.2) and post-expert (Figure **??**) DAGs. Both models were found to be informative ($0/20$ permutations in the same Markov equivalence class, $p = 0.00$) and consistently outperformed permuted DAGs, confirming that their structure is not attributable to random noise. The pre-expert DAG violated 14 out of 52 LMCs, while the post-expert DAG violated 12 out of 40. This reduction, suggests that the refinement introduced by domain knowledge improved the agreement between the causal structure and the logical constraints derived from the data. These findings highlight a trade-off between biological plausibility and statistical coherence: the post-expert model combine medical knowledge more effectively and aligns more closely

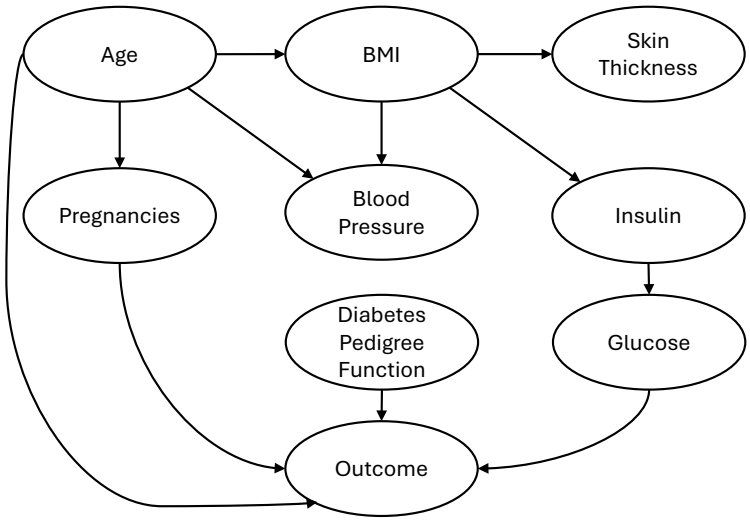


Figure 4.5. Causal graph inferred using the PC algorithm and validated by ChatGPT as domain expert.

with the data constraints. Before proceeding to causal inference, the DAG was fitted to the data to obtain a corresponding SCM, ensuring that the model's structural equations reflect the observed statistical dependencies.

4.5 Individuals and Subgroups

After identifying and validating the causal structure during the causal discovery phase, the analysis proceeds by applying this model at both the individual and subgroup levels to address RQ2 and RQ3 through causal inference techniques. To do this, we randomly extracted four individuals from the dataset, one for all overweight's class, and four subgroups. The characteristics of the individuals are in Table 4.6. The subgroups (by age range) are:

- **S1**: Age = [20-29], diabetics = 48, healthy = 190
- **S2**: Age = [30-39], diabetics = 35, healthy = 43
- **S3**: Age = [40-49], diabetics = 26, healthy = 20

- **S4**: Age $\geq$ 50, diabetics = 21, healthy = 9

Id	P	G	BP	ST	I	BMI	DPF	Age	Outcome
I1	1	109	56	21	135	25.2	0.833	23	0
I2	2	124	68	28	205	32.9	0.875	30	1
I3	6	134	70	23	130	35.4	0.542	29	1
I4	13	153	88	37	140	40.6	1.174	39	0

Table 4.6. Individuals Characteristics

4.6 RQ2: Causal Effects of Lifestyle Changes on Glucose

An intervention aims to evaluate the impact of actively manipulating a (set of) variables on the effect. For example we can simulate the effect of the healthy lifestyle (diet and exercise). We simulate the impact of the intervention, for individual and subgroups of interest, meaning that we assess the impact on a single individual, with her own characteristics, or the *average* effect on a subgroup of individuals. To evaluate the effects of the interventions, we will rely on the variable **glucose** rather than **outcome** – since there is no cure for diabetes, what can be done is to manage the disease by controlling **blood glucose levels** [91]. For both, individuals and subgroups, in accordance with common diabetes prevention strategies, we set, for illustrative purpose, the BMI as normal weight as a target, to be achieved through diet and exercise. Thus, the intervention is to set **BMI=22**. The effect of the intervention is computed using the `DoWhy` library [11]. Considering the probabilistic nature of causal models, we ran 5,000 repetitions for each individual and subgroup. Table 4.8 reports the assessed effect of the intervention, in terms of variation of **glucose post-intervention** (PIGlucose) with respect to the initial **initial glucose** (IGlucose) values. As we expected the model responds well to the intervention. The majority of the median differences are positive, suggesting that the intervention generally leads to a reduction in glucose levels.

However, for *S1* and *S3*, the average difference is negative. In the case of *S1*, which contains the largest number of observations, the change in glucose levels before and after the intervention is minimal – confidence interval (CI) is close to contain 0. For *S3*, the effect of other factors (e.g., blood pressure), exacerbated by the small size of the group, may be contributing to the observed glucose trends.

Id	Characteristic	IGlucose-PIGlucose	95% CI
I1	BMI = 25.2	-1.43	-62.80, 34.21
I2	BMI = 32.9	13.58	-49.12, 48.76
I3	BMI = 35.9	23.28	-40.06, 58.70
I4	BMI = 40.6	41.93	-19.52, 77.00
S1	Age = [20-29]	-15.62	-18.89, -12.30
S2	Age = [30-39]	0.30	-5.04, 6.12
S3	Age = [40-49]	10.50	2.61, 17.89
S4	Age $\geq$ 50	42.25	33.16, 50.74

Table 4.8. Result of intervention BMI=22 on Individuals and Subgroups on 5000 repetition

4.7 RQ3: Counterfactual Outcomes for Diabetes Diagnosis

With counterfactuals, we retrospectively examine what would have happened if the patient had different characteristics. For illustrative purpose, assume we want to answer the question *"Would the patient have become ill with diabetes even if she maintained a healthy lifestyle?"* For individual-level analysis, we selected individuals *I2* and *I3* from Table 4.6. At the subgroup level, we considered only the **diabetic** patients from subgroups *S1–S4*. In both cases, we investigated how outcomes would change under modified conditions, specifically altering the values of **BMI** or **Pregnancies**. The results are in Table 4.10. "#Healthy" denotes the number of patients (median over repetitions) for which the outcome would change from *diabetic* to *healthy* under the set BMI Counterfactual queries at the

individual level indicate that the specific individuals *I2* and *I3* would still have developed diabetes, hence suggesting other co-causes besides BMI and . While, at subgroup level, in all subgroups *S1–S4* there are patients who would have changed to healthy under the counterfactual conditions.[1]

Id	#Diabetic	Query	#Healthy	95% CI
I2	1	BMI = 22	0	0, 0
I3	1	BMI = 22	0	0, 0
S1	48	BMI = 22	5	5, 5
S2	35	BMI = 22	13	13, 13
S3	26	BMI = 22	11	11, 11
S4	21	BMI = 22	7	7, 7

Table 4.10. Counterfactual on Individuals and Subgroups on 5000 repet.

4.8 Statistical Validation

Following the counterfactual and interventional analyses, a validation phase was conducted to assess the robustness of the estimated causal effect of BMI on diabetes incidence. Specifically, four refutation strategies from the DoWhy library were applied to evaluate the stability of the results and the model's sensitivity to potential biases or data variations. For all the tests, the estimated effect of BMI on diabetes was 0.0117, meaning that an increase of one point in BMI leads to about a 1.2% higher chance of developing diabetes. This value aligns with previous findings in the literature, which report a 1.5% increase in diabetes risk per BMI unit [35]. The results of the refutation tests, summarized in Table 4.12, showed minimal variation in the estimated effect, with no statistically significant deviations. These findings support the robustness and internal validity of the causal inference.

After confirming the robustness and internal validity of the estimated causal effect of *BMI* on diabetes incidence through multiple refutation

[1] Repetitions here gave consistently the same result (CI equals the median).

Refutation Test	Estimated Effect	New Effect	p-value	Description
Dummy Outcome Refuter	0.0	-0.0003	0.94	Replaces the actual outcome with a random variable to check for spurious associations.
Random Common Cause Refuter	0.0117	0.0117	0.98	Adds a random confounder to assess sensitivity to irrelevant variables.
Placebo Treatment Refuter	0.0117	-0.0004	0.78	Permutes the treatment to break any real causal relationship and verify the effect disappears.
Data Subset Refuter	0.0117	0.0121	0.2998	Re-estimates the effect on random subsets to evaluate stability under sampling variation.

Table 4.12. Summary of Refutation Tests for BMI Effect on Diabetes

tests, the analysis proceeds to estimate the treatment effect using advanced causal inference models capable of capturing individual-level heterogeneity in responses. To evaluate the effect of **BMI**, as treatment, on the **Outcome**, we used the Causal Forest model. Causal forests are honest tree ensembles tailored to learn heterogeneous treatment effects. By combining sample-splitting and effect-focused splits, they average local effects across trees to deliver individual-level estimates with valid inference [113]. In this analysis, we employed the Causal Forest estimator as provided by EconML library [25]. The CATEs distribution shows marked dispersion: most individual effects are negative, with a few close to zero, and long negative tail,

indicating that the impact of the treatment varies across the subjects. The *ATE*, obtained by averaging the distribution of CATE, has a value equal to -0.143. This negative average treatment effect indicates that moving individuals from the non-normal weight to the normal weight group is associated, on average, with a reduction of about 14 percentage points in the probability of diabetes. The direct ATE estimate provided by the model coincides with the mean of the CATE distribution, and the 95% confidence interval $[-0.286; 0.001]$ confirms a robust negative tendency, though statistical significance is borderline. Validation tests[2] confirmed that the estimated CATEs captured genuine heterogeneity with results reported in Tab.4.13.

Table 4.13. Validation of CATE estimates with DRTester (BMI → Outcome)

Metric	Estimate	SE	p-value	Interpretation
BLP slope	3.857	1.132	0.001	Positive and significant, CATEs capture real heterogeneity.
Calibration R^2	0.231	–	–	23% of variance in group effects explained; moderate calibration.
QINI (uplift)	0.058	0.017	< 0.001	Significant uplift, model ranks beneficiaries better than random.
AUTOC	0.122	0.045	0.003	Significant ranking ability across subgroups.

[2] metrics details https://www.pywhy.org/EconML/_autosummary/econml.validate.DRTester.html

4.9 Threats to validity

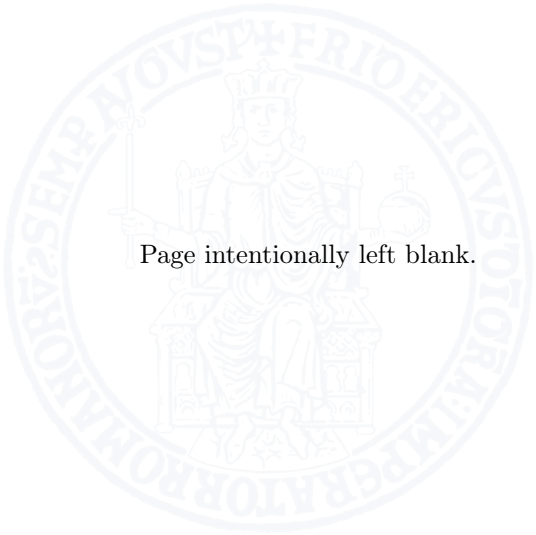
Despite the encouraging results and the methodological rigor adopted in this study, several factors may threaten the validity of the findings.

Internal validity may be limited by unobserved confounders not included in the PIMA dataset, which could bias causal relations and treatment effect estimates. Although LLMs supported the biological validation of the causal graph, their feedback depends on the scope and quality of training data, introducing potential epistemic bias.

External validity is limited by the demographic composition of the PIMA dataset, which mainly includes adult women of Pima Indian heritage. As a result, the identified causal relationships and intervention effects may not fully generalize to other populations. Broader and more diverse datasets are needed to validate and extend these findings.

Construct validity may be affected by the use of proxy variables such as BMI, glucose, and insulin to represent complex physiological processes. These simplifications, though necessary for interpretability, may not fully capture the multi-factorial nature of diabetes, introducing measurement uncertainty into the causal estimates.

Overall, although refutation and falsification analyses strengthen confidence in the causal graph and effect estimates, these limitations underline the importance of interpreting the results with caution. Future research should aim to mitigate these threats by incorporating richer clinical information, validating models across independent datasets, and exploring longitudinal designs to better capture temporal dynamics in disease progression.



Page intentionally left blank.

Chapter 5

A Real-World Case Study: Analysis of Occupational Health Data in the Maritime Domain

This chapter presents an observational study in the maritime domain. Starting from health protocols, it explores, in cooperation with an Italian company in charge of medical examinations (CML Vesuvio), which aspects of workers' health and habits are most strongly associated with *fitness/unfitness* for work in the maritime field. The analysis leverages machine learning techniques to identify the most common patterns. Moreover, data gathered from two international maritime companies are made publicly available [1]. The method defined harnesses both machine learning and causal inference. The former aims to infer the most relevant associations in terms of features and ranges of values more likely impacting the Fitness for work (FFW). The latter aims to infer the relevant *causal relations* between such identified factors, if any, to obtain a causal model to be queried for planning potential improvement actions. In the following, *positive fitness* refers to the worker's complete suitability for a specific task as assessed by a medical examination, while *negative fitness* includes

[1] https://github.com/Patrizia40/MaritimeDataset.git

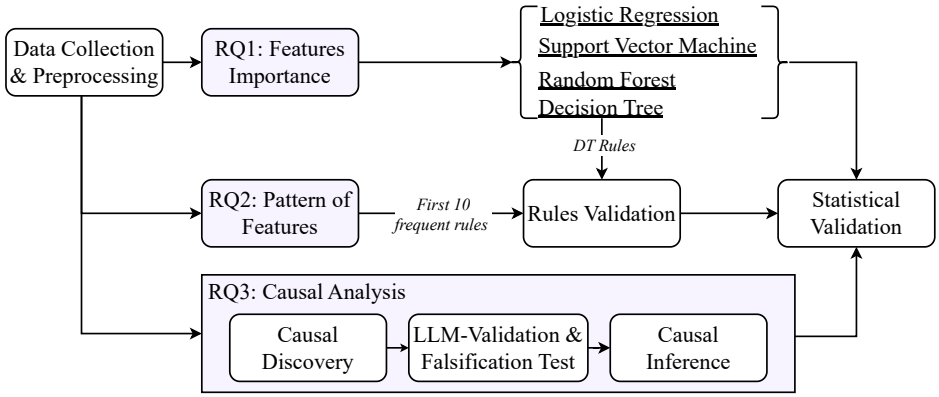


Figure 5.1. The Extended Workflow

all other cases, such as conditional fitness (temporary or permanent) and fitness with restrictions. The above objectives are formulated as RQ:

- **RQ1** What are the most important features involved in the FFW prediction?
- **RQ2** What are the patterns of the features' values more likely associated to the FFW?
- **RQ3** What are the causal relations between the features and the FFW?

To answer the research questions, the following workflow was designed and articulated into four main steps, each discussed in the next sections.
This chapter applies the methodological framework described in Chapter 3, introducing a minor extension to incorporate an additional analysis step. This addition was designed to deepen the investigation and address specific aspects not covered in the general workflow. The extended workflow, is show in figure 5.1. For clarity, the components shared with the general framework have been simplified in the figure, while the additional step is explicitly highlighted.

5.1 The Role of Health Surveillance in Occupational Health

Workplace health and safety has a significant impact - in terms of human lives and well-being, as well as in terms of financial cost - on employees, businesses, and society as a whole. In 2016, there were 2.4 million cases in the EU of non-fatal injuries and illnesses [27]. The estimated cost of treating diseases and injuries related to the workplace in Europe is 3.3% of Gross Domestic Product, according to a study conducted between 2015-2016 by the European Agency for Safety and Health at Work (EU-OSHA) [28]. For workers, companies, governments, and society as a whole, these occupational illnesses, injuries, and fatalities have severe financial consequences caused from the loss of qualified workers, excessive medical expenses and insurance premiums, absenteeism, presenteeism. Prevention through health surveillance plays an important role to reduce the exposure to risk factors [105],[58] and also to reduce the costs for all the players involved. *To properly implement prevention actions, understanding what factors are potentially linked to work-impeding health problems and what is the expected impact of their change is essential.* Health surveillance aims to identify risks, evaluate prevention, and link workplace exposures to health outcomes [1]. Through health protocols' definition, it aims to promote safety through preventive measures, to detect early disease, and to monitor illness patterns in worker populations. Health surveillance allows unusual patterns of illness to be observed among specific populations of workers. These patterns need to be validated through strong statistical evidence linking the marker with disease incidence which can require significant effort and time. For instance, among the most famous risks identified through health surveillance, there is the case of exposure to asbestos, for which the process of recognizing the exposure as a health risk was complex and long-lasting [8].

5.2 Related Work

While ML is routinely applied for health analytics, there are relatively few studies about occupational health, mainly because of scarce datasets and of scarce quality. Among these, in [16], the authors developed ML models for

FFW prediction, concluding that logistic regression is the best models for the binary Fit/Unfit classification. They conducted an observational study to investigate the relation between the work-entry medical examination and FFW. The authors found that this relation does not have a great predictive power, and found that the best features for prediction were glucose, age, hemoglobin, BMI and cholesterol. In addition they also noted, and we confirm, a lack of similar studies. Fadel *et al.* [**?**] also highlight a lack of data caused by the risks of the sharing of medical data, proposing the use of a synthetic dataset. In our study, we use and share a real-world occupational health dataset in the maritime domain. We also apply ML models for a correlation analysis; this is then enriched with pattern mining and especially with causal analysis to enable simulation and planning of improvement actions.

5.2.1 RCTs and Observational Studies

The Randomized Controlled Trial (RCT) is the gold standard for validating patterns, as it avoids many issues of observational studies and requires minimal assumptions or prior information [23]. In the occupational health literature, Serra et al. conduct a systematic review of 39 studies, deriving five main criteria including worker's risk and worker's capacity in relation to the workplace conditions [93]. RCTs are, however, expensive, and not always applicable, e.g., for ethical or privacy issues [92]. The wide spread of Electronic Health Records, and Electronic Medical Records [5] provides an opportunity to enhance the surveillance efforts of public health agencies [10]. The creation of numerous large observational healthcare databases globally is providing enhanced data assets to support observational studies [69], which stand now as a valuable alternative to RCTs, significantly reducing cost and time, as well as ethical concerns. Observational studies allows the investigation of prevalence, incidence, associations, causes, and outcomes, and are strongly bolstered by machine learning techniques, which effectively identify complex patterns in large datasets [9], [94].
In the field of occupational health, very few observational studies have been conducted. In [16], the authors develop a machine learning models for FFW prediction with the conclusion that logistic regression is the best models for binary classification Fit/Unfit. Charapaqui-Miranda et al. [16] conducted an observational study to investigate the relation between work-

entry assessment and the FFW. They found that this relation does not have a great predictive power, and identified that the highest predictive value were glucose, age, hemoglobin, BMI and cholesterol.

5.3 Data Collection and Preprocessing

This step involves cleaning raw data from spurious text (e.g., grammar and syntactical issues), anonymizing records, and transforming them into a format suitable for machine learning algorithms.

5.3.1 Data Collection

We extracted and collected the employees' medical examination data from two maritime companies in Italy for the period 2021-2022. The health protocol [2], [3] of these companies prescribes the following medical tests:

- **Laboratory Tests**: azotemia, blood exam, glycemia, GPT/ALT, GOT/AST, g-GT
- **Diagnostic tests**: audiometric test, spirometry, ECG
- **Alcohol Tests**: alcohol test, AUDIT
- **Drugs Tests**: amphetamine, buprenorphine, cocaine, MDMA, methadone, cannabinoid, methamphetamine, opiate
- **Physical Examination**: Spinal evaluation
- **Vaccination**: anti-tetanus

Furthermore, the health protocol establish which of these tests are mandatory for each job category, as reported in Tables A.1 and A.2.

[2] Dlgs 81/08 "Protection of health and safety in the workplace": https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:2008;81 art306

[3] Dlgs 271/99 "Adjustment of regulations on the safety and health of maritime workers on board national fishing merchant vessels": https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:1999-07-27;271

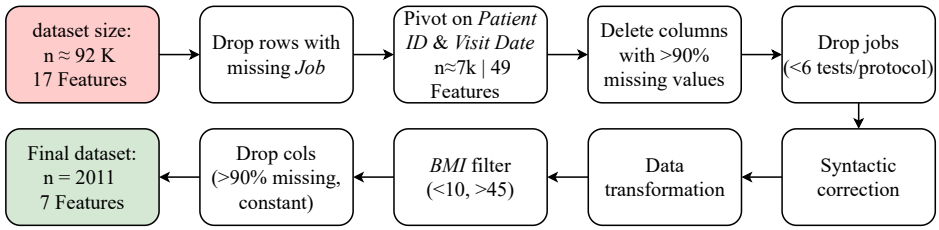


Figure 5.2. Data Preprocessing Workflow

5.3.2 Data Preprocessing

Data have been pre-processed according to the workflow in Figure 5.2. First, we computed the age of each patient at the date of the examination and populated the corresponding field. Then, we deleted all the rows that have a missing value in the **Job** column; for the column **Outcome** of the medical test, we replaced the missing values with "Not inserted" (the test was made, but the outcome is not inserted in the DB). Data are then anonymized by creating a unique ID for each patient from the tuple {Name, Surname, Date of birth}. An excerpt of the raw data we collected is in Table 5.1. Then, a pivoting operation has been applied to transform the

Table 5.1. Pre-Transformation Data Information.

Patient ID	Visit date	Visit type	OUTCOME
1	26/07/2021	Periodic visit	Fit for specific job
1	26/07/2021	Blood count	Within reference range
1	26/07/2021	Spirometry	Normal
1	26/07/2021	Glycemia	Within reference range
2	14/07/2021	Periodic visit	Fit for specific job
2	14/07/2021	Spirometry	Normal
2	14/07/2021	Spinal evaluation	
2	04/08/2022	Periodic visit	Fit for specific job
2	04/08/2022	Spirometry	Normal
2	04/08/2022	Spinal evaluation	

Table 5.3. Post-Transformation Data Information

ID	Visit date	Visit type	Blood count	Spirometry	Glycemia	Spinal evaluation	Outcome
1	26/07/21	Periodic visit	In reference range	normal	In reference range		Fit
2	14/07/21	Periodic visit		normal		Not inserted	Fit
2	04/08/22	Periodic visit		normal		Not inserted	Fit

data format in order to have one entry per patient, getting to a dataset with 7348 entries and 49 columns from 92k $\times$ 17 columns dataset. An extract of the pivoting operation results is report in Table 5.3.
After correcting for syntactic anomalies via *regex* (e.g., removing illegal characters, aligning the values to the expected field type) , a data transformation step is run to either fix the type heterogeneity in the information reported (e.g., mixing numerical and textual data in a field where numbers only numbers are expected) and to derive meaningful features from raw features contributing no information alone. These transformations led to the following new features:

- *Audiometry Class* contains the class of audiometric tracings in which the impairment was found (the audiometric tests were classified in conformity with the Merluzzi-Pira-Bosio method [13]).
- *Audiometry Score* contains the value, in percentage, of the damage (if a subject has a normal hearing condition, this value is 0).
- *Transaminase* is a synthetic feature based on GOT/AST and GTP/ALT values.
- *Drug Test* is a synthetic feature based on the values of individual test values for specific substances.
- *Job Category* is a synthetic feature that captures the similarity between jobs, grouping them into five categories:

- **Deck Officers (DO)**: Bosum, Deck Boy, Deck Cadet, Deck Officer, Ordinary Seaman
- **Professional Certification and Credentials (PCC)**: Able Deck Seaman, Able Seafarer Engine, Able Seaman
- **Engineering Positions (EP)**: Chief Engineer, 1st Engineer, 2nd Engineer, 3rd Engineer, Engineer Rating, Engineer Officer, Engineer Deck
- **Engineering Department (ED)**: Wiper, Engineer Cadet, Fitter, Plumber, Motorman, Electrician
- **Shipyard Jobs (SJ)**: Carpenter, Electrotechnical Officer, Trainee Electrician, Electrician Assistant

Tables 5.5,5.7 report about the distribution of some features, highlighting two problems:

- The majority of features are extremely unbalanced;
- The values have very low variability.

We thus discard all the features that do not bring a real contribution (e.g., all values being equal or missing, like g-GT).

Table 5.5. Values Distribution of Laboratory Values

	Azotemia	**Blood Count**	**g-GT**	**Glycemia**	**Transaminase**	**Drugs Test**
Negative	95.82 %	99.95 %	97.06 %	99.95 %	99.95 %	50.72 %
Positive	0 %	0 %	0 %	0 %	0 %	0 %
Missing	4.17 %	0.05 %	2.93 %	0.05 %	0.05 %	49.27 %

For the **BMI** features we consider a range for the overweight and underweight, we have excluded from the dataset all the workers which have BMI 45 (obesity threshold suggest by Yudin is 40 [118]) and a BMI $\leq$ 10 (the threshold for the underweight is 18.5). The last phase of preprocessing is the missing value handling. Authors of [65] show that missing values are typically divided into three categories: missing completely at

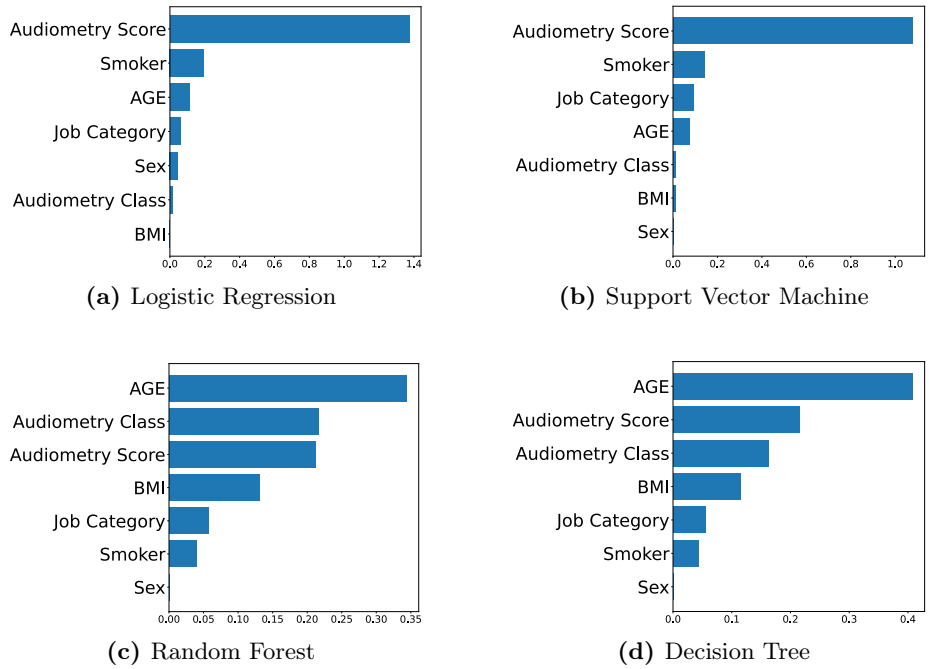
Table 5.7. Distribution of Instrumental Test Values

	Spirometry	**Audiometry**
Normal	97.96 %	84.23 %
Alteration	0 %	19.24 %
Missing	2.03 %	0 %

random (**MCAR**), missing at random (**MAR**), and missing not at random (**MNAR**). In our data, the majority of missing values are **MCAR**, the characteristics of this type of missing value, as reported in [42] and [62], is that causes of missingness are not related to any characteristics of the dataset. According to the protocol A.1-A.2, the left missing values in our dataset are only unnecessary tests for that specific job, so it is possible to delete the rows.
The result of this phase is a dataset with 2,011 entries and the following features:*Job, Job Category, Age, BMI, Sex, Smoker, Audiometry Score, Audiometry Class.*

5.4 RQ1: Data Analysis for Feature Importance Ranking

The goal in this phase is to identify, and rank by importance, features that most impact the FFW prediction. We split the dataset in training set (70%) and test set (30%). We adopt multiple widely-used ML models: LR, Decision Tree (DT), RF and SVM. The features importance scores, as identified by each algorithm, are reported in Figure 5.3.
LR and SVM: LR is widely used in medical research[12]. The most significant feature in this case (Fig. 5.3a) is judged to be **Audiometry Score**, followed by **Smoker** and **Age**. A similar result is obtained by SVM [47] (Fig. 5.3b). Smoking is one of the factors confirmed to be impactful, e.g., as also assessdd by Kumar *et al.*, who found a significant association between smoking and neural hearing loss of workers [55].
DT and RF: Slightly different results are obtained with DT [102] and RF [3]. RF hightlights, as most important features, the **Age** followed by



(a) Logistic Regression **(b)** Support Vector Machine

(c) Random Forest **(d)** Decision Tree

Figure 5.3. Features Importance

the **Audiometry Score** and the **Audiometry Class**. This condition is at the base of the process of the FFW judgement, since with aging the probability to have a negative FFW judgement increase for different causes like the major exposure to some dangerous factors to the health status (e.g., vibrations or noises). Furthermore, as known from the literature, there is a physiological degeneration of hearing that occurs with ageing [43].
All the algorithms confirm the same features, with audiometric-related features being always among the top-3, age always in the top-4, and smoker status being in the top-2 by two algorithms.
Job category analysis In the previous analysis, we grouped jobs by **Job Category**. Figure 3 shows, for each **Job Category**, the distribution of Fit or Unfit workers. For most jobs except Desk Officer, *fit* workers are more than *unfit*. We observe that most workers are fit, except for the DO. Among the DO, most workers are unfit, with the disparity between fit and

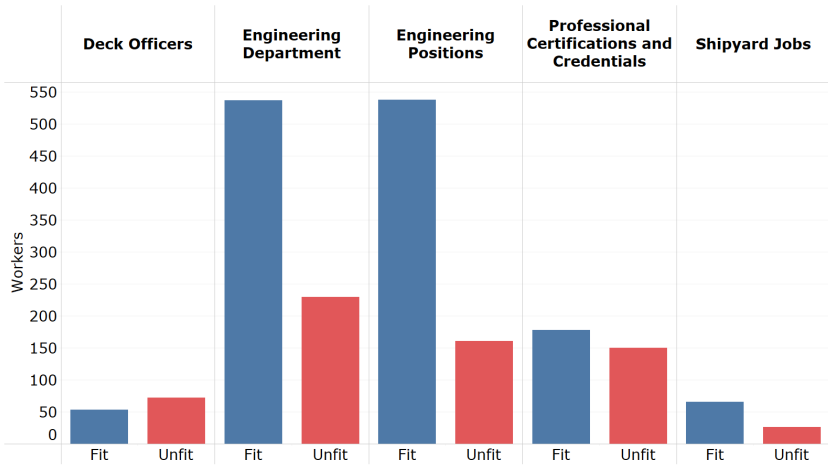


Figure 5.4. Distribution of Fit & Unfit vs Category

unfit workers being less than 25. In addition, to we have calculated the prevalence, respect of Unfit subject, for each category and listed along with the number of instances in the category [76]:

- **DO**: Prevalence = 0.576, number of instance = 125
- **ED**: Prevalence = 0.299, number of Instance = 767
- **EP**: Prevalence = 0.230, number of Instance = 699
- **PCC**: Prevalence = 0.457, number of Instance = 328
- **SJ**: Prevalence = 0.282, number of Instance = 92

We examined data split by each *Job Category* individually, to see if the important features differ depending on the job category. We repeated the previous analysis considering individual jobs. For the DT and RF, the results in terms of feature importance were similar to what we reported in Fig. 5.3 – we just obtained a swap, for several jobs categories, between *Audiometry class* and *Audiometry score*, which are two strongly correlated features. In the LR, except for the *Desk Officer* jobs having a swap between *BMI* and *AGE* and for *Shipyard* jobs having a swap between *Smoker* and *Job*, the important features are unchanged. In the SVM model, the

only significant change is observed in the *Professional Certifications and Credentials* category, for which *Smoker* becomes the third most important factor. Results are consistent with the established relationship between smoking status and neural hearing loss [55]. It is important to highlight that the prescription of examinations and subsequent determination of fitness depend on the specific task. Therefore, the changes identified may vary depending on the health protocol followed. Overall, the identified differences by job category are marginal, and the important features are roughly confirmed to be the same regardless the job category, with few exceptions.

5.5 RQ2: Data Analysis for Patterns Mining

In this Section, we aim to discriminate the conditions under which a subject is judged *fit* or *unfit*. The goal is to identify the features' patterns associated with a positive or negative qualification judgment. To this aim, we used the *top-k class* algorithm [31][4], whose aim is to discover the k most frequent class association rules. A class association rule between a set of antecedents X and consequents Y is a special type of association rule where Y contains a single item that must be chosen from a set allowed items, in our case the *fit* or *unfit* outcome. With $k = 10$, we get the rules reported in Table 5.9. In the previous Section 5.4, we have used, among others, a decision tree as a model. We have used the result obtained from the tree to double-check the validity of the top-k rules. The top-10 rules identified have all at least one rule in the decision tree with the same features and the same clauses. In the following, we report a set of rules of the decision tree matching the top-10 rules found by the Top-k class algorithm:

- AGE $\leq$ 46.50 & Audiometry Score $\leq$ 0.15 & **Audiometry Class** =[Class2 Class2a, Class2 Class2b , **Class2 Class3a**, Class2 Class3b, Class2 Class6, Class2 Class7, Class2 Class8, Class2 Class9, Class3 Class1a] & Job Category = [DO, ED, EP] $\Rightarrow$ Unfit. $\rightarrow$ Matching the rule **Unfit_2**

[4]We used the SPMF library implementation [30]

Table 5.9. Top 10 Rules for Fit/Unfit Prediction

Rule ID	Features Condition			#Istance	Confidence
Unfit_0	Age_Range: [47, 59]	Audiometry Class: Class9		52	0.881
Unfit_1	Age_Range: [47, 59]	Audiometry Class: Class9	Sex: M	52	0.881
Unfit_2	Audiometry Class: Class2 Class3a			53	0.929
Unfit_3	Audiometry Class: Class2 Class3a	Sex: M		53	0.929
Unfit_4	Audiometry Class: Class9	Smoker: Smoker		57	0.890
Unfit_5	Audiometry Class: Class9	Smoker: Smoker	Sex: M	57	0.890
Unfit_6	Age_Range:[60, 71]			88	0.880
Unfit_7	Age_Range:[60, 71]	Sex: M		88	0.880
Unfit_8	Audiometry Class: Class9			97	0.881
Unfit_9	Audiometry Class: Class9	Sex: M		97	0.881
Fit_0	Age_Range: [18, 32]	Audiometry Class: Class0 Class1a	Sex: M	531	0.970
Fit_1	Age_Range: [18, 32]	Audiometry Class: Class0 Class1a		534	0.970
Fit_2	Age_Range: [18, 32]	Sex: M		690	0.936
Fit_3	Age_Range: [18, 32]			693	0.936
Fit_4	Audiometry Class: Class0 Class1a	Audiometry Score : 0,0	Sex: M	878	0.874
Fit_5	Audiometry Class: Class0 Class1a	Audiometry Score : 0,0		881	0.874
Fit_6	Audiometry Score : 0,0	Sex: M		966	0.879
Fit_7	Audiometry Score%: 0,0			969	0.880
Fit_8	Audiometry Class: Classe0 Classe1a	Sex: M		979	0.847
Fit_9	Audiometry Class: Class0 Class1a			982	0.848

- AGE $>$ 43 & **Audiometry Class** = [Class7, **Class9**, classunknown] & Smoker = Smoker & 1.55 $\leq$ Audiometry Score $<$ 2,95, Job_Category = [PCC, SJ] $\Rightarrow$ Unfit. $\rightarrow$ Matching the rule **Unfit_4**, **Unfit_8**

- **AGE $>$ 61** & Smoker= Smoker & 22,5$<$ BMI $\leq$ 26,5 & Audiometry Class= [Class0 Class0, Class0 Class 1a, Class0 Class7] $\Rightarrow$ Unfit. $\rightarrow$ Matching the rule **Unfit_6**

- **19$\leq$ AGE $<$ 29** & **Audiometry Class = Class0 Class1a** & BMI $\leq$ 19,5 & Job_Category= [ED, EP, PCC] $\Rightarrow$ Fit $\rightarrow$ Matching the rules **Fit_1**, **Fit_3**

- 45$\leq$ AGE $<$ 50 & **Audiometry Score % $\leq$ 0.25** & **Audiometry Class** =[Class0 Class0, **Class0 Class1a**, Class0 Class7] & Smoker = Ex-Smoker $\Rightarrow$Fit $\rightarrow$ Matching the rules **Fit_5**, **Fit_7**, **Fit_9**

Unfit Rule	RR	OR
Unfit_0	15.949	17.274
Unfit_2	28.449	30.931
Unfit_4	17.483	19.097
Unfit_6	15.745	18.100
Unfit_8	16.020	18.708

Fit Rule	RR	OR
Fit_1	15.544	24.812
Fit_3	6.867	12.855
Fit_5	3.256	7.305
Fit_7	3.418	9.235
Fit_9	2.598	6.623

Table 5.10. Relative Risk (RR) and Odd Ratio (OR)

5.5.1 Relative Risks and Odds Ratio

We examined some of the above rules also using the *Relative Risk (RR)* [104]. The purpose of this comparison is to further validate the association between the features highlighted as most important and the Fit/Unfit prediction. We use the rules without the *sex* feature in Table 5.9 to distribute the population according to the feature values. The *Relative Risk (RR)* quantifies how much more (or less) likely the outcome is in the exposed group compared to the unexposed group. Formally, it is defined as:

$$RR = \frac{P(Outcome = 1|\text{Rule predicts positive})}{P(Outcome = 1|\text{Rule predicts negative})} = \frac{TP/(TP + FP)}{FN/(FN + TN)}$$

where TP are the true positives, FP the false positives, FN the false negatives, and TN the true negatives. To complement the interpretation of the RR, we also report the *Odds Ratio (OR)*, which measures the odds of the outcome occurring in the exposed group compared to the unexposed one. OR is defined as $OR = \frac{TP/FP}{FN/TN} = \frac{TP \cdot TN}{FP \cdot FN}$. While the RR is generally more intuitive in expressing risk ratios, the OR is often preferred in statistical modeling and case–control studies because of its mathematical properties and symmetry. Our results, shows in Table 5.10, suggest that the pattern of features obtained from the *top-k* algorithm has a strong relation with the target class, as all the RR values are greater than 2 – Andrade sets the range of clinical significance for the RR as ≤ 0.50 RR ≥ 2.0 [4].

5.6 RQ3: Causal Relations Behind Fitness For Work

For the development of this step, the interpretation of the Fitness for Work variable was revised. Previously, due to the limited number of cases labeled as Unfit, the category Fit with restriction/regulation had been merged under the same Unfit label. In this phase, however, for the causal analysis, the labels were kept separate in order to preserve the informational specificity of each category and allow for a more accurate modeling of the causal relationships.

5.6.1 Causal Discovery and LLM validation

As CD algorithm, we adopted DirectLiNGAM[101].where the model is assumed to be linear and non-Gaussian. The causal relationships were inferred directly from the data, complemented by limited prior domain knowledge: *AGE*, *Sex* and *Job* were constrained to have no incoming edges, and "Job" type was assumed to influence work fitness. The resulting DAG, shows that *AGE* plays a central role, influencing the FFW judgment both directly and indirectly through the *BMI* and *Audiometry* pathways. The model also highlights that *Job* and *Smoker* have direct effects on FFW, while *Audiometry Class* affects *Audiometry Score*, which in turn influences the FFW outcome. Overall, the structure depicts a coherent relationship among demographic, occupational, and auditory factors determining fitness for work.
In this phase, the CausalDAG audit was performed using *GPT-4*, configured as a domain-informed expert in causal inference and graphical models. The model operated exclusively on the DOT representation of the graph and on the accompanying domain description derived from the maritime company's health-surveillance dataset. *GPT-4* produced structured and reproducible outputs, each associated with a confidence score, in accordance with the predefined analytical checklist. The procedure, as detailed in Chapter 3, entailed: (i) validation of the global topological ordering with respect to domain plausibility; (ii) evaluation of local causal relations, classifying edges as direct, mediated, or implausible; (iii) verification of acyclicity and proposition of corrective edge modifications when neces-

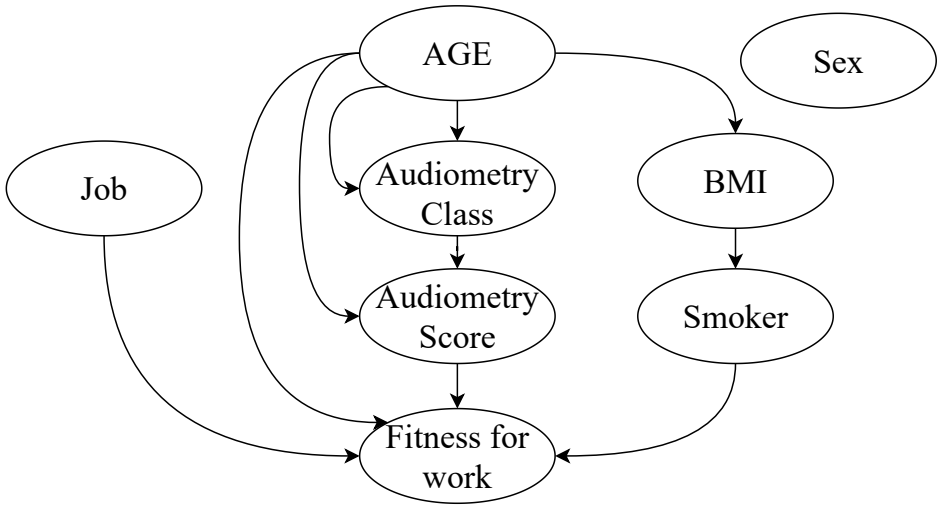


Figure 5.5. Causal graph learned via LiNGAM on seafarer dataset

sary; and (iv) assessment of minimal valid adjustment sets for estimating treatment–outcome causal effects within the public-health framework. This structured audit ensured methodological transparency and internal consistency in the causal modeling process. The exact prompt used for the audit procedure is reported in Figure 5.6.

You are an expert in causal inference and graphical models. I will provide you with a causal graph in DOT format.
Important instructions:

- Only use the information from the DOT graph and the provided domain description.
- Explicitly report any variable that appears isolated (i.e., has no incoming or outgoing edges) and discuss whether this is plausible in the given domain.
- If something is unclear or missing, explicitly say *"I don't know"* instead of guessing.
- For every answer, give a confidence score (0-100%).
- Use structured outputs (bullet points or tables where possible).

Domain Description:

This graph comes from Public Health domain. It was composed by the health surveillance record of a maritime company. The variables including:

- AGE.
- Audiometry Class: The class of damage categorizes the type and origin of hearing loss.
- Audiometry Score: The score of damage quantifies the severity of hearing impairment. In this case we have a percentage score.
- BMI.
- FFW: feature indicating Fit, Fit with restriction/regulation, and Unfit for a specific Job. Fit= 0, fit with restriction/regulation =1, Unfit= 2. This variable is the *Outcome*.
- Job: Job position.
- Sex.
- Smoker: Smoker habits. Possible values: Smoker, Ex-smoker, Not Smoker

Your task is to audit the DOT graph step by step following this checklist:

1. **Global order validation**
    - Extract a topological order of the variables from the DOT graph.
    - Is this order plausible given the domain *Public Health*, partucular *Occupational Health*?
    - Explicitly report any variable that appears isolated (i.e., has no incoming edges) and discuss whether this is plausible in the given domain.
    - Point out up to 3 variable pairs that may be incorrectly ordered.
    - Provide an overall confidence score.

2. **Local edge validation**
    - For each edge $X \to Y$ in the DOT graph:
    - Is the effect plausibly direct, mediated, or implausible?
    - Answer in the format: [direct / mediated / implausible] + short explanation + confidence.

3. **Cycle check**
    - Identify if any cycles exist in the DOT graph.
    - If cycles exist, are they plausible in a causal model?

- Suggest which edge(s) should be removed or reversed to restore acyclicity, with justification.

4. **Adjustment set validation**
   - Suppose we want to estimate the causal effect of *BMI* on *Fitness for Work*.
   - What is the minimal valid adjustment set to block all backdoor paths?
   - Is this set plausible in the domain [insert domain description]?
   - If not, suggest corrections (variables to remove/add) with reasoning.

Figure 5.6. Prompt for Causal Domain Validation of Occupational Health DAG

The complete output generated by GPT-4 in response to the audit prompt is reported, in Figure 5.7 below in its original form. The audit shows the global topological order of the variables, identifying Sex as an isolated node and detecting three questionable causal directions, such as BMI → Smoker. It then assesses the plausibility of each local causal edge, distinguishing between direct, mediated, and implausible relationships, each supported by confidence levels. The analysis confirms that the graph is acyclic and therefore consistent with a valid DAG structure. Finally, it defines the minimal adjustment sets required for causal estimation, determining that *AGE* and *Audiometry Class* should be controlled for when estimating the effect of Audiometry Score on FFW, and *AGE* alone for the effect of Audiometry Class on FFW. Overall, the audit highlights a coherent and interpretable causal structure, suitable for inclusion in scientific reporting. The audit also demonstrates methodological transparency by explicitly acknowledging cases of uncertainty or incomplete information.

Causal Graph Audit (Public Health, maritime company)
1) Global order validation

- *Topological order (one valid example):* AGE → {BMI, Job, *Audiometry Class*} → Smoker → *Audiometry Score* → *Fitness for work*.
- *Plausibility (Occupational / Public Health):*
  - AGE early is plausible (age precedes most health/work variables).

  - Audiometry Class before Audiometry Score is defensible if "class" represents underlying pathology that determines measured severity; however, in many workflows class is derived from the score.
  - BMI →Smoker ordering (BMI causes smoking) is questionable; typically smoking influences BMI, not vice-versa.
  - Smoker→ Job Category is also questionable; job often influences smoking norms/exposure rather than smoking determining job assignment.

- *Isolated Variables*: The variable *Sex* has no incoming or outgoing edges. This is **not plausible**, as sex is a known causal factor influencing BMI, hearing loss, and work fitness. It should therefore be connected to other variables (confidence 95
- *Potentially mis-ordered pairs (max 3):*
  - **BMI → Smoker** is implausible, since smoking typically affects BMI, not the other way around (confidence 90%).
  - **AGE → Audiometry Score** appears likely mediated, because age affects hearing primarily through deterioration (captured by Audiometry Class), rather than directly influencing the score (confidence 80%).
  - **AGE → Fitness for work** also seems mediated, as age influences health and functional capacity (via BMI or hearing), which in turn determine fitness (confidence 75%).
- *Overall confidence:* **80%**

2) Local edge validation Each entry includes a classification — direct, mediated, or implausible — plus explanation and confidence.

- $BMISmoker$: implausible direction; smoking influences BMI, not vice versa (confidence 90%).
- $AGEAudiometryClass$: plausible direct causal link; age-related hearing loss is well established (confidence 95%).
- $SmokerFitnessforwork$: plausible direct effect; smoking impacts health and functional ability (confidence 85%).
- $JobFitnessforwork$: plausible direct relationship; job demands directly constrain fitness evaluations (confidence 95%).
- $AGEFitnessforwork$: likely indirect, mediated through health variables such as BMI or hearing (confidence 75%).
- $AudiometryScoreßFitnessforwork$: plausible direct link; hearing capacity influences job performance (confidence 90%).

- $AGEBMI$: well-documented direct relationship; body weight tends to vary with age (confidence 90%).
- $AudiometryClassßAudiometryScore$: direct and coherent relationship; class defines type, score defines severity (confidence 95%).
- $AGEAudiometryScore$: likely mediated through Audiometry Class, making the direct edge redundant (confidence 75%).

3) Cycle check: No cycles are present in the graph. The structure is a valid DAG, and no edge removals are required. *Confidence:100%*

4) Adjustment set validation for Y = *Fitness for work*

1. 1: Estimate the effect of Audiometry Score on Fitness for work. The relevant backdoor paths are:
   - Audiometry Score ← Audiometry Class ← AGE → Fitness
   - Audiometry Score ← AGE → Fitness
   - Audiometry Score ← AGE → BMI → Smoker → Fitness

   The minimal adjustment set required to block these paths includes AGE and Audiometry Class. AGE is a common cause of both Score and Fitness, while Audiometry Class is a parent of Score and an ancestor of Fitness. This adjustment is considered plausible, with a confidence of 90%.
2. : Estimate the effect of Audiometry Class on Fitness for work. The main backdoor paths are:
   - Audiometry Class ← AGE → Fitness
   - Audiometry Class ← AGE → BMI → Smoker → Fitness

   The minimal adjustment set required is AGE, a known confounder. This adjustment is considered highly plausible, with a confidence of 95%.

Overall confidence: 90% **Closing notes**

- **Most critical edge:** BMI → Smoker (direction should be reconsidered).
- **Disconnected variable:** Sex (likely relevant in practice, but uninformative in this graph).
- **Adjustment practice:** AGE is often the minimal confounder for effects on *Fitness for work*.

Overall audit confidence: 83%

Figure 5.7. Gpt-4 Answers of Prompt for Causal Domain Validation of Occupational Health DAG

These adjustments would enhance both the epidemiological validity and

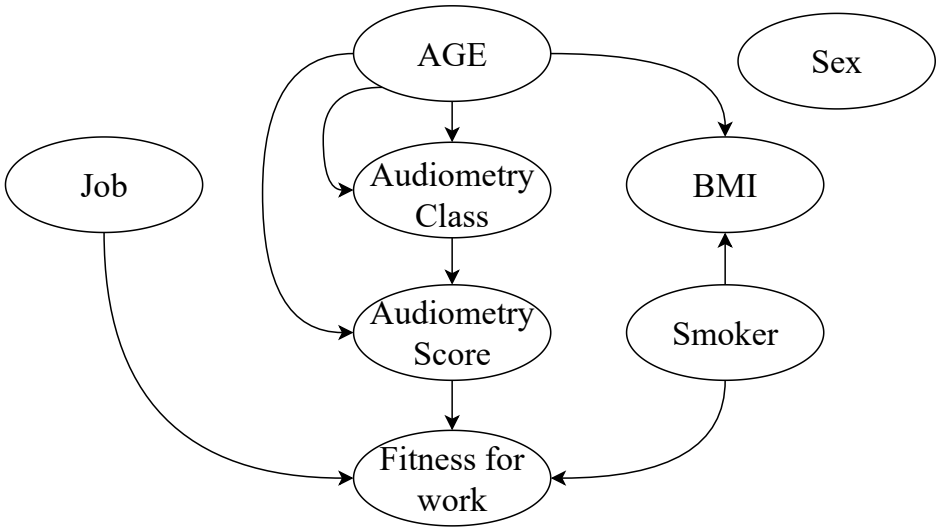


Figure 5.8. Occupational Causal Graph post LLM validation

the interpretability of the graph within the occupational health domain. All other relations were retained as validated, and the overall causal structure was considered coherent with the public-health context of the study. The updated causal graph, reflecting these structural adjustments, is presented in Figure 5.8 and was subsequently used as the reference model for the following analyses. The falsification analysis indicates that the proposed DAG is informative, as none of the 20 permuted structures fall within its Markov equivalence class. Although the DAG violates 14 out of 24 local Markov conditions, it performs better than 95% of the permuted DAGs ($p = 0.05$), suggesting that its causal structure is substantially supported by the data. Given the predefined significance level and the graph's informativeness, the DAG is not rejected. Overall, the DAG is falsifiable but not falsified, supporting its plausibility as a representation of the underlying causal relationships captured in the dataset. Before proceeding to causal inference, the DAG was fitted to the data to obtain a corresponding SCM, ensuring that the model's structural equations reflect the observed statistical dependencies.

5.6.2 Causal Inference Job Category-based

The causal inference step was conducted on five distinct groups of subjects, defined according to the *Job Category*. Although all groups belong to the same occupational context, each is associated with a slightly different health surveillance protocol and with variations in the type of personal protective equipment (PPE) used, reflecting differences in both medical assessment procedures and preventive measures adopted. These controlled differences were introduced to make the effects of the query *intervention* and *counterfactual* simulations more evident. In both cases, drastic conditions were applied, by setting the feature *Audiometry Score* $= 0$, thereby simulating a condition of complete absence of damage. This approach made it possible to isolate the potential impact of this condition and to evaluate its influence on the distribution of fitness outcomes within each protocol category.
Each group is characterised by the following initial distribution of FFW:

- **DO**: Fit: 53, Fit with restriction/regulation: 70, Unfit: 2
- **ED**: Fit: 537, Fit with restriction/regulation: 228, Unfit: 2
- **EP**: Fit: 538, Fit with restriction/regulation: 159, Unfit: 2
- **PCC**: Fit: 178, Fit with restriction/regulation: 148, Unfit: 2
- **SJ**: Fit: 66, Fit with restriction/regulation: 25, Unfit: 1

These distributions represent the ground truth, and serve as the reference point for evaluating the effects of the subsequent intervention and counterfactual analyses.

Intervention The analysis of the causal intervention on the Audiometry Score revealed a substantial change in the distribution of FFW. Estimates were obtained with 5000 repetitions. For each category, the median and 95% confidence interval (CI) were computed to assess both the central tendency and the variability of the intervention effect. As reported in Table 5.11, the intervention leads to a general increase in the proportion of fit individuals, accompanied by a reduction in the fit with condition category. The effect is expressed in the column Δ *Variation vs. Ground*

Truth as the absolute difference, in percentage, relative to the observed (real) distribution. The **DO** group exhibits the most pronounced change: the proportion of fully fit individuals rises from 42.4% in the observed data to 84.7% under intervention, corresponding to a Δ of +42.4 %, while the fit with restriction/regulation category decreases by 40.8 %. Similar trends, with different magnitudes, are observed in the **PCC** (+29.0 %), **ED** (+19.8 %), SJ (+12.0 %, and **EP** (+8.9 %) groups, all characterized by relatively narrow and consistent confidence intervals. Although the percentage of unfit cases appears very small, this is due to the extremely limited number of non-fit individuals in the real dataset (only one or two per group). After the intervention, no individual remains unfit, indicating that the model reassigns all subjects toward higher levels of fitness.
Overall, the results indicate that the intervention produces a positive and expected effect, increasing the likelihood of full fitness and reducing the occurrence of conditional restrictions. This behavior confirms that the causal model reproduces the observed dynamics realistically and coherently, supporting its consistency with empirical evidence.

Counterfactual The counterfactual analysis estimated the distribution of Fitness for work states under the hypothetical condition *Audiometry Score = 0*. For each professional group (DO, ED, EP, PCC, SJ), the proportions of fit, fit with restriction/regulation, and unfit individuals were computed with 5000 repetitions, from which the median and 95% confidence intervals were derived. As shown in Table 5.13, the counterfactual produces a consistent increase in the number of fit individuals across all groups, accompanied by a corresponding reduction in the fit with condition category. The largest variation occurs in group **DO** (+24.6 %), while smaller but positive effects are observed in **ED**, **SJ**, **EP**, and **PCC**. Changes in the unfit category are negligible, reflecting the very limited number of non-fit subjects in the observed data. Overall, these counterfactual results behave as expected: setting the Audiometry Score to zero leads to a general improvement in work fitness, confirming that the causal model captures the expected relationship between hearing ability and occupational fitness and reproduces realistic, stable outcomes under hypothetical interventions.

Table 5.11. Fitness for work distribution after do(Audiometry Score = 0), on 5000 repet.

Job	FFW	Median of istance	95% Confidence Intervals	Δ Variation vs. Ground Truth
DO	Fit	103	[99, 107]	+42.4%
	Fit with condition	22	[18 , 26]	−40.8%
	Unfit	-	-	−1.6%
ED	Fit	698	[687, 709]	+19.8%
	Fit with condition	69	[63 , 89]	−19.6%
	Unfit	-	-	−0.3%
EP	Fit	624	[610, 636]	+8.9%
	Fit with condition	75	[63, 89]	−8.6%
	Unfit	-	-	−0.3%
PCC	Fit	275	[269, 282]	+29.0%
	Fit with condition	53	[46, 59]	−28.4%
	Unfit	-	-	−0.6%
SJ	Fit	77	[71, 82]	+12.0%
	Fit with condition	15	[10, 21]	−10.9%
	Unfit	-	-	−1.1%

5.7 Statistical Validation

Several robustness tests from the DoWhy framework were applied to evaluate the stability and credibility of the estimated causal model. Each test introduces a controlled perturbation to the data or to the model structure, allowing the verification of whether the detected causal relationship persists only when it truly exists. The results presented in Table 5.16 show that the identified causal relationship is stable, specific, and statistically consistent. Similarly, the Random Common Cause test shows that the estimated effect remains virtually unchanged even when random confounders are introduced, indicating that the model is resilient to irrelevant sources of variation. A complementary robustness check was also performed by

Table 5.13. Fitness for work distribution under the hypothetical scenario Audiometry Score = 0, on 5000 repet.

Job	FFW	Median of istance	95% Confidence Intervals	Δ Variation vs. Ground Truth
DO	Fit	82	[70, 88]	+24.6%
	Fit with condition	39	[29, 46]	−24.7%
	Unfit	4	[0, 5]	+0.1%
ED	Fit	650	[616, 671]	+16.1%
	Fit with condition	115	[85, 125]	−16.1%
	Unfit	2	[0, 4]	+0.0%
EP	Fit	577	[543, 588]	+6.5%
	Fit with condition	119	[93, 135]	−6.5%
	Unfit	3	[0, 5]	+0.0%
PCC	Fit	190	[165, 206]	+4.2%
	Fit with condition	133	[108, 147]	−5.2%
	Unfit	5	[2, 9]	+1.0%
SJ	Fit	75	[60, 90]	+13.3%
	Fit with condition	15	[6, 16]	−13.4%
	Unfit	2	[0, 3]	+0.4%

repeating the estimation on multiple random subsets of the dataset (95% of the data, 10 repetitions). The results remained consistent across samples, suggesting that the model's conclusions are stable and generalizable. The *p-values* obtained across all tests align with the expectation that such artificial perturbations should not materially affect a valid causal relationship. Overall, these robustness tests provide strong empirical evidence for the methodological soundness and stability of the causal model, forming a solid foundation for the subsequent analysis of conditional treatment effects (CATE) presented in the next section. After confirming the robustness and internal validity of the estimated causal effect of *Audiometry Score* on FFW through multiple robustness tests, the analysis proceeds to estimate the treatment effect using advanced causal inference models capable

of capturing individual-level heterogeneity in responses. To do this, several CATE estimators were implemented, including linear interaction models, meta-learners, and double machine learning frameworks. To complement these, the Causal Forest model from the EconML library [25] was employed as the primary estimator for individual-level effects. Causal forests are honest tree ensembles specifically designed to estimate heterogeneous treatment effects by combining sample-splitting and effect-focused splits, thus yielding valid inference at the individual level [113]. The resulting CATE distribution was narrowly centered around zero, indicating small overall effects but localized heterogeneity across subjects. Most individual estimates were close to zero, with both slightly positive and negative devia-

Refutation Test	**Estimated Effect**	**New Effect**	**p-value**	**Interpretation**
Dummy Outcome Refuter	0.0	-0.0009	0.94	Random outcomes eliminate the effect, confirming the model's true causality.
Random Common Cause Refuter	0.0822	0.082	0.98	Model is stable; a random confounder leaves the effect unchanged.
Placebo Treatment Refuter	0.0822	0.00006	0.96	Effect vanishes with randomized treatment, confirming specificity.
Data Subset Refuter	0.0822	0.0820	0.437	Stable effect across subsamples shows good generalizability.

Table 5.16. Summary of Refutation Tests for Audiometry Score Effect on FFW

tions, suggesting that while the average effect is modest, specific subgroups experience more pronounced responses. The ATE, obtained by averaging the CATE distribution, was approximately 0.011, with a 95% confidence interval of $[-0.027; 0.050]$. This reflects a near-neutral average impact of hearing performance on work fitness when considered over the entire population, yet consistent with the direction and variability observed in previous estimators. The accuracy and reliability of the estimated CATE were further evaluated through doubly robust diagnostics using the DRTester framework [5], which assesses how well the model captures true heterogeneity and ranks individuals according to treatment response. Table 5.17 summarizes the results. The BLP slope (0.96, $p < 0.001$) demonstrated a strong alignment between predicted and realized effects, while the calibration R^2 (0.727) indicated that approximately 73% of the observed effect variance was explained by the model. Positive QINI (0.004, $p = 0.005$) and AUTOC (0.014, $p = 0.003$) values further supported the internal consistency and ranking capability of the estimator. The findings indicate that the causal relationship between hearing ability and work fitness exhibits mild but structured heterogeneity across individuals. The small average magnitude of the effect may be partly explained by the imbalanced distribution of the outcome variable, where the predominance of "fit" subjects limits the variability of the outcome and may lead causal estimators to underestimate the true underlying effect. Importantly, this apparent contrast—strong robustness of the overall causal model but weaker evidence at the individual level—is fully consistent with expectations: while refutation tests validate the structural soundness of the ATE, CATEs inherently require greater sample balance and variability to capture subgroup-level differences. Thus, a model can exhibit strong internal validity and reliable ATE estimation while showing limited precision in detecting individual-level heterogeneity, especially under outcome imbalance.

[5] Metrics detail: https://www.pywhy.org/EconML/_autosummary/econml.validate.DRTester.html

Table 5.17. Validation of CATE estimates with DRTester (T= Audiometry Score)

Metric	Estimate	SE	p-value	Interpretation
BLP slope	0.96	0.172	<0.001	Positive and significant, CATEs capture real heterogeneity.
Calibration R^2	0.727	–	–	73% of variance in group effects explained; moderate calibration.
QINI (uplift)	0.004	0.001	0.005	Significant uplift, model ranks beneficiaries better than random.
AUTOC	0.014	0.005	0.003	Significant ranking ability across subgroups.

5.8 The Role of *Sex* as a Feature

In the Table 5.9 we reported some rules depend on the value of feature Sex, but no match was found for these rules with the DT rules. The reason of this lack is that the feature Sex is not an important feature for 3 of 4 model used (with features importance equal to 0), as reported in the Fig 5.3. A proof of this is given by the number of instances in the Table 5.9: for the Unfit rules the number of instance is the same with or without the Sex features, the Fit rules the difference is at least of 3 instances. However the lower importance of this feature can be caused by a greater presence of male individuals (2008) compared to female individuals (3). For this reason, we have bring back all the previously steps without the *Sex* feature. For **RQ1**, nothing has changed except a minimum change in the LR, but the absence of the *Sex* doesn't impact the importance of the other features. For the **RQ2**, all the rules without *Sex* have not changed. The new rules, and some of the tree-matching, are reported as follows:

- **No_Sex_Unfit_0**: AGE_Range: [47, 59], Audiometry Class: Class9, Smoker: Smoker with #Instance= 31, Confidence= 0.911764
- **No_Sex_Unfit_1**: AGE_Range: [47, 59], Audiometry Class: Class2 Class3a with #Instance= 35, Confidence= 0.945945
- **No_Sex_Unfit_2**: AGE_Range: [60,71], Job Category: Engineering Department with #Instance=36, Confidence= 0.947368
- **No_Sex_Unfit_3**: Audiometry Class: Class9, Job Category: Engineering Department with #Instance= 42 , Confidence= 0.913043
- **No_Sex_Unfit_4**: AGE_Range: [60,71], Smoker: Smoker with #Instance= 44, Confidence= 0.936170
- **No_Sex_Fit_0**: Audiometry Class: Class0 Class1a & Audiometry Score %: 0.0 & Smoker: Smoker with #Instance= 448, Confidence= 0.854961
- **No_Sex_Fit_1**: AGE_Range: [18,32] & Audiometry Class: Class0 Class1a & Audiometry Score %: 0.0 with #Instance= 483, Confidence= 0.969879
- **No_Sex_Fit_2**: Audiometry Class: Class0 Class1a & Smoker: Smoker with #Instance 495, Confidence 0.859375
- **No_Sex_Fit_3**: Audiometry Score %: 0.0 & Smoker: Smoker with #Instance= 500, Confidence= 0.827814
- **No_Sex_Fit_4**: AGE_Range: [18,32] & Audiometry Score %: 0.0 with #Instance= 510, Confidence= 0.967741

- **AGE > 56.5** & BMI >29.50 & **Job Category**= [Deck Officers, **Engineering Department**,Engineering Positions] & Audiometry Score % > 0.20 ⇒ Unfit. → Matching the rule **No_Sex_Unfit_2**.
- AGE > 46.50 &**Audiometry Class**=[Class4 Class1a, Class4 Class2a, Class4 Class4a,
  Class4 Class4b, Class4 Class5a, Class4 Class6, Class4 Class7, Class4 Class9, Class5 Class1a, Class5 Class2a, Class5 Class3a, Class5

Class4b, Class5 Class5a, Class5 Class6, Class5 Class8, Class5 Class9, Class6, Class7, Class7 Class1a, Class7 Class4a, Class7 Class7, **Class9**, classunknow] & 0.95 < Audiometry Score % ≤ 2.25 & 27 ≤ BMI ≤ 28.50 & **Job Category**= [Deck Officers, **Engineering Department**] ⇒ Unfit. → Matching the rule **No_Sex_Unfit_3**.

Therefore, the male predominance in these roles introduces a bias that limits its contribution in this domain. However, the issue emerges already in the CD phase, where the DAG in Figure 5.5 shows *Sex* as an isolated variable, disconnected from the rest of the model. This lack of connections represents a red flag, revealing an **implausible** disconnection between variables, as also highlighted by Gpt-4 during the validation phase. This confirms that the limited relevance of *Sex* is not due to its intrinsic insignificance, but rather to a bias in the data distribution. At first glance, given such low variability, the *Sex* feature might have been discarded before the analysis. However, doing so would have been a mistake. Documented biological differences between men and women make *Sex* a potentially relevant variable, and its removal would have meant ignoring an important factor that could affect health outcomes. In today's context, where the evaluation of sex-related effects is increasingly recognized as crucial in occupational health, excluding *Sex* would risk overlooking an important dimension of analysis.

5.9 Threats to validity

This chapter explored the contributing factors of FFW judgments in the maritime domain. Despite the limited variability of the data and the lack of prior domain expert, the preliminary causal model produced realistic outcomes. However, serveral limitations emerged that may affect the validity of the model and the interpretation of its results. The main threats identified in this phase are following reported.

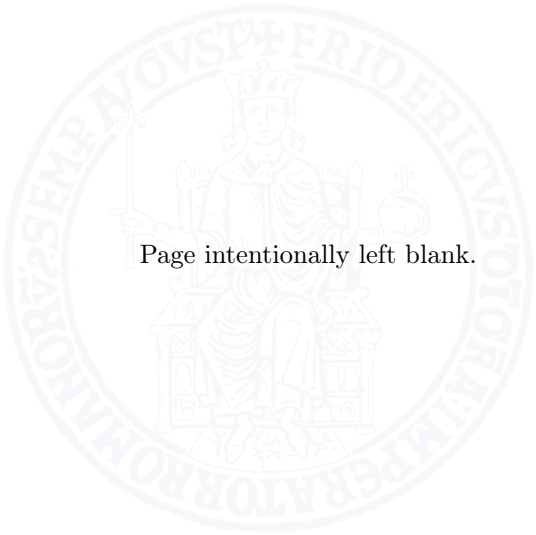
Internal validity concerns the reliability of some causal dependencies may be affected by the presence of data bias and unbalanced sampling. In particular, the unbalancing in the *Sex* feature introduces a selection bias that may affect the learned relationships and lead to partial or incomplete causal structures. These aspects influence the internal consistency of the

model and its ability to represent genuine and real causal effects derived from the observed data. Moreover, the strong imbalance in the outcome variable, with a predominance of "fit" subjects, reduces the variability of the response and may attenuate the detection of individual-level heterogeneity. These aspects influence the internal consistency of the model and its ability to represent genuine and real causal effects derived from the observed data.

External validity is limited by the reliance on a single maritime dataset, whose population composition does not fully represent the diversity of the broader workforce. This constraint restricts the external applicability of the results and highlights the importance of integrating new and more balanced data sources to support generalization beyond the current context.

Construct validity refers to the way in which the concepts and variables used in the model represent the real-world phenomena they intend to capture. Some complex constructs, such as FFW, were represented through simplified categorical attributes. This may overlook intermediate or latent dimensions (e.g., psychological, behavioral, or environmental factors) that influence real-world outcomes. Such simplifications can limit the ability of the model to capture the multidimensional nature of occupational health assessments.

Overall, the main limitations identified in this chapter derive from the quality, balance of the available data, rather than from methodological inadequacy. Nevertheless, this phase of the study provided valuable insights that represent a crucial step toward more comprehensive and reliable causal framework for maritme health surveillance.



Page intentionally left blank.

Chapter 6

Conclusions

The world is nothing but an infinite chain of causes and effects.

Blaise Pascal

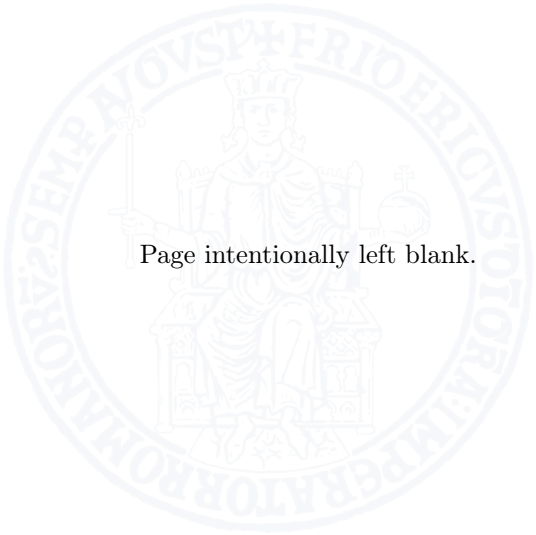
This thesis has explored the challenge of applying causal reasoning to real-world clinical data, where observational nature, data where observational nature, data heterogeneity, and limited prior knowledge often limit the identification of true causal mechanisms. The proposed framework was designed to bridge the gap between theoretical CI and its practical use in healthcare research, integrating methods for CD, domain validation, and statistical inference into a single pipeline. Through the two case studies presented, the framework demonstrated its ability to extract plausible causal relationships, simulate intervention and counterfactual scenarios, and provide clinically interpretable insights.

Advantages The proposed framework builds on the inherent strengths of observational studies, which provide several practical and ethical advantages over RCTs. While RCTs remain the gold standard for causal inference, they are often expensive, time-consuming, and limited to carefully selected and homogeneous populations. In contrast, observational data enable research at lower cost and without the ethical constraints associated with randomization, allowing the evaluation of risk factors or treatment effects that cannot be directly tested on human subjects. Moreover, they make it possible to study outcomes in under-studied or excluded popula-

tions, such as ethnic minorities or individuals who fall outside typical trial inclusion criteria.
Within this context, the proposed framework enhances the analytical potential of observational data by integrating causal discovery, domain validation, and statistical inference into a unified workflow. By combining data preprocessing, machine learning–based feature selection, LLM-assisted validation, and statistical robustness tests, the methodology enables the extraction and validation of plausible causal structures directly from real-world data. This approach supports both interventional and counterfactual analyses, allowing the simulation of clinical protocols or preventive actions before real-world implementation, and offering an interpretable, ethical, and cost-effective alternative for exploratory causal reasoning in healthcare.

Limitation Despite its strengths, several methodological limitations must be acknowledged. First, the quality and reliability of the inferred causal structures are inherently dependent on data completeness and preprocessing accuracy. Missing, noisy, or biased variables can compromise both the discovery and inference phases. Second, while LLM-based validation enhances interpretability and domain coherence, it remains constrained by potential model biases, prompt sensitivity, and limited access to specialized medical context. For this reason, expert domain evaluation remains essential to ensure the clinical reliability and plausibility of the inferred causal relationships. Another relevant limitation concerns the availability and standardization of *Electronic Health Records* (EHRs). Although EHR systems are increasingly adopted worldwide, their accessibility, completeness, and interoperability still vary across healthcare institutions and countries, making large-scale and reproducible causal analyses difficult. Additionally, the proposed framework involves several computationally intensive stages — including feature selection, causal discovery, and statistical validation — which may increase the overall computational cost, particularly for high-dimensional clinical datasets. These factors define the current boundaries of the proposed methodology and highlight opportunities for improvement.

Future work Future research could focus on improving the discovery phase by selecting algorithms based on the relationships among individual features, as well as refining the prompting process through the inclusion of domain-specific key phrases that help guide the model toward more accurate and context-aware responses. Furthermore, applying the proposed framework to larger and more diverse datasets - including those related to different pathologies or research contexts - could not only support the development of a more generalizable global model but also lead to progressive improvements in the framework's overall design and performance.



Page intentionally left blank.

Appendix A

Health Surveillance Protocol

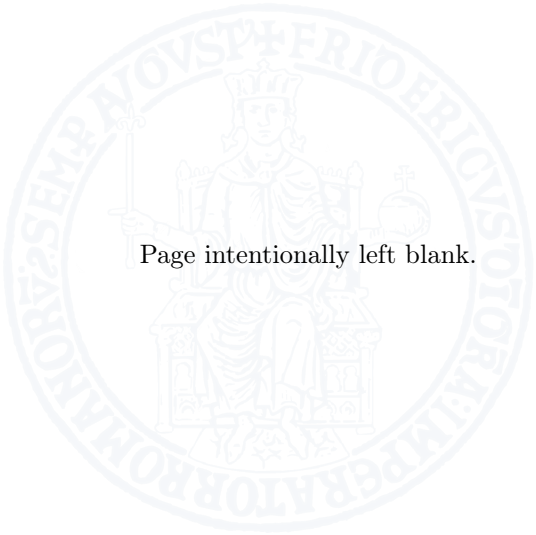
This appendix summarizes the health protocol shown below is the one applied by the companies considered in this study. The table below lists the required health checks for each job role.

Table A.1. Health protocol

JOB	Audiometry	Spirometry	ECG	AUDIT	Alcohol test	Rachis physical examination
1st Engineer	Yearly	Yearly	Biennial	Yearly	Yearly	Yearly
2nd Engineer	Yearly	Yearly	Biennial	Yearly	Yearly	Yearly
2nd Mate	-	Yearly	Biennial	Yearly	Yearly	Yearly
3rd Engineer	Yearly	Yearly	Biennial	Yearly	Yearly	Yearly
3rd Mate	-	Yearly	Biennial	Yearly	Yearly	Yearly
Able Seafarer Deck	Yearly	Yearly	-	Yearly	Yearly	Yearly
Able Seafarer Engine	Yearly	Yearly	Biennial	Yearly	Yearly	Yearly
Able Seaman	Yearly	Yearly	Biennial	Yearly	Yearly	Yearly
Barman /Bartender	-	Yearly	-	-	-	Yearly
Bosun	Yearly	Yearly	Biennial	Yearly	Yearly	Yearly
Brassworker	Yearly	Yearly	Biennial	Yearly	Yearly	Yearly
Cadet Electrician	Yearly	Yearly	Biennial	Yearly	Yearly	Yearly
Cadet Purser	-	Yearly	-	-	-	Yearly
Carpenter	Yearly	Yearly	Biennial	Yearly	Yearly	Yearly
Catering Boy	-	Yearly	-	-	-	Yearly
Chief Engineer	Yearly	Yearly	Biennial	Yearly	Yearly	Yearly
Chief Mate	-	Yearly	Biennial	Yearly	Yearly	Yearly
Cook	-	Yearly	-	-	-	Yearly
Deck Boy	-	Yearly	Biennial	Yearly	Yearly	Yearly
Deck Cadet	-	Yearly	Biennial	Yearly	Yearly	Yearly
Deck Officer	-	Yearly	Biennial	Yearly	Yearly	Yearly
Doctor	-	Yearly	-	Yearly	Yearly	Yearly
Driver	Yearly	Yearly	Biennial	Yearly	Yearly	Yearly
Electrician	Yearly	Yearly	Biennial	Yearly	Yearly	Yearly
Electrotechnical Officer	Yearly	Yearly	Biennial	Yearly	Yearly	Yearly
Engine Officer	Yearly	Yearly	Biennial	Yearly	Yearly	Yearly
Enginee Rating	Yearly	Yearly	Biennial	Yearly	Yearly	Yearly
Engineer Cadet	Yearly	Yearly	Biennial	Yearly	Yearly	Yearly
Fitter	Yearly	Yearly	Biennial	Yearly	Yearly	Yearly
Galley Boy/ Utility	-	Yearly	-	-	-	Yearly
Housekeeping Utility	-	Yearly	-	-	-	Yearly
Junior Messman	-	Yearly	-	-	-	Yearly
Master	-	Yearly	Biennial	Yearly	Yearly	Yearly
Messman	-	Yearly	-	-	-	Yearly
Motorman	Yearly	Yearly	Biennial	Yearly	Yearly	Yearly
Ordinary Seaman	-	Yearly	-	Yearly	Yearly	Yearly
Purser	-	Yearly	-	-	-	Yearly
Storekeeper	-	Yearly	-	-	-	Yearly
Trainee Electrician	Yearly	Yearly	Biennial	Yearly	Yearly	Yearly
Trainee Storekeeper	-	Yearly	-	-	-	Yearly
Wiper	Yearly	Yearly	Biennial	Yearly	Yearly	Yearly

Table A.2. Health protocol continue

JOB	GOT/AST	GPT/ALT	Azotemia	Blood count	g-GT	Glycemia	Drug Test
1st Engineer	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly
2nd Engineer	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly
2nd Mate	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly
3rd Engineer	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly
3rd Mate	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly
Able Seafarer Deck	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	-
Able Seafarer Engine	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	-
Able Seaman	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	-
Barman /Bartender	-	-	-	-	-	-	-
Bosun	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	-
Brassworker	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	-
Cadet Electrician	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	-
Cadet Purser	-	-	-	-	-	-	-
Carpenter	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	-
Catering Boy	-	-	-	-	-	-	-
Chief Engineer	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly
Chief Mate	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly
Cook	-	-	-	-	-	-	-
Deck Boy	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	-
Deck Cadet	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	-
Deck Officer	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	-
Doctor	-	-	-	-	-	-	-
Driver	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly
Electrician	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	-
Electrotechnical Officer	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly
Engine Officer	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly
Enginee Rating	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	-
Engineer Cadet	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	-
Fitter	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	-
Galley Boy/ Utility	-	-	-	-	-	-	-
Housekeeping Utility	-	-	-	-	-	-	-
Junior Messman	-	-	-	-	-	-	-
Master	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly
Messman	-	-	-	-	-	-	-
Motorman	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly
Ordinary Seaman	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	-
Purser	-	-	-	-	-	-	-
Storekeeper	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	-
Trainee Electrician	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	-
Trainee Storekeeper	-	-	-	-	-	-	-
Wiper	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	-



Page intentionally left blank.

Bibliography

[1] Occupational health surveillance. *Workplace Health & Safety*, 62(1):3–4, 2014.

[2] 15. management of diabetes in pregnancy: standards of care in diabetes—2024. *Diabetes Care*, 47(Supplement_1):S282–S294, 2024.

[3] Md Zahangir Alam, M Saifur Rahman, and M Sohel Rahman. A random forest based predictor for medical data classification using feature ranking. *Informatics in Medicine Unlocked*, 15:100180, 2019.

[4] Chittaranjan Andrade. Understanding relative risk, odds ratio, and related terms: as simple as it can get. *The Journal of clinical psychiatry*, 76(7):21865, 2015.

[5] Muhammad Anshari. Redefining electronic health records (ehr) and electronic medical records (emr) to promote patient empowerment. *IJID (International Journal on Informatics for Development)*, 8(1):35–39, Jun. 2019.

[6] American Diabetes Association. Gestational diabetes mellitus. *Diabetes Care*, 26(suppl_1):s103–s105, 01 2003.

[7] Susan Athey and Stefan Wager. Estimating treatment effects with causal forests: An application. *Observational studies*, 5(2):37–51, 2019.

[8] P W J Bartrip. History of asbestos related disease. *Postgraduate Medical Journal*, 80(940):72–76, 02 2004.

[9] Rohan Bhardwaj, Ankita R Nambiar, and Debojyoti Dutta. A study of machine learning in healthcare. In *2017 IEEE 41st annual computer software and applications conference (COMPSAC)*, volume 2, pages 236–241. IEEE, 2017.

[10] Guthrie S. Birkhead, Michael Klompas, and Nirav R. Shah. Uses of electronic health records for public health surveillance to advance public health. *Annual Review of Public Health*, 36(Volume 36, 2015):345–359, 2015.

[11] Patrick Blöbaum, Peter Götz, Kailash Budhathoki, Atalanti A. Mastakouri, and Dominik Janzing. Dowhy-gcm: An extension of dowhy for causal inference in graphical causal models. *Journal of Machine Learning Research*, 25(147):1–7, 2024.

[12] Ernest Yeboah Boateng, Daniel A Abaye, et al. A review of the logistic regression model with emphasis on medical research. *Journal of data analysis and information processing*, 7(04):190, 2019.

[13] D Bosio, M Coggiola, A Baracco, P Andreis, and F Perrelli. The issue of medico-legal assessment of noise induced hearing loss: comparison of methods. *Giornale Italiano di Medicina del Lavoro ed Ergonomia*, 33(3 Suppl):142–145, 2011.

[14] M. Caliendo and S. Kopeinig. Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1):31–72, 2008.

[15] Victor Chang, Jozeene Bailey, Qianwen Ariel Xu, and Zhili Sun. Pima indians diabetes mellitus classification based on machine learning (ml) algorithms. *Neural Computing and Applications*, 35(22):16157–16173, 2023.

[16] Saul Charapaqui-Miranda, Katherine Arapa-Apaza, Moises Meza-Rodriguez, and Horacio Chacon-Torrico. *Comparing predictive machine learning algorithms in fit for work occupational health assessments*, pages 218–225. Springer International Publishing, 2019.

[17] Marie-Laure Charpignon, Bella Vakulenko-Lagun, and et al. Causal inference in medical records and complementary systems pharmacology for metformin drug repurposing towards dementia. *Nature communications*, 13(1):7652, 2022.

[18] D. M. Chickering. Learning equivalence classes of bayesian-network structures. *J. Mach. Learn. Res.*, 2:445–498, mar 2002.

[19] Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):39–67, 2020.

[20] D. Colombo, M. H. Maathuis, M. Kalisch, and T. S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 40(1):294–321, 2012.

[21] American Diabetes Association Professional Practice Committee. 2. diagnosis and classification of diabetes: Standards of care in diabetes—2024. *Diabetes Care*, 47(Supplement_1):S20–S42, 12 2023.

[22] Anthony C Constantinou, Neville K Kitson, and Alessio Zanga. Using gpt-4 to guide causal machine learning. *Expert Systems with Applications*, 268:126120, 2025.

[23] Angus Deaton and Nancy Cartwright. Understanding and misunderstanding randomized controlled trials. *Social science & medicine*, 210:2–21, 2018.

[24] Natalie A DiPietro. Methods in epidemiology: observational study designs. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 30(10):973–984, 2010.

[25] MR EconML. Econml: A python package for ml-based heterogeneous treatment effects estimation, 2019.

[26] Elias Eulig, Atalanti A Mastakouri, Patrick Blöbaum, Michaela Hardt, and Dominik Janzing. Toward falsifying causal graphs using a permutation-based test. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26778–26786, 2025.

[27] European Agency for Safety and Health at Work, F Michaelsen, E Tompa, T Bree, M Emmerik, A Moffidi, Y Jung, L Porsch, and S Heuvel. *The value of occupational safety and health and the societal costs of work-related injuries and diseases*. Publications Office, 2019.

[28] Eurostat. European statistics on accidents at work (esaw) summary methodology. https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-RA-12-102, 2013.

[29] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

[30] Philippe Fournier-Viger, Jerry Chun-Wei Lin, Antonio Gomariz, Ted Gueniche, Azadeh Soltani, Zhihong Deng, and Hoang Thanh Lam. The spmf open-source data mining library version 2. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part III 16*, pages 36–40. Springer, 2016.

[31] Philippe Fournier-Viger, Cheng-Wei Wu, and Vincent S Tseng. Mining top-k association rules. In *Advances in Artificial Intelligence: 25th Canadian*

Conference on Artificial Intelligence, Canadian AI 2012, Toronto, ON, Canada, May 28-30, 2012. Proceedings 25, pages 61–73. Springer, 2012.

[32] Luca Giamattei, Antonio Guerriero, Roberto Pietrantuono, and Stefano Russo. Causal reasoning in software quality assurance: A systematic review. *Information and Software Technology*, 178:107599, 2025.

[33] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

[34] Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)*, 53(4):1–37, 2020.

[35] Shivani Gupta and Sangeeta Bansal. Does a rise in bmi cause an increased risk of diabetes?: Evidence from india. *PLoS One*, 15(4):e0229716, 2020.

[36] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1):389–422, 2002.

[37] Uzma Hasan and Md Osman Gani. Kcrl: A prior knowledge based causal discovery framework with reinforcement learning. In *Machine Learning for Healthcare Conference*, pages 691–714. PMLR, 2022.

[38] Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, 13(1):2409–2464, 2012.

[39] M. Á. Hernán, B. Brumback, and J. M. Robins. Marginal Structural Models to Estimate the Causal Effect of Zidovudine on the Survival of HIV-Positive Men. *Epidemiology*, 11(5), 2000.

[40] Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

[41] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*. John Wiley & Sons, 2013.

[42] Zhen Hu, Genevieve B Melton, Elliot G Arsoniadis, Yan Wang, Mary R Kwaan, and Gyorgy J Simon. Strategies for handling missing clinical data for automated surgical site infection detection from the electronic health record. *Journal of biomedical informatics*, 68:112–120, 2017.

[43] Qi Huang and Jianguo Tang. Age-related hearing loss or presbycusis. *European Archives of Oto-rhino-laryngology*, 267:1179–1191, 2010.

[44] G. W. Imbens. Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *The Review of Economics and Statistics*, 86(1):4–29, 02 2004.

[45] Guido W Imbens and Thomas Lemieux. Regression discontinuity designs: A guide to practice. *Journal of econometrics*, 142(2):615–635, 2008.

[46] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

[47] Padmavathi Janardhanan, L. Heena, and Fathima Sabika. Effectiveness of support vector machines in medical data mining. *Journal of Communications Software and Systems*, 2015.

[48] Thomas Jiralerspong, Xiaoyin Chen, Yash More, Vedant Shah, and Yoshua Bengio. Efficient causal graph discovery using large language models. *arXiv preprint arXiv:2402.01207*, 2024.

[49] Ulla Kampmann, Sine Knorr, Jens Fuglsang, and Per Ovesen. Determinants of maternal insulin resistance during pregnancy: an updated overview. *Journal of diabetes research*, 2019(1):5320156, 2019.

[50] Kensaku Kawamoto, Caitlin A Houlihan, E Andrew Balas, and David F Lobach. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *Bmj*, 330(7494):765, 2005.

[51] E. Kiciman and A. Sharma. Causal Reasoning: Fundamentals and Machine Learning Applications. https://causalinference.gitlab.io/, 2019.

[52] Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2023.

[53] Seonwoo Kim. Overview of clinical study designs. *Clinical and Experimental Emergency Medicine*, 11(1):33, 2023.

[54] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes, 2022.

[55] Adesh Kumar, Rajiv Gulati, Sangeeta Singhal, and et al. The effect of smoking on the hearing status-a hospital based study. *Journal of clinical and diagnostic research : JCDR*, 2013.

[56] Kamlesh Lakhwani, Sandeep Bhargava, Kamal Kant Hiran, Mahesh M Bundele, and Devendra Somwanshi. Prediction of the onset of diabetes using artificial neural network and pima indians diabetes dataset. In *2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, pages 1–6. IEEE, 2020.

[57] Thuc Duy Le, Tao Hoang, Jiuyong Li, Lin Liu, Huawen Liu, and Shu Hu. A fast pc algorithm for high dimensional causal discovery with multi-core pcs. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(5):1483–1495, 2016.

[58] Damodar Vishnu Lele. Occupational health surveillance. *Indian journal of occupational and environmental medicine*, 2018.

[59] Xiaoling Liang, Wenhao Song, Weibing Yang, and Zhenhua Yue. Enhancing diabetes risk assessment through bayesian networks: An in-depth study on the pima indian population. *Endocrine and Metabolic Science*, 17:100212, 2025.

[60] Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388, 2010.

[61] Chenxi Liu, Yongqiang Chen, Tongliang Liu, Mingming Gong, James Cheng, Bo Han, and Kun Zhang. Discovery of the hidden world with large language models. *Advances in Neural Information Processing Systems*, 37:102307–102365, 2024.

[62] Mingxuan Liu, Siqi Li, Han Yuan, Marcus Eng Hock Ong, Yilin Ning, Feng Xie, Seyed Ehsan Saffari, Yuqing Shang, Victor Volovici, Bibhas Chakraborty, et al. Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques. *Artificial Intelligence in Medicine*, page 102587, 2023.

[63] Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, page 6449–6459. Curran Associates Inc., 2017.

[64] Gilles Louppe, Louis Wehenkel, Antonio Sutera, and Pierre Geurts. Understanding variable importances in forests of randomized trees. *Advances in neural information processing systems*, 26, 2013.

[65] Christina Mack, Zhaohui Su, and Daniel Westreich. *Managing Missing Data in Patient Registries: Addendum to Registries for Evaluating Patient Outcomes: A User's Guide, Third Edition*. AHRQ Methods for Effective Health Care. Agency for Healthcare Research and Quality (US), Rockville (MD), 2018.

[66] Rheanna Mainzer, Margarita Moreno-Betancur, Cattram Nguyen, Julie Simpson, John Carlin, and Katherine Lee. Handling of missing data with

multiple imputation in observational studies that address causal questions: protocol for a scoping review. *BMJ open*, 13(2):e065576, 2023.

[67] D. Malinsky and D. Danks. Causal discovery algorithms: A practical guide. *Philosophy Compass*, 13(1), 2018.

[68] S. Mani and G. Cooper. Causal discovery from medical textual data. *Proceedings AMIA Symposium*, page 542, 12 2001.

[69] CJ Mann. Observational research methods. research design ii: cohort, cross sectional, and case-control studies. *Emergency medicine journal*, 20(1):54–60, 2003.

[70] Joris M. Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020.

[71] S. L. Morgan and C. Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Analytical Methods for Social Research. Cambridge University Press, 2 edition, 2014.

[72] Abdulazeez Mousa, Waraz Mustafa, Ridwan Boya Marqas, and Shivan HM Mohammed. A comparative study of diabetes detection using the pima indian diabetes database. *J. of Duhok University*, 26(2):277–288, 2023.

[73] Swapna Munnangi and Sameh W Boktor. Epidemiology of study design. 2017.

[74] Narmada Naik, Ayush Khandelwal, Mohit Joshi, Madhusudan Atre, Hollis Wright, Kavya Kannan, Scott Hill, Giridhar Mamidipudi, Ganapati Srinivasa, Carlo Bifulco, et al. Applying large language models for causal structure learning in non small cell lung cancer. In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, pages 688–693. IEEE, 2024.

[75] A. R. Nogueira, A. Pugnana, S. Ruggieri, D. Pedreschi, and J. Gama. Methods and tools for causal discovery and causal inference. *WIREs Data Mining and Knowledge Discovery*, 12(2):e1449, 2022.

[76] Marlies Noordzij, Friedo W Dekker, Carmine Zoccali, and Kitty J Jager. Measures of disease frequency: prevalence and incidence. *Nephron Clinical Practice*, 115(1):c17–c20, 2010.

[77] National Institute of Diabetes, Digestive, and Kidney Diseases. Pima indians diabetes database. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

[78] Chollette C Olisah, Lyndon Smith, and Melvyn Smith. Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Computer Methods and Programs in Biomedicine*, 220:106773, 2022.

[79] J. Pearl and D. Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., USA, 1st edition, 2018.

[80] Judea Pearl. *Causality*. Cambridge university press, 2009.

[81] Judea Pearl. Robustness of causal claims. *arXiv preprint arXiv:1207.4173*, 2012.

[82] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[83] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

[84] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT press, 2017.

[85] Karl Popper. *The logic of scientific discovery*. Routledge, 2005.

[86] Agnesa Preda, Dominic-Gabriel Iliescu, Alexandru Comănescu, George-Lucian Zorilă, Ionela Mihaela Vladu, Mircea-Cătălin Forțofoiu, Tiberiu Stefaniță Țenea Cojan, Silviu-Daniel Preda, Ileana-Diana Diaconu, Eugen Moța, et al. Gestational diabetes and preterm birth: what do we know? our experience and mini-review of the literature. *Journal of Clinical Medicine*, 12(14):4572, 2023.

[87] J. Ramsey, M. Glymour, R. Sanchez-Romero, and C. Glymour. A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International Journal of Data Science and Analytics*, 3, 2017.

[88] Thomas Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003.

[89] P. R. Rosenbaum and D. B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387):516–524, 1984.

[90] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[91] Wendy R Russell, Athanasia Baka, Inger Björck, Nathalie Delzenne, Dan Gao, Helen R Griffiths, Ellie Hadjilucas, Kristiina Juvonen, Sampo Lahtinen, Mirian Lansink, et al. Impact of diet composition on blood glucose regulation. *Critical reviews in food science and nutrition*, 56(4):541–590, 2016.

[92] Sara Saturni, Federico Bellini, Fulvio Braido, Pierluigi Paggiaro, Alessandro Sanduzzi, Nicola Scichilone, PA Santus, Luca Morandi, and Alberto Papi. Randomized controlled trials and real life studies. approaches and methodologies: a clinical point of view. *Pulmonary pharmacology & therapeutics*, 27(2):129–138, 2014.

[93] C Serra, MC Rodriguez, GL Delclos, M Plana, LI Gómez López, and FG Benavides. Criteria and methods used for the assessment of fitness for work: a systematic review. *Occup Environ Med.*, 64:304–12, 2007.

[94] K Shailaja, Banoth Seetharamulu, and MA Jabbar. Machine learning in healthcare: A review. In *2018 Second international conference on electronics, communication and aerospace technology (ICECA)*, pages 910–914. IEEE, 2018.

[95] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pages 3076–3085. PMLR, 2017.

[96] Amarish Kumar Sharma, Sanjeev Singh, Himanshu Singh, Deviyani Mahajan, Prachetha Kolli, Gowtham Mandadapu, Bimlesh Kumar, Dharmendra Kumar, Sudarshan Kumar, and Manoj Kumar Jena. Deep insight of the pathophysiology of gestational diabetes mellitus. *Cells*, 11(17):2672, 2022.

[97] Amit Sharma and Emre Kiciman. Dowhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*, 2020.

[98] Xinpeng Shen, Sisi Ma, Prashanthi Vemuri, M Regina Castro, Pedro J Caraballo, and Gyorgy J Simon. A novel method for causal structure discovery from ehr data and its application to type-2 diabetes mellitus. *Scientific reports*, 11(1):21025, 2021.

[99] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

[100] Shohei Shimizu, Aapo Hyvärinen, Patrik O Hoyer, and Yutaka Kano. Finding a causal ordering via independent component analysis. *Computational Statistics & Data Analysis*, 50(11):3278–3293, 2006.

[101] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvarinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, Kenneth Bollen, and Patrik Hoyer. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research-JMLR*, 12(Apr):1225–1248, 2011.

[102] Yan-Yan Song and Ying Lu. Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry.*, 2015.

[103] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

[104] Steven Tenny and Mary R Hoffman. Relative risk. 2017.

[105] Catherine Thomsen, Jacquelyn McClain, Kenneth Rosenman, and Letitia Davis. Indicators for occupational health surveillance. *Centers for Disease Control and Prevention*, 2007.

[106] SMTI Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 6, 2024.

[107] I. Tsamardinos, L. E. Brown, and C. F. Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, Oct 2006.

[108] Jan Van den Broeck, Solveig Argeseanu Cunningham, Roger Eeckels, and Kobus Herbst. Data cleaning: detecting, diagnosing, and editing data abnormalities. *PLoS medicine*, 2(10):e267, 2005.

[109] W. M. van der Wal and R. B. Geskus. ipw: An R Package for Inverse Probability Weighting. *Journal of Statistical Software*, 43(13):1–23, 2011.

[110] Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth N Balasubramanian, and Amit Sharma. Causal inference using llm-guided discovery. *arXiv preprint arXiv:2310.15117*, 2023.

[111] Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth N Balasubramanian, and Amit Sharma. Causal order: The key to leveraging imperfect experts in causal inference. 2025.

[112] Brian G Vegetabile. On the distinction between" conditional average treatment effects"(cate) and" individual treatment effects"(ite) under ignorability assumptions. *arXiv preprint arXiv:2108.04939*, 2021.

[113] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[114] Lei Wang, Shanshan Huang, Shu Wang, Jun Liao, Tingpeng Li, and Li Liu. A survey of causal discovery based on functional causal model. *Engineering Applications of Artificial Intelligence*, 133:108258, 2024.

[115] Yan Xie, Benjamin Bowe, Andrew K Gibson, Janet B McGill, Geetha Maddukuri, Yan Yan, and Ziyad Al-Aly. Comparative effectiveness of sglt2 inhibitors, glp-1 receptor agonists, dpp-4 inhibitors, and sulfonylureas on risk of kidney outcomes: emulation of a target trial using health care databases. *Diabetes care*, 43(11):2859–2869, 2020.

[116] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–46, 2021.

[117] Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: DAG Structure Learning with Graph Neural Networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 7154–7163. PMLR, 2019.

[118] Jean Yudin and Bruce Kinosian. *Home-Based Medical Care for Older Adults: A Clinical Case Book*, chapter Morbid Obesity, pages 119–124. Springer International Publishing, 2020.

[119] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schoelkopf. Kernel-based conditional independence test and application in causal discovery, 2012.

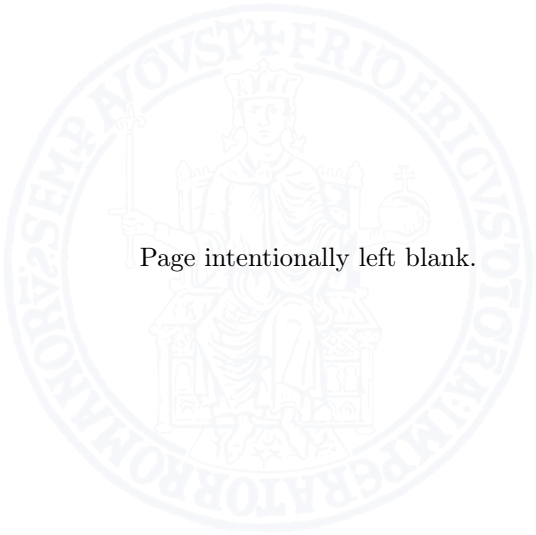
[120] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, pages 1–46, 2025.

[121] X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing. DAGs with NO TEARS: Continuous Optimization for Structure Learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 1–12. Curran Associates, Inc., 2018.

[122] Youran Zhou, Sunil Aryal, and M Reda Bouadjenek. A comprehensive review of handling missing data: Exploring special missing mechanisms. *arXiv preprint arXiv:2404.04905*, 2024.

Author's publications

1. P. Quaranta, R. Pietrantuono and G. Cascone, "Correlation and Causal Analysis of Occupational Health Data in the Maritime Domain," 2025 IEEE 13th International Conference on Healthcare Informatics (ICHI), Rende, Italy, 2025, pp. 739-744, doi: 10.1109/ICHI64645.2025.00112.
2. P. Quaranta and R. Pietrantuono, "Exploring Causal Modeling to Enhance Diabetes Prediction and Management," 2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS), Madrid, Spain, 2025, pp. 725-726, doi: 10.1109/CBMS65348.2025.00150.



Page intentionally left blank.