ID N. AIDR02/18790

UNIVERSITÀ CAMPUS BIO-MEDICO DI ROMA

DEPARTMENT OF ENGINEERING

CONSIGLIO NAZIONALE DELLE RICERCHE

ISTITUTO DI ELETTRONICA E DI INGEGNERIA DELL'INFORMAZIONE E DELLE TELECOMUNICAZIONI

Italian National Ph.D. in Artificial Intelligence
Health and Life Sciences
XXXVIII Cycle

Counterfactual-based Recommendations for Prevention of High-Prevalence Chronic Diseases

Supervisors
Dr. Alessia Paglialonga
Dr. Maurizio Mongelli

Candidate
Marta Lenatti

Feb, 2026

"Perché mai noi umani dovremmo affidarci a macchine morali?"

— F. Varanini —

A Bianco... e a chi come lui

Abstract

Chronic diseases, among the leading causes of death and disability worldwide, are characterized by persistent symptoms that require long-term care. Effective prevention through the management of modifiable risk factors (e.g., lifestyle habits) is key to mitigate the associated health and socioeconomic burden. Artificial Intelligence (AI) predictive models trained on large sets of clinical data can support prevention efforts by identifying individuals at risk and highlighting patterns associated with disease onset and progression.

Explainable AI (XAI) techniques have shown considerable potential in supporting the development of trustworthy predictive models by elucidating the internal mechanisms of their decision-making process. However, the integration of XAI-based medical decision support systems into clinical practice remains limited, as existing approaches often fail to generate explanations that are genuinely interpretable by humans, actionable, and clinically feasible. Counterfactual explanations, a class of purely data-driven, local, post-hoc XAI methods, offer a promising way to address these limitations by generating "what-if" scenarios that illustrate how changes in input features could alter model predictions. Nevertheless, this technique solely relies on associative patterns in data, which may reflect spurious relationships, resulting in recommendations that are impractical or infeasible in real-world settings. Incorporating causal reasoning in medical decision making can help uncover underlying disease mechanisms and produce more reliable and actionable insights. In this regards, causal learning techniques ca be used to estimate the efficacy of hypothetical lifestyle interventions (e.g., improvements in physical activity and diet) by explicitly representing cause–effect relationships, thereby supporting the design of personalized preventive strategies.

This Thesis project investigates advanced methodological frameworks using data-driven and domain expert-driven knowledge for developing predictive models aimed at chronic disease prevention. Counterfactual analysis, combining both purely data-driven XAI methods and causal learning approaches, is applied to two chronic disease prevention case studies (type 2 diabetes and cardiovascular disease) leveraging large-scale routinely collected data from primary care electronic health records and disease progression models. Constraints on feature mutability and adherence to a causal structure are imposed to promote actionability

and clinical feasibility of the outcomes, while also maintaining compliance with quantitative performance metrics. The results demonstrate that the proposed methodologies enhance model transparency while producing clinically relevant insights, illustrating how targeted modifications of modifiable risk factors can improve patient outcomes, support personalized care, and reveal variability in benefits across patient groups.

Contents

List of Figures

List of Tables

List of Algorithms

List of Acronyms

Acronym	Full name
CE	Counterfactual Explanation
CEILS	Counterfactual Explanations as Interventions in Latent Space
COPD	Chronic Obstructive Pulmonary Disease
CPCSSN	Canadian Primary Care Sentinel Surveillance Network
CPT	Conditional Probability Table
CVD	Cardiovascular Disease
dBP	diastolic Blood Pressure
DiCE	Diverse Counterfactual Explanations
DAG	Directed Acyclic Graph
DPP	Diabetes Prevention Program
EHR	Electronic Health Record
FPG	Fasting Plasma Glucose
HbA1c	Glycated hemoglobin
FRS	Framingham Risk Score
HDL	High-Density Lipoproteins
IDF	International Diabetes Federation
LDL	Low-Density Lipoproteins
MC-SVDD	Multi-Class Support Vector Data Description
MC-SVM	Multi-Class Support Vector Machine
ML	Machine learning
MUCH	MUlti-class Counterfactual explanations via Halton Sampling
RCT	Randomized Controlled Trial
sBP	sistolic Blood Pressure

SCM	Structural Causal Model
SHAP	Shapley Additive explanations
SVM	Support Vector Machine
T2D	Type 2 diabetes mellitus
WHO	World Health Organization
XAI	Explainable Artificial Intelligence

Chapter 1

Introduction

1.1 The Burden of Chronic Diseases

Chronic diseases, often referred to as noncommunicable diseases, are illnesses characterized by persistent symptoms, therefore requiring long-term control [1]. The four primary categories of chronic diseases, including cardiovascular diseases, cancer, chronic respiratory diseases, and diabetes, are the leading causes of death and disability worldwide. Indeed, the impact of chronic diseases, driven by population growth and aging, grew from causing 61% of global deaths in 2000 to causing 74% of global deaths in 2019 and 75% of non-pandemic-related deaths in 2021 [2]. In addition to such determinants, chronic diseases are the result of a combination of individual risk factors including [3]:

- behavioral factors (leading risk factors) such as unhealthy diets, alcohol abuse, insufficient physical activity, and tobacco use;
- metabolic factors such as elevated blood pressure (the leading metabolic risk factor globally, to which 25% of global chronic disease deaths are attributed), obesity, high blood glucose, abnormal blood lipids;
- environmental factors such as indoor and outdoor air pollution and exposure to chemical hazards.

Globally, a person aged 30 years in 2019 had a 17.8% (13.3–23.1% uncertainty interval) chance of dying from one of the aforementioned primary chronic diseases before the age of 70 years [2]. Although the annual premature mortality rate is declining, regional disparities remain, and the overall pace of reduction is still slow (Figure 1.1). If the current average annual reduction rate in premature mortality from chronic diseases persists, no region will

meet the Sustainable Development Goals (SDGs) endorsed by the WHO by 2030, i.e., one-third reduction in mortality from chronic diseases through prevention and treatment [4].

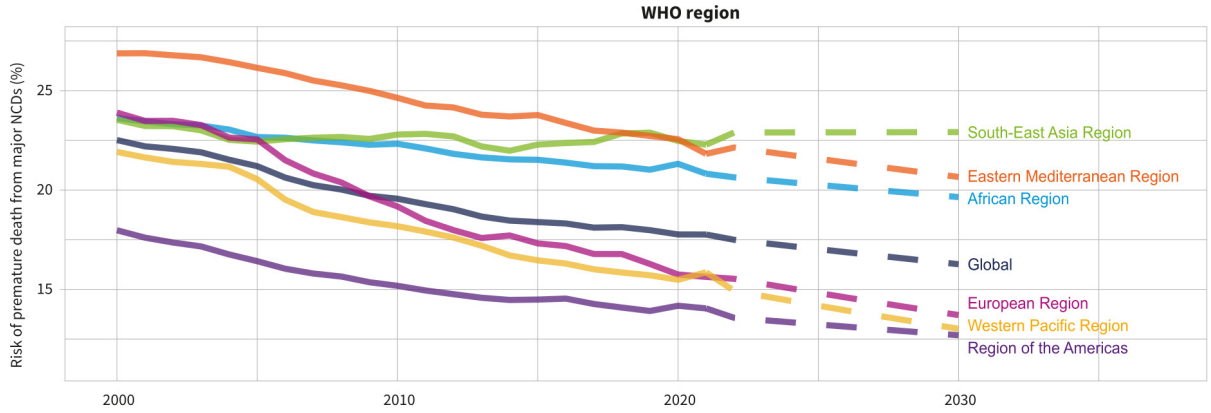


Figure 1.1: Probability of dying from the four primary chronic diseases (cardiovascular diseases, cancers, chronic respiratory diseases and diabetes) at premature ages (30–69 years). Source: [4]. Solid lines: past trends; dashed lines: projected trends up to 2030.

Chronic diseases not only pose a threat to individual health but also have impact on the community due to their socioeconomic burden. The healthcare system in the US, for instance, faces costs of around one trillion dollars per year due to chronic diseases [5]. Hence, prioritizing prevention and management policies is of paramount importance. Aside from specific genetic disorders and environmental exposure, which are considered non-modifiable risk factors, behavioral or lifestyle choices, known as modifiable risk factors, generally play a key role in either improving or worsening health outcomes [1]. Therefore, a logical initial step is to foster prevention and management of chronic diseases by targeting individual behaviors, since what a person does or fails to do often has a major influence on the onset of such conditions [5]. Medical professionals play a key role in supporting these behavioral changes and proposing personalized strategies informed by both their expertise and established medical guidelines [1, 6].

Recently, there has been growing interest in developing Artificial Intelligence (AI) systems for the prediction and management of chronic diseases, aiming to improve diagnosis, therapy, and patient care. AI systems, by identifying intricate patterns within large volumes of data -such as health records, patient histories, genetic information, and diagnostic imaging- can in principle support prevention, by promoting timely interventions, and a reduction in complications [7, 8]. Furthermore, these systems hold considerable promise for enhancing precision medicine, moving beyond the traditional "one-size-fits-all" approach by creating individualized treatment plans, assisting clinicians in tailoring therapies to specific patients' characteristics and needs [9].

1.2 Explainable Artificial Intelligence

Artificial Intelligence (AI) has become deeply embedded in our daily lives, driven by the need to analyze vast amounts of data for the purpose of knowledge discovery. Understanding and trusting the decisions made by AI systems is of paramount importance, particularly in sensitive and safety-critical areas such as healthcare. Opacity of medical decision support systems can create hesitation in adopting AI technologies, ultimately undermining trust [10]. Indeed, stakeholders (e.g., clinicians, patients, etc...) are understandably reluctant to accept the outcomes of AI systems without insight into how those conclusions were reached [11]. To address these issues, eXplainable AI (XAI) has emerged as a new research field. It aims to provide humans with explanations that help them understand the reasoning behind an AI system's decision-making process, overcoming the limitations of black-box models. In other words, the goal of XAI is to allow end-users to understand the underlying explanatory factors of why an AI decision is taken [12]. Several motivations for the use of XAI can be identified [13, 10]:

- *explain to justify*: explanations may serve to justify decisions made by AI systems, ensuring that these decisions are fair, ethical, and defensible, especially when outcomes are unexpected.
- *explain to control*: explanations may serve to help detect vulnerabilities in the AI system, preventing errors through better transparency and debugging.
- *explain to improve*: explanations may facilitate the continuous improvement of AI systems, as insights into how outputs are generated allow for more effective refinement.
- *explain to discover*: explanations may serve as a tool for gaining new knowledge on relationships and patterns in the data, uncovering new insights about the problem domain.

XAI offers clear rationales for AI decisions, enhancing trust and facilitating adherence to recently developed regulatory standards [14, 15, 16, 17]. Among these, the EU AI Act [17], entered into force on 1 August 2024 as the first-ever comprehensive legal framework on AI. Such regulation poses emphasis on the critical role of explainability and transparency, especially in high risk AI systems, requesting the need to openly share details about data sources, algorithms, and decision-making processes to address bias and foster accountability.

Over the past decade, several review articles have been published in the field of AI explainability and interpretability [18, 10, 19, 20], with a few focusing specifically on applications, emerging trends and challenges in the medical field [21, 22]. Such works highlighted

the marked increase in research publications from 2019 onwards, mainly driven by the publication of the European General Data Protection Regulation (GDPR)[14] and its demand for increased transparency when treating sensitive data.

Hakkoum et al. [22] selected 179 articles investigating AI interpretability in medicine published from August 1994 to December 2020 from six digital libraries (ScienceDirect, IEEE Xplore, ACM Digital Library, SpringerLink, Wiley, and Google Scholar). The focus of this work was directed toward research and review articles that investigated interpretability techniques applied to medical tasks, proposed novel interpretability methods within medical contexts, or conducted comparative analyses of existing interpretability approaches in such settings. In terms of medical task, most of the selected studies (75%) focused on diagnosis, with only a few studies focusing on prognosis, treatment, management, screening and monitoring. In terms of medical disciplines, oncology, endocrinology and cardiology were the most frequently addressed. Classification was the most used objective (92%), and the most interpreted black-box models were artificial neural networks (70%) followed by Support Vector Machines (10%). According to this review, the majority of studies up to 2020 concentrated on numerical and categorical data, whether sourced from tabular datasets or radiomic features, followed by images, while only a small portion examined time series and text. A more recent review by Frasca et al. [21] queried articles published between January 2013 and October 2023, focusing on the analysis on 488 articles specifically related to interpretability and explainability of AI algorithms in the medical area that were indexed in both Scopus and Web of Science. This review article identified a shift in the literature with respect to [22]: more recent work was predominantly focused on medical imaging analysis in the context of the pandemic, while the emergence of large language models has directed increasing attention toward natural language processing. More specifically, based on keyword occurrence, most articles focused on deep learning, particularly convolutional neural networks for medical image analysis, followed by natural language processing techniques. In terms of medical disciplines and pathologies, the main focus areas were COVID-19, Alzheimer's disease, cardiology, and oncology (brain and breast cancer) [21].

The extensive development of XAI techniques has paved the way for data-driven clinical decision support systems that aim to incorporate transparency into automated decision pathways, beyond maximizing performance metrics [22]. Nevertheless, there are several challenges to the practical implementation of these tools, particularly in relation to concerns around privacy, scalability, fairness and accountability, which have the potential to erode the trustworthiness of the system [23, 21]. Moreover, several of the currently available XAI techniques struggle to produce explanations that are interpretable in human terms and that can be used to provide readily applicable and actionable interventions [24].

1.3 Causal learning

Most existing AI systems are either descriptive or predictive in nature, focusing on estimating current or future outcomes based on available data without accounting for potential interventions or changes. These *predictive approaches* essentially act as risk scoring functions. When considering chronic diseases, for instance, they can be used to estimate the probability of disease diagnosis to identify patients who are at elevated risk [25]. However, advancing preventive precision medicine requires *interventional approaches* capable of selecting and personalizing the treatment most likely to yield a favorable outcome, i.e., approaches that are truly actionable and can support clinical decision-making [26]. Indeed, interventional approaches allow us to answer different kind of questions, such as to assess how modifying a subject's lifestyle would impact their future risk of developing a chronic disease.

Traditional predictive models merely use data to approximate a function mapping input to the output of interest, reflecting associations that do not necessarily hold a causal meaning [27]. In contrast, interventional approaches necessitate an explicit consideration of the underlying causal structure. The gold standard for testing causal hypotheses in clinical research is the Randomized Controlled Trial (RCT). By design, well-conducted RCTs evenly distribute both known and unknown patient factors across treatment and control groups, thereby minimizing bias arising from confounding variables. Therefore, under conditions of proper randomization and full compliance, differences in outcomes between groups can be causally attributed to the intervention [28]. However, conducting RCTs is often impractical due to financial, temporal, and ethical constraints. Also, their external validity can be limited, as results may not generalize beyond the study population as a consequence of restrictive inclusion criteria that fail to capture real-world diversity [28, 29].

Causal learning provides a set of tools for simulating interventions from data, overcoming the challenges related to the implementation of RCTs [30, 31]. The scope of causal learning can be broadly categorized into two main tasks: learning causal structures from data (causal discovery) and estimating causal quantities from data (causal inference) [32]. Real-world observational data, such as Electronic Health Records (EHRs), are a promising source of data for causal learning as they offer large-scale, diverse, and accessible data across a broad range of diseases and patient populations. For such reasons, causal learning from observational data has recently become an active area of research [33, 28, 34, 35] with substantial potential in the management and evaluation of chronic diseases; practical applications can be found in diabetes care [36, 37, 38, 39], oncology [29, 40, 41, 42], neurological and mental health care [43, 39]. From current literature, it is evident that a causal understanding of chronic disease onset is critical for translating predictive insights into actionable prevention.

However, interventional approaches (e.g., treatment effect estimation) based on causal learning from observational data introduce challenges that do not arise in traditional predictive models. These difficulties stem from the fundamental problem of causal inference: for any given individual, we can only ever observe the outcome under the treatment they actually received (factual). The outcome under any alternative treatment (counterfactual) is inherently unobservable and therefore hypothetical [31]. Therefore, inferring causal relationships from observational data requires strong methodological assumptions [34]. Unlike predictive models, which rely solely on observed patterns to forecast outcomes, causal inference models must estimate these hypothetical alternatives, making the task substantially more difficult both to perform and to evaluate. Furthermore, observational data, which are generally less accepted by regulatory bodies than RCT data, are affected by issues such as missingness, selection bias, and unmeasured confounders [44].

1.4 Research gap and main objectives

XAI techniques have shown promising results in supporting clinical decision-making by unravelling the inner mechanisms of predictive models [22, 21]. Despite so, the adoption of XAI-based medical decision support systems in clinical practice is hampered by the fact that these solutions may fail to produce explanations that are actually interpretable in human terms (e.g., as they often provide explanations in terms of complex and multi-dimensional representations [45]) and actionable [24].

Counterfactual explanations [46] help address this issue by producing counterfactual samples (i.e., "what-if" scenarios) that explain a model in terms of how modifications to the input data would result in a different model prediction. An example of a counterfactual explanation in the context of chronic disease prevention models could be the following: "if systolic blood pressure had been reduced from 150 mmHg to 125 mmHg, and total cholesterol from 6.8 mmol/L to 4.8 mmol/L, the model would have predicted a low cardiovascular risk instead of a high-risk classification". Such "contrary to-fact" explanations are more aligned with human reasoning with respect to those provided by other XAI techniques like for example feature importance, making them especially suitable for decision-support or recommendation tasks [47]. However, these explanations usually rely on extracting associative relationships from data, potentially leading to unfeasible decisions in practice (e.g., excessive reduction in blood biomarkers) [48, 12, 49]. The study of causation beyond association may be helpful to uncover data patterns and shed light on the mechanisms of chronic disease onset and development [37]. In particular, causal learning techniques are valuable tools for modeling how different hypothetical actions can produce varied outcomes by following a chain

of cause-and-effect [50, 31]. Thus, they can be effective in guiding the design of personalized prevention strategies through actionable interventions [26, 34]. An emerging body of research seeks to empower XAI by integrating knowledge of the causal relationships underlying the phenomenon of interest, using insights from causal learning to inform the generation of explanations [12].

Currently, there is still a lack of explainable and causally grounded AI approaches for chronic disease prevention that target individualized changes in modifiable risk factors. This PhD project focuses on the investigation of XAI and causal learning approaches applied on observational data from EHR and data from disease progression models to identify individuals and stratified groups of subjects at risk of developing chronic diseases and provide *data-driven recommendations in the form of counterfactual samples*, for lowering this risk, leveraging data and domain knowledge.

From a methodological point of view, the Project consists of three objectives. *OBJ 1* involves the analysis of the XAI declination of counterfactual samples, investigating purely associative counterfactual explanations. The focus is on generating explanations that suggest minimal changes to modifiable features (e.g., blood biomarkers), ensuring that the suggested changes are both feasible and capable of altering model predictions. *OBJ 2* leverages causal learning techniques to perform counterfactual inference for intervention design. It places specific emphasis on using counterfactual analysis to guide lifestyle modification interventions that identify targeted, causally grounded changes in modifiable risk factors to prevent disease onset. Together, these objectives advance methodologies that integrate interpretability, causality, and actionable minimal-change interventions, rather than merely analyzing existing techniques. The final overarching objective *OBJ 3* is to integrate these two classes of methods, investigating the interplay between XAI and causality, with a special emphasis on techniques that integrate causal constraints into the generation of counterfactual explanations.

Observational data for two case studies, each addressing a different chronic clinical condition, were curated as part of the Project: cardiovascular diseases - CVDs (*Case Study 1*) and Type 2 diabetes – T2D (*Case Study 2*). Analyses were performed on large original datasets of routinely collected measures extracted from the 2000-2015 portion of the Canadian Primary Care Sentinel Surveillance Network (CPCSSN,[51]), a nation-wide database of de-identified primary care EHRs. Additional data for *Case Study 2* included simulations from a computationally efficient version of a dynamical model of Type 2 diabetes progression [52].

1.5 Thesis organization

This Chapter introduced the AI landscape that frames my doctoral research. The remainder of the Thesis is structured as follows:

- Chapter 2 provides fundamental notions on the state of the art in XAI and causal learning, with particular emphasis on defining essential concepts and the mathematical notation adopted throughout the Thesis;
- Chapter 3 describes a novel method for generating counterfactual explanations from Support Vector Data Descriptors [53] in multi-class classification tasks and its validation on a real-world dataset drawn from observational EHR data (*Case Study 1*), using both standard and newly developed evaluation metrics;
- Chapter 4 presents a causal learning framework for observational EHR data (*Case Study 2*) that combines causal discovery with the integration of prior expert knowledge and applies causal inference to estimate the impact of hypothetical lifestyle interventions on the onset of the chronic disease, in presence of latent variables;
- Chapter 5 introduces a novel framework for counterfactual inference in dynamic settings to assess the effect of hypothetical lifestyle interventions on the onset of the chronic disease, based on a large set of simulations generated with a dynamic model of chronic disease progression (*Case Study 2*);
- Chapter 6 presents an exploratory study on the application of methods integrating causal knowledge during the generation of counterfactual explanations for chronic disease prevention, using observational EHR data (*Case Study 2*);
- Chapter 7 concludes the dissertation by providing final remarks and discussing future directions of research.

  Appendices. Appendix A provides a more in-depth extension of the analyses presented in Chapter 3, evaluating the proposed method across three benchmark datasets. Appendix B contains a brief description of the CPCSSN [51], a nation-wide database of EHRs, which I had the opportunity to access during the project via a visiting researcher agreement with Toronto Metropolitan University. Appendix C summarizes the key contributions of additional research activities undertaken during my doctoral studies at CNR-IEIIT, which, although not directly related to the core research of this Thesis, provided important experience and original results. Finally, Appendix D includes a comprehensive list of my Journal and Conference publications.

Chapter 2

Background

This chapter provides an overview of the methodological background of the two primary AI fields investigated during my PhD, with the goal of developing chronic disease prevention models: XAI (Section 2.1) and causal learning (Section 2.2).

2.1 Explainable Artificial Intelligence (XAI)

2.1.1 Overview

XAI encompasses a variety of techniques designed to make the decision-making processes of AI systems more transparent. Despite its widespread adoption, a consensus among researchers on several XAI-related definitions has yet to be reached. Several taxonomies have been suggested to systematically categorize XAI approaches and clarify the terminology. Speith et al. [19] suggest a unified taxonomy that integrates existing classification systems (e.g., [18, 20, 54, 55, 56]). In such a combined taxonomy, XAI approaches are classified in a structured way based on various criteria, namely, stage, scope, applicability, result, functioning, and output format of the explanation.

In terms of *stage*, XAI approaches are often divided into ante-hoc or post-hoc methods. Ante-hoc approaches are related to AI systems that inherently enable a direct comprehension of their operational mechanisms, hence transparent-by-design or white-box models (e.g., decision trees, rule-based models, linear regression models). In contrast, post-hoc approaches deal with the extraction of explanatory information from existing black-box systems such as deep neural networks, random forests, and support vector machines. Therefore, post-hoc approaches are applied after the AI system has been trained to clarify its decision-making processes. Examples of post-hoc approaches include Permutation Feature Importance, Shapley Additive Explanations (SHAP, [57]), Local Interpretable Model-agnostic

Explanations (LIME, [58]), Partial Dependence Plots (PDP) and individual conditional expectation (ICE) [59].

Based on *applicability*, XAI approaches can be classified into model-specific or model-agnostic. Model-specific approaches provide explanations for a specific class of AI models. For example, gradient-based methods like Grad-CAM [60] can be applied to differentiable models only, while some methods such as treeSHAP [61] can be applied to tree-based AI models only. In contrast, model-agnostic approaches, like LIME and SHAP, can be applied to any AI model.

In terms of provided *result*, XAI approaches can be classified into feature importance scores, surrogate models, or examples [62]. Feature importance approaches focus on uncovering how input features influence an outcome. Surrogate models aim to approximate the original model (or a specific part of such model) with a simpler, ante-hoc explainable one. Finally, some approaches (e.g., endogenous counterfactual explanations) try to explain a model by providing representative examples of instances with their relative predicted output. It is worth noting that the taxonomy proposed by [19] do not assume mutually exclusive categories. For instance, LIME provides a feature importance score by leveraging a surrogate model.

In terms of *functioning* [56], that is, the way an XAI approach extracts information from the AI system, we can distinguish methods based on local perturbations, methods leveraging structures, meta-explanations, and architecture modifications. XAI approaches that rely on local perturbations, like LIME, slightly modify the inputs of a model to determine the impact of these inputs on the model's predictions. Methods leveraging the structure exploit specific properties of the AI system they are supposed to explain to construct the explanation (e.g., gradient-based methods), while methods based on architecture modification try to simplify and explain complex models by altering their architecture. Meta-explanations build upon explanations produced by different explainability techniques and combine them with the aim of providing a better explanation than each of the used methods individually. Finally, the functioning-based categorization also includes examples, aligning with the example category identified in the result-based categorization.

In terms of *scope*, XAI approaches can be referred to as local or global. Local approaches (e.g., counterfactual explanations) provide explanations for a single model prediction, for example, examining the impact of the input features on the model prediction for a certain test sample. On the other hand, global approaches (e.g., permutation feature importance and PDP) provide general explanations of the whole logic of a AI system, for example, examining the impact of the input features on the model decisions for all test samples.

Another noteworthy dimension in the taxonomy is the *output format* of the explana-

tions, that is, whether they provide, for example, numerical, rule-based, textual, visual (e.g., heatmaps), simplified models (e.g., in architecture modifications) or mixed-types explanations.

Lastly, XAI methods can vary depending on the underlying *type of problem*, i.e., whether it is classification or regression, and the *type of input data* (tabular data, text, images, time series).

In this thesis, XAI approaches for tabular input data are examined, with a focus on counterfactual explanations. According to the taxonomy presented in this section, counterfactual explanations can be categorized as an example-based, local, post-hoc XAI technique, model-specific or model-agnostic, depending on their generation process. While most XAI approaches focus on explaining *why* a black-box model predicted a particular outcome, counterfactual explanations address this issue differently by helping stakeholders understand *which* input features would need to be modified to obtain a desired outcome, thus addressing "what-if" scenarios [48]. The following section provides an overview of counterfactual explanations, focusing on main definitions, desired properties, and examples of techniques that characterize the current state of the art.

2.1.2 Counterfactual explanations

Counterfactual explanations (CEs), first introduced by Wachter et al. [46], provide explanations of an AI system by specifying the modifications necessary in the input to produce a desired target prediction. For instance, consider an individual whose diabetes onset has been predicted by an AI system. A counterfactual explanation offers a conditional (what-if) statement that specifies the changes in the individual's characteristics or behavior that could have resulted in a different predicted outcome, e.g.: "If you had exercised for 30 minutes daily and reduced your sugar intake, the AI system would have predicted a lower risk of diabetes".

CEs generation methods, or counterfactual explainers, may be designed to handle different input data like tabular data, images or text and may deliver explanations in different forms including numerical values, regions of pixels, and linguistic expressions, as remarked in a recent survey [63].

Definition 2.1.1. (*Counterfactual explanations*) Given a black-box classification or regression model B that provides a certain output $y = B(\boldsymbol{x})$ for an instance $\boldsymbol{x}$, called factual, a counterfactual explanation consists of an instance $\boldsymbol{x}^{cf}$ such that the decision for B on $\boldsymbol{x}^{cf}$ is different from y, i.e., $B(\boldsymbol{x}^{cf}) \neq y$, and such that the difference between $\boldsymbol{x}$ and $\boldsymbol{x}^{cf}$ is minimal. Minimality is measured through a certain distance function, $\text{dist}(\boldsymbol{x}, \boldsymbol{x}^{cf})$.

Definition 2.1.2. (*Counterfactual explainer*) A counterfactual explainer is a function $\phi(\cdot)$ that takes as input a model B and a given instance of interest $\boldsymbol{x}$, and with its application $C = \phi(\boldsymbol{x}, B)$ returns a set $C = \{\boldsymbol{x}^{cf'}, \boldsymbol{x}^{cf''}, \cdots, \boldsymbol{x}^{cf^h}\}$ of $h \leq k$ counterfactual samples where k is the number of CEs required. Often k is set equal to 1.

The distinction between counterfactual explainers lies in how they retrieve $\boldsymbol{x}^{cf}$ [63]. The majority of counterfactual explainers found in the literature generate CEs by *addressing an optimization problem*, which is typically formulated by defining a loss function that aims to ensure a set of desired properties. Other counterfactual explainers adopt *heuristic search strategies* to generate CEs where the solution is iteratively updated with the objective of minimizing a cost function based on heuristic, greedy choices designed to ensure desired properties, allowing for faster computation compared to exact optimization. Explainers based on optimization strategies or heuristic search are often referred to as exogenous explainers because they obtain CEs through interpolation and/or random data generation, without guaranteeing the presence of naturally occurring feature values. In contrast, instance-based counterfactual explainers retrieve CEs by *selecting examples from a given dataset*, and thus are usually referred to as endogenous explainers. Samples generated with such methods naturally ensure realistic and feasible CEs, as they are derived from a real reference group.

Counterfactual explainers should be ideally designed to satisfy a set of properties [64, 63, 65, 66], including:

- *Availability.* A counterfactual explainer should return k CEs for each factual sample $\boldsymbol{x}$;
- *Validity.* A CE should change the model outcome with respect to the factual one, i.e., $B(\boldsymbol{x}^{cf}) \neq B(\boldsymbol{x})$;
- *Proximity (Similarity).* In order to be feasible in practice, a CE should be similar to (i.e., close to) the factual observation . Given a distance function $\text{dist}(\cdot, \cdot)$, the distance between $\boldsymbol{x}$ and $\boldsymbol{x}^{cf}$ should be as small as possible, i.e., $\text{dist}(\boldsymbol{x}, \boldsymbol{x}^{cf}) < \epsilon$, where ϵ is a predefined maximum distance threshold;
- *Actionability.* In order to be feasible in practice, a CE should never change immutable features (e.g., age). All differences between $\boldsymbol{x}$ and $\boldsymbol{x}^{cf}$ should refer to modifiable features only;
- *Sparsity (Minimality).* In order to be easily interpreted, a CE should change the predicted outcome by acting on a small number of features. There should not be any other valid counterfactual $\hat{x}^{cf}$ such that the number of different feature values between $\boldsymbol{x}$ and $\boldsymbol{x}^{cf}$ is higher than the number of different feature values between $\boldsymbol{x}$ and $\hat{x}^{cf}$;

- *Discriminative Power.* In order to be easily interpreted and implemented in practice, the end user should be able to distinguish a CE from its factual by simply looking at them;

- *Plausibility.* In order to be realistic, a CE should adhere to the observed correlations between features, that is, $\boldsymbol{x}^{cf}$ should not be labeled as an outlier with respect to the instances in the reference population of instances with output $B(\boldsymbol{x}^{cf})$. Usually, the training data distribution is used as the reference population.

- *Diversity.* If $h > 1$, the difference among all the CEs found for $\boldsymbol{x}$ in the retrieved set C should be maximized. Also, different factual observations should have different CEs, that is, $\boldsymbol{x}_1 \neq \boldsymbol{x}_2 \implies \boldsymbol{x}_1^{cf} \neq \boldsymbol{x}_2^{cf}$;

- *Causal feasibility.* In order to be feasible in practice, a CE must adhere to established cause and effect relationships between features. If one feature is modified, other dependent features may need to be adjusted accordingly to preserve the underlying mechanisms of the phenomenon.

These properties are widely consolidated in the literature as desiderata for explanation generation; however, standardized methods for their assessment remain lacking. Therefore, specific metrics will be defined on a case-by-case basis throughout this Thesis. Besides, an ideal CE should satisfy all the above-mentioned properties, however, optimizing the generation process for one property may require compromises in others [64]. For example, ensuring actionability and causal feasibility often comes at the cost of availability and proximity.

Presented below are some examples of counterfactual explainers, each characterized by their unique strategy for explanation retrieval and by the specific properties they aim to pursue. For example, Wachter's original formulation [46] relies on the minimization of a loss function $\mathcal{L}$ that incorporates two terms, a measure of the distance between the model prediction on the candidate CE and the desired outcome ($\mathcal{L}_{\text{validity}}$), as well as a measure of the distance between the original instance and the candidate CE ($\mathcal{L}_{\text{proximity}}$). That is,

$$\mathcal{L} = \lambda \mathcal{L}_{\text{validity}} + \mathcal{L}_{\text{proximity}} = \lambda(B(\boldsymbol{x}^{cf}) - y_{target}) + \text{dist}(\boldsymbol{x}, \boldsymbol{x}^{cf}),$$

where the hyperparameter λ balances the contribution of the two terms. Hence, the objective is to retrieve CEs that change the predicted class while remaining at minimum distance from the factual instance.

More complex terms may be incorporated into the process, for example, to promote diversity of the retrieved CEs [65], to maintain correlations between features [67], to foster

plausibility [68] or causal consistency [69]. For example, Diverse Counterfactual Explanation (DiCE, [65]) extends the loss function to search for k different CEs and adds an additional regularization term to the loss function ($\mathcal{L}_{\text{diversity}}$) that ensures variability in the set C of returned CEs, while keeping proximity with the factual, namely:

$$\mathcal{L}_{DICE} = \frac{1}{k}\sum_{i=1}^{k}\mathcal{L}_{\text{validity}} + \frac{\lambda_1}{k}\sum_{i=1}^{k}\mathcal{L}_{\text{proximity}} - \lambda_2\mathcal{L}_{\text{diversity}}(\boldsymbol{x}^{cf'}, \boldsymbol{x}^{cf''}, \cdots, \boldsymbol{x}^{cf^k})$$

where λ_1 and λ_2 balance the three parts of the loss function, and $\mathcal{L}_{\text{diversity}}$ penalizes CEs which are too similar.

The Contrastive Explanation Method (CEM, [68]) retrieves CEs as pertinent negatives as a result of optimization of the following loss function:

$$\mathcal{L} = \lambda\mathcal{L}_{\text{validity}} + \lambda_1\mathcal{L}_{\text{sparsity}} + \lambda_2\mathcal{L}_{\text{reconstruction}}$$

where $\mathcal{L}_{\text{validity}}$ is a validity term that encourages class flipping, $\mathcal{L}_{\text{sparsity}}$ is an elastic net regularizer combining $L1$ and $L2$ regularization to enhance sparsity, and $\mathcal{L}_{\text{reconstruction}}$ is an error in $L2$ reconstruction of $\boldsymbol{x}^{cf}$, evaluated by an autoencoder, encouraging plausibility with respect to data in the counterfactual class.

Mahajan et al. [69] proposes a causal proximity regularizer that can be used instead of the standard $\mathcal{L}_{\text{proximity}}$ of [46], defined as:

$$\mathcal{L}_{\text{proximity,causal}} = \sum_{\mathbf{u}\in\mathbf{U}}\text{dist}(\boldsymbol{x}_{\mathbf{u}}, \boldsymbol{x}^{cf}_{\mathbf{u}}) + \sum_{\mathbf{v}\in\mathbf{V}}\text{dist}(\boldsymbol{x}^{cf}_{\mathbf{v}}, \boldsymbol{x}^{cf}_{\mathbf{v},\mathbf{pred}})$$

where the first term is the standard proximity loss, i.e., ℓ^1 or ℓ^2 distance between factual observation and CEs, to be used for exogenous features $\mathbf{u}$ (features that are not caused by any other feature in $\boldsymbol{x}$), while the second term is a causally constrained proximity loss terms that considers the distance between the counterfactual value for each endogenous feature $\mathbf{v}$ and its expected value as predicted by a causal model. This way, CEs are constrained to be near the factual data sample not only based on the Euclidean distance between them, but also based on the known causal relationships.

A limitation of these approaches is that gradient-based methods are model-specific and limited to differentiable algorithms, such as deep neural networks. As an alternative approach, gradient-free optimization based on heuristic search strategies have been proposed in the context of non-differentiable models. An example of model-agnostic counterfactual explainer that exploits gradient-free optimization by leveraging quasi-random sampling of bounded classification region will be presented in Chapter 3. Another example of gradient-

free technique is provided by DiCE that also offers a model-agnostic framework implementing three distinct heuristic search strategies [70]: an independent random sampling of features, genetic algorithms, and K-dimensional trees. Random search involves sampling a finite set of points by randomly perturbing the initial observation $\boldsymbol{x}$ and selecting the perturbed sample that minimizes $\mathcal{L}_{DICE}$ while being classified in the target class. The genetic search, generates a set of candidate CEs through random perturbations. The candidates that minimizes $\mathcal{L}_{DICE}$ are selected and further processed through recombination and mutation to create new refined instances. Selection, recombination and mutation are repeated until convergence. Lastly, K-dimensional tree strategy consists of a space-partitioning data structure for efficient nearest neighbor search, enforcing diversity.

Despite the search that typically occurs in the original feature space, recent efforts have been made to recover counterfactual explanations in a transformed latent space, with optimized dimensions [71, 72]. For example, Crupi et al. [72] developed the Counterfactual Explanations as Interventions in Latent Space (CEILS) method, a framework that can be applied to any explainer for generating CEs that satisfy causal feasibility. The approach searches for CEs that lie at minimum distance from the factual within a constrained latent space that incorporates a set of causal dependencies, defined by a given causal model. An example illustrating the application of this method for chronic disease prevention models will be presented in Chapter 6.

2.2 Causal learning

2.2.1 Overview

Unlike traditional predictive models that rely on associative analysis to detect patterns or correlations between variables, causal learning is concerned with uncovering whether one factor actively produces changes in another, following a chain of cause and effect. This distinction is essential for making reliable predictions, guiding effective decision-making, and designing interventions.

In the domain of causal learning, two primary tasks can be identified: causal discovery (Section 2.2.2) and causal inference (Section 2.2.3). These two tasks address fundamentally different questions: causal discovery seeks to identify which factors determine the outcome, whereas causal inference aims to quantify the effect of a specific factor on the outcome [32]. In other words, causal discovery is the process of learning the causal structure from (observational) data, while causal inference is the process of quantifying causal effects, assuming that the causal structure is already known or has been inferred from the data.

Causal learning is largely centered around the representation of a phenomenon by means of a causal graph, typically a Directed Acyclic Graph (DAG). Presented below are some fundamental definitions related to the main graphical structures considered in this thesis (see Figure 2.1 for an example); a more in-depth discussion is available in [73, 74, 75].

Definition 2.2.1. *(Graph)* A graph $\mathcal{G} = (\mathbf{X}, \mathcal{E})$ is a mathematical object consisting of a set of nodes (or vertices) $\mathbf{X} = \{X_1, \ldots, X_n\}$ and a set of edges (or arcs) $\mathcal{E} \subseteq \mathbf{X} \times \mathbf{X}$ connecting them. A node X_i is a parent of X_j if $(X_i, X_j) \in \mathcal{E}$, i.e., if X_i is a direct cause of X_j. The set of parents of X_j is denoted by $Pa(X_j)$. A pair of nodes X_i, X_j is connected by a directed edge $X_i \to X_j$ if $(X_i, X_j) \in \mathcal{E}$ and $(X_j, X_i) \notin \mathcal{E}$ or by an undirected edge $X_i - X_j$ if $(X_i, X_j) \in \mathcal{E}$ and $(X_j, X_i) \in \mathcal{E}$. A cycle is a directed path $X_1, \cdots, X_k$ where $X_1 = X_k$.

Definition 2.2.2. *(Directed Acyclic Graph, DAG)* A graph $\mathcal{G}$ is a directed graph if all its edges are directed, that is, for each edge $(X_i, X_j) \in \mathcal{E}$, X_i is a direct cause of X_j and X_j is a direct effect of X_i. A directed graph is called a DAG if it contains no cycles.

Definition 2.2.3. *(Partially-directed acyclic graph, PDAG)*, A graph $\mathcal{G}$ is called a partially-directed acyclic graph (PDAG) if it can contain both undirected ($-$) and directed ($\to$) edges.

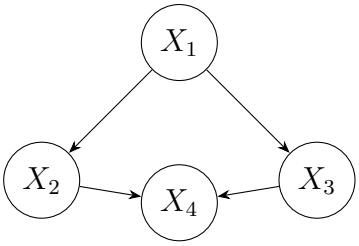
Definition 2.2.4. *(Skeleton)* The skeleton of $\mathcal{G}$ is the undirected graph formed by converting every directed edge in $\mathcal{G}$ into an undirected edge.

Definition 2.2.5. *(V-structure)* Given three nodes X_1, X_2, X_3, a v-structure is a graphical structure in which one node (the collider node) receives edges from two other nodes, e.g., $X_1 \to X_2 \leftarrow X_3$.

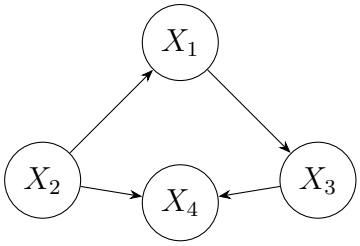
Definition 2.2.6. *(Observational Equivalence).* Two DAGs $\mathcal{G}$ and $\mathcal{H}$ are observationally Markov equivalent, i.e., $\mathcal{G} \equiv \mathcal{H}$, if they have the same skeleton and the same v-structures, therefore they cannot be distinguished purely from observational data. A Markov equivalent class represents the set of possible DAGs that are observationally equivalent. A completed PDAG (CPDAG, also called essential graph) is a PDAG representing the Markov equivalent class of a DAG.

2.2.2 Causal discovery

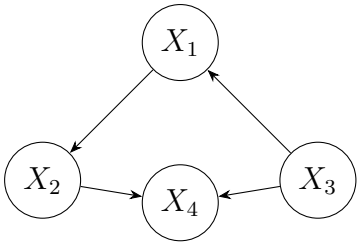
When the causal structure of the phenomenon under study is unknown or disputed, one should establish cause-effect relationships from existing data. Knowledge derived from data should combined together with prior knowledge about the phenomenon, whenever possible. As anticipated, the process of learning graphical structures with causal interpretation is



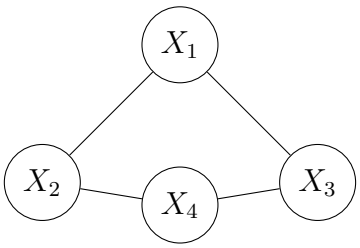
(a) DAG 1



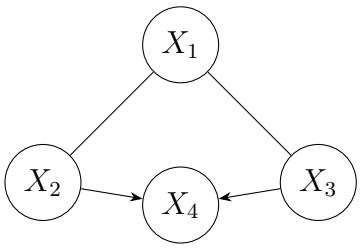
(b) DAG 2



(c) DAG 3



(d) Skeleton of DAG 1, DAG 2 and DAG 3



(e) CPDAG of DAG 1, DAG 2 and DAG 3

Figure 2.1: Examples of three observationally Markov equivalent DAGs (panels 2.1a, 2.1b, 2.1c, DAG 1 $\equiv$ DAG 2 $\equiv$ DAG 3), their Skeleton (panel 2.1d) and CPDAG (panel 2.1e).

known as causal discovery. Formally, let $\mathbb{G}$ be the set of all possible graphs defined over the variables of a certain dataset $\mathcal{D}$ and $\mathcal{G}^* \in \mathbb{G}$ be the true but unknown graph from which $\mathcal{D}$ has been generated. The causal discovery problem consists of estimating the true graph $\mathcal{G}^*$ (i.e., the ground truth) from the given observational data set $\mathcal{D}$ [76]. It is worth noting that most causal discovery algorithms do not recover the true $\mathcal{G}^*$, rather, they output its observationally equivalent class (i.e., a CPDAG, see Definition 2.2.6), leaving some edge directions undetermined from the data. Moreover, causal discovery algorithms usually impose a set of restrictive assumptions, which can result in the learned graph not accurately reflecting reality when using real-world data. These assumptions include [77, 78]:

- Independent and Identically Distributed (i.i.d.) samples: observational data in $\mathcal{D}$ are independent and identically distributed;
- Causal sufficiency: all common causes of the observed variables are included in $\mathcal{D}$; this implies that there are no hidden confounders;
- Causal Sufficiency of Measurements: observed data in $\mathcal{D}$ are assumed to be measured correctly, i.e., there are no measurement or discretization errors and the observed variables accurately reflect the underlying causal variables;
- Acyclicity: the causal graph is a Directed Acyclic Graph (DAG);

- Causal Markov assumption: Each variable is conditionally independent of its non-descendants given its direct causes (its parents in the graph);
- Faithfulness: All conditional independencies in the data are entailed by the causal structure represented in the graph;
- Linearity/Additive Noise Model (ANM) assumption: some methods assume that each observed variable is a linear combination of its direct causes (its parents in the causal graph), with an independent noise term added.

In practice, some algorithms can operate under weaker versions of these assumptions. A comprehensive description of the sets of assumptions in causal discovery is available in [77].

Structure learning methods for causal discovery

Structure learning methods are divided into two classes: constraint-based and score-based [32]. Additionally, there is a third category known as hybrid methods, which combines the features of both constraint-based and score-based approaches to mitigate the shortcomings of each approach.

Constraint-based methods Constraint-based methods try to recover the causal graph by exploiting a set of statistical conditional independence tests to identify dependence and independence relationships among variables, from a set of observational data. Then, they use graphical properties to infer the final graphical structure [32]. Since a set of independence relationships is usually compatible with more than one DAG, constraint-based algorithms generally return a CPDAG instead of converging to the true graph $\mathcal{G}^*$.

The most widely-recognized constraint-based method is the Peter-Clark (PC) algorithm [79], which consists of two steps: skeleton construction and orientation. In the first phase, the algorithm starts with a fully connected undirected graph, and a sequence of conditional independence tests is performed among pairs of adjacent variables. For each pair of tested variables, if conditional independence holds, the undirected edge is removed. This phase yields a skeleton. In the second phase, the remaining edges are oriented according to a set of rules, producing a CPDAG. Several variants of this algorithm have been proposed, for example, to relax the causal sufficiency assumption and handle latent variables (Fast Causal Inference-FCI, [80]) or to handle a high-dimensional data set (Max-Min Parent and Child Algorithm-MMPC, [81]).

Score-based methods Score-based algorithms are usually structured around the maximization of a measure of fitness of a graph $\mathcal{G}$ through a space of possible graphs $\mathbb{G}$ for the observed dataset $\mathcal{D}$, following a defined scoring criterion $S(\mathcal{G}, \mathcal{D})$ [32]:

$$\mathcal{G}^* = \underset{\mathcal{G} \in \mathbb{G}}{argmax}\, S(\mathcal{G}, \mathcal{D})$$

Typically used scores are the Akaike Information Criterion (AIC) [82] or the Bayesian Information Criterion (BIC) [83]. The Hill climbing algorithm [84] represents the simplest as well as one of the most widely used score-based algorithms for structure learning. The starting point of the algorithm is typically an empty or a random graph. Hill climbing investigates the space $\mathbb{G}$ by performing local modifications (additions, deletions or reversals of edges) at each iteration, to improve a score function, and stops when no single local change improves the score, returning a DAG. Hill climbing is an approximate algorithm that often gets stuck at a local maximum. Another famous score-based algorithm is the Greedy Equivalent Search (GES) [85]. The algorithm is composed by two phases: the forward search and the backward search. In the forward search phase, the algorithm starts from an empty graph and iteratively add edges that most improve the score, until no addition improves the score. During the backward search phase, the algorithm systematically eliminates edges from the graph identified in the prior phase if this action improves the score. The algorithm stops when no further edge removal improves the score, returning a CPDAG.

Structural Expectation-Maximization (Structural EM, [86]) is an iterative algorithm that jointly learns the structure and parameters of a causal graphical model (i.e., a Bayesian network) from incomplete data, namely missing data and/or unobserved (latent) variables. It uses a variant of the EM algorithm [87] to facilitate efficient search over large number of candidate structures, relaxing the assumption of causal sufficiency. Given a partially observed set of data, an initial graph $\mathcal{G}_0$ (e.g., a random graph over the set of observed features) and an initial set of parameters for $\mathcal{G}_0$, the algorithm alternates an expectation step (E) imputing missing values from the current fitted Bayesian network and a maximization step (M) that searches for the best graph structure and parameters that maximize the expected score (e.g., BIC) over the completed data. The E- and M-steps are repeated until convergence. An example of the use of Structural EM to learn a causal graph from observational EHR data in presence of latent factors will be presented in Chapter 4.

Integrating prior knowledge in causal discovery

Human knowledge about the existence or non-existence of some causal relationships, which can be obtained through expert elicitation, can be integrated into causal discovery by cre-

ating a knowledge base of *hard* or *soft constraints* [78]. Traditionally, human knowledge is presented to the causal discovery algorithm as predefined knowledge before beginning the learning process [88]. Hard constraints represent strict rules that the algorithm must follow during learning (e.g., forbidden or required edges). Any graph violating such constraints is automatically discarded. Prior knowledge based on hard constraints can be easily encoded in a knowledge base K, defined by whitelists and blacklists, i.e. sets of edges that should and should not be in the learned graph, respectively. In contrast, soft constraints do not enforce a specific causal structure; instead, they guide the learning process toward more plausible or preferred structures, while still allowing the data to influence and override those preferences if necessary. Prior knowledge based on soft constraints, usually associated with score-based approaches, can be incorporated by assigning probabilistic priors to certain edges or by introducing penalty terms into the learning process [78].

Evaluation metrics

There are typically two methods of assessing structure learning algorithms, depending on whether a real-world data set or a synthetic data set is used [32]. In the presence of real data, the ground truth causal graph $\mathcal{G}^*$ is typically unknown, therefore the learned graph can be evaluated by how well it explains the data set or by assessing its predictive capabilities, using traditional performance metrics (e.g., accuracy, sensitivity, specificity, Area under the ROC Curve-AUC). Alternatively, one can assume $\mathcal{G}^*$ and use it to generate a synthetic data set, learning a graph from that synthetic data set. In this configuration, the effectiveness of the structure learning algorithm can be evaluated by how closely it can approximate the ground truth. Frequently utilized evaluation metrics are the Structural Hamming Distance (SHD) [76] and the Structural Intervention Distance (SID)[89]. SHD quantifies the discrepancy between two graphical models by counting the number of operations (additions, removals, reversals of edges) needed to transform the learned graph into $\mathcal{G}^*$. SID, on the other hand, is characterized by interventional distributions and quantifies the number of interventional distributions that are incorrectly inferred from the graph. These metrics are usually called pattern metrics as they search for common patterns between the ground-truth graph and the learned graph.

2.2.3 Causal inference

Causal inference estimates causal effects between variables, once we already have (or assume) the underlying causal structure of the phenomenon of interest [32]. There are two widely accepted mathematical frameworks in the literature for causal inference: the potential out-

come framework (POF), also called Neyman-Rubin causal framework [90], and the Structural Causal Model (SCM) framework proposed by Pearl [91]. The former approach focuses on reasoning about potential outcomes for each unit under various treatments and directly defines causal effects through these outcomes. In contrast, the latter method formalizes causal effects using structural equations and the related joint distributions over interventions. Interventions are implemented by modifying the graph using a set of algebraic rules known as do-calculus. Although these frameworks are often regarded as equivalent, their relationship remains the subject of ongoing debate [92]. For the remainder of this thesis, I will focus on the SCM framework.

In "The Book of Why" [93] Pearl and Mackenzie introduced a causal hierarchy (or "ladder of causation"), which consists of three layers (or rungs) encoding different aspects of the underlying reality: the associational, the interventional, and the counterfactual aspect, roughly corresponding to different intuitive notions of human cognition, namely, "seeing", "doing", and "imagining". Following such hierarchy, causal inference can be differentiated into three levels of increasingly complex queries that require more advanced methods to address them effectively. The bottom layer embodies the notion of "seeing," which involves observing a certain phenomenon unfold and potentially drawing conclusions from it. This layer addresses information that is exclusively observational (factual) in nature. The middle layer embodies the notion of "doing," which involves intervening in the world to evaluate the effect of certain actions. Hence, this layer encodes information about what would hypothetically happen if some intervention were to be performed. Finally, the top layer embodies the notion of "imagining," which involves contemplating different alternative ways for how the world might be, even if these possibilities are at odds with the current state of the world. Hence, this layer involves queries about what would have happened, in a counterfactual sense, considering that another event (factual) actually took place. The hierarchy establishes formal frameworks in terms of the questions that they can represent and, ideally, answer (see Table 2.1). After establishing the causal hierarchy, we proceed to explore SCMs, which provide the mathematical structure necessary to rigorously define each level.

Structural causal models

The ladder of causation clearly defines a hierarchy that goes from observing associations, to reasoning about interventions, and finally to imagining counterfactual worlds. Structural Causal Models (SCMs) [91], combining features from structural equation models used in socioeconomics, the POF framework, and probabilistic graphical models, translate these concepts into a graphical and algebraic framework for causal inference [94].

A causal model M in the SCM framework is a 4-tuple $\langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$ where

Layer	Human action	Typical Query	Example
$\mathcal{L}_1$: Associational $P(Y \mid x)$	Seeing	How would observing $X = x$ change my knowledge of Y?	What does observing a symptom tell us about the disease?
$\mathcal{L}_2$: Interventional $P(Y \mid \text{do}(x), z)$	Doing	What if I act on X? How does intervening on $X = x$ influence Y given $Z = z$?	What if I start a low-carb diet, will my fasting glucose levels improve?
$\mathcal{L}_3$: Counterfactual $P(Y_x \mid x', y')$	Imaging	Was it X that caused Y? What if I had acted differently? $\rightarrow$ What would have been Y under $X = x$ given that we have observed $Y = y'$ under $X = x'$?	If I had started exercising daily last year, would I have avoided insulin therapy now?

Table 2.1: Pearl's causal hierarchy [93].

- $\mathbf{U} = \{U_1, U_2, \ldots, U_n\}$ is a set of exogenous variables, that are determined by factors outside the model, namely, background conditions for which no explanatory mechanism is encoded in model M;
- $\mathbf{V} = \{V_1, V_2, \ldots, V_m\}$ is a set of endogenous variables, that are determined by other variables in the model, i.e., variables in $\mathbf{U} \cup \mathbf{V}$;
- $\mathcal{F} = \{f_1, ..., f_m\}$ is a set of functions, also called structural equations. Each f_i, $i \in 1, \cdots, m$ maps variables in $\mathbf{U} \cup (\mathbf{V} \backslash V_i)$ to a specific endogenous variable V_i; the entire set $\mathcal{F}$ maps $\mathbf{U}$ to $\mathbf{V}$. In other words, each f_i tells us the value of V_i given its parents $Pa(V_i) \subseteq \mathbf{V} \backslash V_i$ and its exogenous variables $U_i \subseteq \mathbf{U}$, namely $V_i = f_i(Pa(V_i), U_i)$. For example, under the Additive Noise Model (ANM) assumption [74], $V_i = f_i(Pa(V_i)) + U_i$;
- $P(\mathbf{U})$ is a probability function defined over the domain of $\mathbf{U}$, summarizing the "state of the world" outside the variables of interest;

Every instantiation $\mathbf{U} = \mathbf{u}$ of the exogenous variables uniquely determines the values of all variables in $\mathbf{V}$. The vector $\mathbf{U} = \mathbf{u}$ can also be interpreted as an experimental "unit" which can stand for an individual subject, since it describes all factors needed to make $\mathbf{V}$ a deterministic function of $\mathbf{U}$. Their randomness, encoded in $P(\mathbf{U})$, induces variations in the endogenous set. Notably, when exogenous variables in $\mathbf{U}$ take fixed known (deterministic SCM), or parameterized (parametric SCM) values rather than values drawn from a probability distribution, then the probabilistic component $P(\mathbf{U})$ is not needed and M simplifies to a 3-tuple $\langle \mathbf{U}, \mathbf{V}, \mathcal{F} \rangle$. An example of parametric SCM will be discussed in Chapter 5.

M can be graphically represented by a directed graph $\mathcal{G}(M)$, characterized by nodes and edges. The set of nodes comprises exogenous and endogenous variables. Each directed edge represents a causal relationship between a couple of nodes. This graph identifies relationships among variables but does not specify the functional form of f_i.

We next define how SCMs provide the formal basis for the causal hierarchy presented above, as discussed in [95].

Layer 1 - Seeing

This level is about evaluating the probability of certain events occurring, leveraging statistical associations that arise naturally from the data-generating process encoded by the SCM. Formally, an SCM M induces an observational joint probability distribution $P^M(\mathbf{V})$ such that for each $\mathbf{Y} \subseteq \mathbf{V}$:

$$P^M(\mathbf{y}) = \sum_{\{\mathbf{u}|\mathbf{Y}(\mathbf{u})=\mathbf{y}\}} P(\mathbf{u})$$

where $\mathbf{Y}(\mathbf{u})$ is the solution for $\mathbf{Y}$ after evaluating $\mathcal{F}$ for $\mathbf{U} = \mathbf{u}$.

Layer 2 - Doing

This layer is about intervening on certain variables rather than merely observing them. A modification of an SCM gives natural valuations for quantities of this kind, as defined next.

Definition 2.2.7. *(Submodel).* Let M be a causal model, $\mathbf{X}$ be a set of variables in $\mathbf{V}$, and $\mathbf{x}$ be a particular realization of $\mathbf{X}$. A submodel $M_{\mathbf{x}} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}_{\mathbf{x}}, P(\mathbf{U}) \rangle$ of M is a causal model where $\mathcal{F}_{\mathbf{x}} = \{f_i : V_i \notin \mathbf{X}\} \cup \{\mathbf{X} = \mathbf{x}\}$. Hence, $\mathcal{F}_{\mathbf{x}}$ is formed by deleting from $\mathcal{F}$ all functions f_i corresponding to members of set $\mathbf{X}$ and replacing them with the set of constant functions $\mathbf{X} = \mathbf{x}$.

Submodels (or interventional SCM) are useful for representing the effect of local interventions (or actions) and hypothetical changes, including those dictated by counterfactual antecedents. Performing an external intervention is modeled through the replacement of the original mechanisms associated with some variables $\mathbf{X}$ with a constant $\mathbf{x}$, which is represented by the do-operator. If we interpret each function f_i in $\mathcal{F}$ as an independent physical mechanism and define the action $do(\mathbf{X} = \mathbf{x})$ as the minimal change in M required to make $\mathbf{X} = \mathbf{x}$ hold true under any $\mathbf{u}$, then $M_{\mathbf{x}}$ represents the model that results from such a minimal change, since it differs from M by only those mechanisms that directly determine the variables in $\mathbf{X}$. In other words, the effect of action $do(\mathbf{X} = \mathbf{x})$ on M is given by the submodel $M_{\mathbf{x}}$ [91].

The impact of the intervention on an outcome variable $\mathbf{Y}$ is called potential response:

Definition 2.2.8. *(Potential response).* Let $\mathbf{X}$ and $\mathbf{Y}$ be two sets of variables in $\mathbf{V}$, and $\mathbf{u}$ be a unit. The potential response of $\mathbf{Y}$ to action $do(\mathbf{X} = \mathbf{x})$, denoted $\mathbf{Y}_{\mathbf{x}}(\mathbf{u})$, is the solution for $\mathbf{Y}$ of the set of equations $\mathcal{F}_{\mathbf{x}}$.

An SCM M induces a family of joint probability distributions over $\mathbf{V}$, one for each possible intervention $do(\mathbf{X} = \mathbf{x})$. For a subset $\mathbf{Y} \subseteq \mathbf{V}$, the post-intervention distribution is:

$$P^M(\mathbf{y}_{\mathbf{x}}) = \sum_{\{\mathbf{u}|\mathbf{Y}_{\mathbf{x}}(\mathbf{u})=\mathbf{y}\}} P(\mathbf{u})$$

This distribution is often written as $P(\mathbf{Y}|do(\mathbf{x}))$.

Layer 3 - Imaging

This level allows us to reason about hypothetical scenarios. It captures individual-level counterfactual reasoning. Counterfactuals are defined as what-if questions, i.e., conditional assertions whose antecedent is false and whose consequent describes how the world would have been if the antecedent had occurred.

Definition 2.2.9. *(Interventional Counterfactual).* Let $\mathbf{Y}$ be a variable (or a subset) in $\mathbf{V}$, and let $\mathbf{X}$ be a subset of $\mathbf{V}$. The counterfactual sentence "The value that $\mathbf{Y}$ would have obtained, had $\mathbf{X}$ been $\mathbf{x}$ in unit (or situation) $\mathbf{U} = \mathbf{u}$, given the specific evidence $\mathbf{e}$ we have about the factual world" is interpreted as denoting the potential response $\mathbf{Y}_{\mathbf{x}}(\mathbf{u})$. Letting $M_{\mathbf{x}}$ be a modified version of M, with the equation(s) of set $\mathbf{X}$ replaced by $\mathbf{X} = \mathbf{x}$, the formal definition of the counterfactual $\mathbf{Y}_{\mathbf{x}}(\mathbf{u})$ reads $\mathbf{Y}_{\mathbf{x}}(\mathbf{u}) = \mathbf{Y}_{\mathbf{M}_{\mathbf{x}}}(\mathbf{u})$. Hence, the counterfactual $\mathbf{Y}_{\mathbf{x}}(\mathbf{u})$ in model M is defined as the solution for $\mathbf{Y}$ in the "modified" submodel $M_{\mathbf{x}}$.

Interventional counterfactual queries can be computed following a three-step procedure [96]:

1. Abduction: update $P(\mathbf{u})$ by the evidence $\mathbf{e}$, to obtain $P(\mathbf{u}|\mathbf{e})$.
2. Action: Replace the equations determining the variables in the set $\mathbf{X}$ by $\mathbf{X} = \mathbf{x}$, i.e., modify M by the action $do(\mathbf{x})$, to obtain the submodel $M_{\mathbf{x}}$.
3. Prediction: Use the updated probability $P(\mathbf{u}|\mathbf{e})$ in conjunction with $M_{\mathbf{x}}$ to compute the probability of the counterfactual consequence $\mathbf{Y} = \mathbf{y}$.

Drawing values of exogenous variables $\mathbf{U}$ following the distribution $P(\mathbf{U})$ over an SCM M induces a family of joint probability distributions over counterfactual events $\mathbf{Y}_{\mathbf{x}} = \mathbf{y}, \cdots, \mathbf{Z}_{\mathbf{w}} = \mathbf{z}$ (for short, $\mathbf{y}_{\mathbf{x}}, \cdots, \mathbf{z}_{\mathbf{w}}$). For subsets $\mathbf{Y}, \cdots, \mathbf{Z}, \mathbf{X}, \mathbf{W} \subseteq V$, the distribution

over a given collection of counterfactual events is defined as:

$$P^M(\mathbf{y_x}, \cdots, \mathbf{z_w}) = \sum_{\{\mathbf{u}|\mathbf{Y_x}(\mathbf{u})=\mathbf{y}, \cdots, \mathbf{Z_w}(\mathbf{u})=\mathbf{z}\}} P(\mathbf{u})$$

Balke and Pearl [97] proposed the use of an auxiliary structure called *twin network*, to solve counterfactual queries. A twin network is an SCM where the endogenous nodes and their structural equations are duplicated while remaining driven by the same exogenous variables. The twin network includes (i) a factual world, which describes the relationships between the observed endogenous variables (what actually happened), and (ii) a counterfactual world, which describes the relationships between alternative versions of the same variables (a hypothetical scenario), governed by the same structural equations of the factual world. Thus, the twin network defines a joint model over the factual and counterfactual worlds, linking them through shared exogenous variables. This representation allows us to compute the probability of a given outcome in both worlds and to compare them at the individual level, meaning the underlying external factors influencing the system do not change.

In practice, the counterfactual query cannot always be uniquely determined. In such cases, the counterfactual query is said to be non-identifiable. For example, the presence of latent (unobserved) exogenous variables, incorporating confounding effects, may lead to non-identifiability in counterfactual inference. In such cases, the query cannot be precisely estimated from observational data without bias [31]. Hence, the resulting target counterfactual probability can only be known to belong to an interval that captures all possible scenarios that could be inferred from data. This means that, given a causal diagram $\mathcal{G}$, observational data, and a collection of interventional distributions, the optimal bound $[l, r]$ over an arbitrary counterfactual probability $P(\mathbf{y_x}, \cdots, \mathbf{z_w})$ can be determined by solving the following optimization problem on the twin network [98] :

$$[l, r] = \left[\min_{M \in \mathcal{M}(\mathcal{G})} P^M(\mathbf{y_x}, \ldots, \mathbf{z_w}), \max_{M \in \mathcal{M}(\mathcal{G})} P^M(\mathbf{y_x}, \ldots, \mathbf{z_w}) \right]$$

where $\mathcal{M}(\mathcal{G}) = \{M | \mathcal{G}_M = \mathcal{G}\}$ denotes the set of all SCMs with graph $\mathcal{G}$, i.e., the set of all SCMs that are compatible with the available observational and interventional distributions but may differ in their latent variable distributions or structural equations. One way to solve this optimization problem numerically, once assuming that the domains of $\mathbf{V}$ and $\mathbf{U}$ are discrete and finite, consists of using the Expectation-Maximization for Causal Computation (EMCC) method [99], a framework that extends the EM scheme to causal inference, for counterfactual estimation under uncertainty. EMCC is a sampling method that generates a set of points within the exact counterfactual bounds by making inferences on Bayesian

networks that share the topology of the original model. The lower and upper limits of these points serve as an approximation of the exact range.

In practical scenarios (e.g., see Chapter 4), we often do not know explicitly the set of structural equations $\mathcal{F}$ describing the structural mechanisms in M. In these scenarios, a relaxed version of the EMCC, described in [100], can be adopted to approximate counterfactual bounds in the presence of unknown deterministic structural equations, by approximating them with Conditional Probability Tables (CPTs). Each CPT can be thought of as a probabilistic surrogate for the unknown structural equation. The EMCC algorithm can learn the best set of CPTs to fit observed and counterfactual data.

Chapter 3

Multi-class Counterfactual Explanations Using Support Vector Data Description

As discussed in earlier chapters, counterfactual explanations can enhance transparency and trust in predictive models by identifying how changes in key variables could alter their outcomes. These explanations can also offer guidance on managing modifiable risk factors for chronic disease prevention. Their value is particularly evident in multi-class settings, where predictions span multiple risk levels. In such settings, counterfactual explanations could help clarify which factors drive progressive transitions between risk categories, thereby improving model interpretability and supporting informed and targeted preventive decision-making.

Building on these considerations, this Chapter outlines an original methodological framework for generating counterfactual explanations in multi-class classification problems[1]. Part of the work presented in this Chapter has been published in two Journal articles [101, 102]. The original contributions of this study can be summarized as follows:

- formulation of an extension of the Support Vector Data Descriptor [53] to Multi-Class classification scenarios (MC-SVDD, Section 3.1) and development of an algorithm to control its classification error (Section 3.1.2);
- development of a counterfactual explainer based on bounded classification regions derived from MC-SVDD (Section 3.2);

[1]Code available at: https://github.com/lenattimarta/MUCH

- formulation of a counterfactual conformity measure, leveraging conformal predictions [103] guarantees to filter counterfactual explanations that do not reach the desired level of confidence (Section 3.2.3);

- application and comparison of the proposed counterfactual explainer with state-of-the-art approaches, using benchmark datasets (Appendix A) and real-world data drawn from primary care EHRs targeting chronic disease prevention (Section 3.3), i.e., the creation of personalized strategies to reduce the risk of developing cardiovascular diseases (CVDs, *Case Study 1*) in patients diagnosed with Chronic Obstructive Pulmonary Disease (COPD).

3.1 Multi-Class Support Vector Data Description (MC-SVDD)

Multi-class classification is the task of assigning a new instance into one among at least three classes. Different strategies exist for tackling multi-class problems. Some algorithms, such as decision trees and DNNs, natively support multiple output classes, while others, like logistic regression, provide exclusively binary outputs. For the latter, additional adaptations are required to extend them to multi-class tasks. We can distinguish two types of multi-class adaptations [104]: *One-Vs-One* (OvO) and *One-Vs-Rest* (OvR). In OvO techniques, the task is solved by training $m(m-1)/2$ binary classifiers, where m is the number of classes and each binary classifier distinguish between two classes. Then, classification is performed through majority voting i.e., the final prediction is set as the class most frequently returned by the binary classifiers. Due to its incremental adaptation to multiple outputs, the OvO approach lacks a comprehensive view of the relationships among all classes. In contrast, OvR techniques train m binary classifiers, each focusing on differentiating a target class from the rest. The instance is ultimately assigned to the class that achieves the highest probability score.

The multi-class algorithm proposed in this Chapter generalizes the well-known SVDD of Tax and Duin [105, 53] to the multi-class case. The SVDD is a state-of-the-art algorithm for outlier detection that is capable of enclosing meaningful data points within an hyperspherical region. Points outside such region are considered outliers with respect to the data distribution. Hence, SVDD is generally considered a one-class or semi-supervised learning method [106]. Similarly to Support Vector Machine (SVM), it exploits a kernel trick to separate data in a high-dimensional feature space. But instead of finding a separating hyperplane, it finds a minimal-volume hypersphere that describes data.

Several studies have provided extension of the SVDD algorithm. Some of them, e.g., [107], focused on the detection of anomalous objects rather than providing canonical classification. Huang et al. [108] proposed Two Class-SVDD (TC-SVDD), an extension of SVDD to binary classification tasks, by building two hypersphere-shaped boundaries simultaneously. The framework proposed by Duan et al. [109], integrates SVDD with binary decision trees to handle multi-class classification problems, while Guo et al. [110] proposed a multi kernel learning adaptation to SVDD (MKL-SVDD) to design the kernel weights for multiple kernels and obtain the optimal kernel combination. Hou et al. [111] developed a multi-class SVDD algorithm to classify multiple classes of planetary gear faults based on the method proposed by [112] that minimizes the radius of each hypersphere while maximizing the distance between them. However, the boundary between couples of classes is optimized for each pair of centers, without including additional constraints inherent to the other classes. The *Multi-Class SVDD* (MC-SVDD) approach proposed in this Chapter solves the problem in one shot, without repetitive adaptations, and provides the weights for classification as an exact solution to an optimization problem.

3.1.1 Model formulation

MC-SVDD enables the identification of minimal-volume hyperspheres in a kernel space, that can distinctively partition the data into m classes. In practice, we minimize the total volume of the m hyperspheres with the constraint that (*C1*) for each training object, the distance between the center of one hypersphere and the instance is smaller than the radius of that hypersphere, that is, the instance belongs to a specific output class and (*C2*) the instance do not fall inside other hyperspheres, that is, the instance should not belong to other output classes. This formulation follows a OvR scheme.

Let $\mathcal{D}_{tr} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n} \in \mathcal{X}_{tr} \times \mathcal{Y}_{tr}$ be a training set that contains a total of n records, divided among m possible output classes, with $n_1, n_2, \ldots, n_m$ records in each class. Records in $\mathcal{D}_{tr}$ are labeled and ordered according to their output class, i.e., $\mathbf{y} = [1 \ldots 1\ 2 \ldots 2\ m \ldots m]^\top$. Analogously, we define a test set $\mathcal{D}_{ts} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n_{ts}} \in \mathcal{X}_{ts} \times \mathcal{Y}_{ts}$.

Let $\mathbf{a}_k$ and R_k denote the center and radius of a given hypersphere k. To allow a flexible description of the hyperspheres we introduce $\varphi : \mathcal{X} \longrightarrow \mathcal{V}$, a *feature map* from the space of the input features $\boldsymbol{x} \in \mathcal{X}$ to a higher dimensional inner product space $\mathcal{V}$. Such feature map allows us to exploit kernels $K_{i,j} = K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \varphi(\boldsymbol{x}_i)^\top \varphi(\boldsymbol{x}_j),\ i \in [n],\ j \in [n]$ that satisfy the Mercer's theorem [113]. Searching for hyperspheres of minimum volume that satisfy *C1* and

C2 means finding the solution of the following optimization problem:

$$\min\ F(R_k; \mathbf{a}_k) = \sum_{k=1}^{m} R_k^2 \tag{3.1a}$$

$$\text{s.t.}\quad \left\|\varphi(\boldsymbol{x}_i^k) - \mathbf{a}_k\right\|^2 \leq R_k^2,\ i \in [n_k], \forall k \tag{3.1b}$$

$$\left\|\varphi(\boldsymbol{x}_i^k) - \mathbf{a}_h\right\|^2 \geq R_h^2,\ i \in [n_k], \forall h \neq k \tag{3.1c}$$

We can follow the approach used in the original SVDD formulation [105], which consists in reducing Eq. 3.1 to a quadratic programming problem. To prevent constraints from being overly restrictive in the presence of outliers, we introduce slack variables $\boldsymbol{\xi}^{kk}, \boldsymbol{\xi}^{kh}$ that permit controlled violations. Indeed, we soften the *C1* and *C2* constraints, adding a penalty term to the optimization. The minimization problem transforms into

$$\min\ F(R_k, \mathbf{a}_k, \xi^{kh}) = \sum_{k=1}^{m} R_k^2 + \sum_{k=1}^{m}\sum_{h=1}^{m} C_{kh} \sum_{i=1}^{n_k} \xi_i^{kh} \tag{3.2a}$$

$$\text{s.t.}\quad \left\|\varphi(\boldsymbol{x}_i^k) - \mathbf{a}_k\right\|^2 \leq R_k^2 + \xi_i^{kk},\ i \in [n_k],\ \forall k \tag{3.2b}$$

$$\left\|\varphi(\boldsymbol{x}_i^k) - \mathbf{a}_h\right\|^2 \geq R_h^2 - \xi_i^{kh},\ i \in [n_k],\ \forall h \neq k \tag{3.2c}$$

$$\xi_i^{kk} \geq 0,\ i \in [n_k],\ \forall k \tag{3.2d}$$

$$\xi_i^{hk} \geq 0,\ i \in [n_k],\ \forall h \neq k \tag{3.2e}$$

where the parameter C_{kh} controls the misclassification error between classes.
In the relaxed formulation (Eq. 3.2), the distance from an object $\varphi(\boldsymbol{x}_i^k)$, belonging to class k, to the class center $\mathbf{a}_k$ should ideally not exceed R_k^2, with larger distances being penalized. Similarly, the distance from $\varphi(\boldsymbol{x}_i^k)$ to any other class center $\mathbf{a}_h$ ($h{\neq}k$) should ideally exceed R_h^2, with smaller distances being penalized.

Then, we incorporate constraints (3.2b)-(3.2e) into (3.2a), with the introduction of Lagrange multipliers. Hence, we define the Lagrangian function

$$
\begin{aligned}
&L(R_k; \mathbf{a}_k; \boldsymbol{\xi}^{kk}, \boldsymbol{\xi}^{kh}; \boldsymbol{\alpha}^{kk}, \boldsymbol{\alpha}^{kh}; \boldsymbol{\gamma}^{kk}, \boldsymbol{\gamma}^{kh}) \\
&= \sum_{k=1}^{m} R_k^2 + \sum_{k=1}^{m} \sum_{h=1}^{m} C_{kh} \sum_{i=1}^{n_k} \xi_i^{kh} \\
&- \sum_{k=1}^{m} \sum_{i=1}^{n_k} \alpha_i^{kk} \Big(R_k^2 + \xi_i^{kk} - \left\| \varphi(\boldsymbol{x}_i^k) - \mathbf{a}_k \right\|^2 \Big) \\
&- \sum_{h \neq k} \sum_{i=1}^{n_h} \alpha_i^{kh} \Big(\left\| \varphi(\boldsymbol{x}_i^k) - \mathbf{a}_h \right\|^2 - R_h^2 + \xi_i^{kh} \Big) \\
&- \sum_{k=1}^{m} \sum_{i=1}^{n_k} \gamma_i^{kk} \xi_i^{kk} - \sum_{h \neq k} \sum_{i=1}^{n_h} \gamma_i^{kh} \xi_i^{kh}
\end{aligned}
\tag{3.3}
$$

with the Lagrange multipliers $\boldsymbol{\alpha}^{kk}, \boldsymbol{\alpha}^{kh}, \boldsymbol{\gamma}^{kk}, \boldsymbol{\gamma}^{kh} \geq 0$ (3.4). In the dual form, L should be maximized with respect to the Lagrange multipliers, so setting the partial derivatives to zero gives the new constraints (3.5)-(3.6) and (3.7)-(3.8):

$$
\frac{\partial L}{\partial R_k} = 0 \Rightarrow \sum_{i=1}^{n_k} \alpha_i^{kk} - \sum_{h \neq k} \sum_{i=1}^{n_h} \alpha_i^{kh} = 1 \tag{3.5}
$$

$$
\frac{\partial L}{\partial \mathbf{a}_k} = 0 \Rightarrow \mathbf{a}_k = \sum_{i=1}^{n_k} \alpha_i^{kk} \varphi(\boldsymbol{x}_i^k) - \sum_{h \neq k} \sum_{i=1}^{n_h} \alpha_i^{kh} \varphi(\boldsymbol{x}_i^h) \tag{3.6}
$$

$\forall k \in [m]$ and $\forall h \neq k$. With respect to the slack variables

$$
\frac{\partial L}{\partial \xi_i^{ss}} = 0 \Rightarrow C_{ss} - \alpha_i^{ss} - \gamma_i^{ss} = 0 \Rightarrow 0 \leq \alpha_i^{ss} \leq C_{ss} \tag{3.7}
$$

$$
\frac{\partial L}{\partial \xi_i^{st}} = 0 \Rightarrow C_{st} - \alpha_i^{st} - \gamma_i^{st} = 0 \Rightarrow 0 \leq \alpha_i^{st} \leq C_{st} \tag{3.8}
$$

$\forall s \in [m]$ and $\forall t \neq s$.

Substituting (3.5) and (3.6) into Eq. 3.3, the Lagrangian is expressed as follows

$$\begin{aligned} L &= \sum_{k=1}^{m}\sum_{i=1}^{n_k}\alpha_i^{kk}\big(\varphi(\boldsymbol{x}_i^k)\cdot\varphi(\boldsymbol{x}_i^k)\big) \\ &- \sum_{h\neq k}\sum_{i=1}^{n_k}\alpha_i^{kh}\big(\varphi(\boldsymbol{x}_i^k)\cdot\varphi(\boldsymbol{x}_i^k)\big) \\ &- \sum_{i=1}^{m}\sum_{i,j=1}^{n_k}\alpha_i^{kk}\alpha_j^{kk}\big(\varphi(\boldsymbol{x}_i^k)\cdot\varphi(\boldsymbol{x}_j^k)\big) \\ &- \sum_{h\neq k}\sum_{i,j=1}^{n_k}\alpha_i^{kh}\alpha_j^{kh}\big(\varphi(\boldsymbol{x}_i^k)\cdot\varphi(\boldsymbol{x}_j^k)\big) \\ &+ 2\sum_{h\neq k}\sum_{i=1}^{n_k}\sum_{j=1}^{n_h}\alpha_i^{kk}\alpha_j^{kh}\big(\varphi(\boldsymbol{x}_i^k)\cdot\varphi(\boldsymbol{x})_j^h\big) \end{aligned} \tag{3.9}$$

The maximization of (3.9) provides the set of Lagrangian multipliers $\boldsymbol{\alpha}^{kk}, \boldsymbol{\alpha}^{kh}$ $\forall k \in [m]$, $\forall h \neq k$, while $\boldsymbol{\gamma}^{kk}$ and $\boldsymbol{\gamma}^{kh}$ can be eliminated by exploiting their positivity and first-order conditions on slack variables. The Lagrange multipliers assume different values based on whether the training samples satisfy or violate constraints (3.2b) and (3.2c), namely:

$$\begin{aligned} \big\|\varphi(\boldsymbol{x}_i^k) - \mathbf{a}_k\big\|^2 < R_k^2 &\;\Rightarrow\; \alpha_i^{kk} = 0 \\ \big\|\varphi(\boldsymbol{x}_i^k) - \mathbf{a}_h\big\|^2 > R_h^2 &\;\Rightarrow\; \alpha_i^{kh} = 0 \\ \big\|\varphi(\boldsymbol{x}_i^k) - \mathbf{a}_k\big\|^2 = R_k^2 &\;\Rightarrow\; 0 < \alpha_i^{kk} < C_{kk} \\ \big\|\varphi(\boldsymbol{x}_i^k) - \mathbf{a}_h\big\|^2 = R_h^2 &\;\Rightarrow\; 0 < \alpha_i^{kh} < C_{kh} \\ \big\|\varphi(\boldsymbol{x}_i^k) - \mathbf{a}_k\big\|^2 > R_k^2 &\;\Rightarrow\; \alpha_i^{kk} = C_{kk} \\ \big\|\varphi(\boldsymbol{x}_i^k) - \mathbf{a}_h\big\|^2 < R_h^2 &\;\Rightarrow\; \alpha_i^{kh} = C_{kh} \end{aligned} \tag{3.10}$$

$\forall k \in [m]$ and $\forall h \neq k$ respectively. In line with the original SVDD formulation [105], the instances $\boldsymbol{x}_i^k$ with $\alpha_i^{kk} > 0$ and $\alpha_i^{kh} > 0$ are called *support vectors* for class k. By definition, the radius R_k is the distance from the center $\mathbf{a}_k$ of the hypersphere to any of the support vectors of class k with Lagrange multipliers strictly minor than the parameters $C_{k\{\cdot\}}$. Therefore,

$$
\begin{aligned}
R_k^2 &= \left\|\varphi(\boldsymbol{x}_s^k) - \mathbf{a}_k\right\|^2 = \left(\varphi(\boldsymbol{x}_s^k) \cdot \varphi(\boldsymbol{x}_s^k)\right) \\
&- 2\sum_{i=1}^{n_k} \alpha_i^{kk} \left(\varphi(\boldsymbol{x}_s^k) \cdot \varphi(\boldsymbol{x}_i^k)\right) \\
&+ 2\sum_{h\neq k} \sum_{i=1}^{n_h} \alpha_i^{kh} \left(\varphi(\boldsymbol{x}_s^k) \cdot \varphi(\boldsymbol{x}_i^h)\right) \\
&+ \sum_{i,j=1}^{n_k} \alpha_i^{kk} \alpha_j^{kk} \left(\varphi(\boldsymbol{x}_i^k) \cdot \varphi(\boldsymbol{x}_j^k)\right) \\
&- 2\sum_{h\neq k} \sum_{i=1}^{n_k} \sum_{j=1}^{n_h} \alpha_i^{kk} \alpha_j^{kh} \left(\varphi(\boldsymbol{x}_i^k) \cdot \varphi(\boldsymbol{x}_j^h)\right) \\
&+ \sum_{h\neq k} \sum_{i,j}^{n_h} \alpha_i^{kh} \alpha_j^{kh} \left(\varphi(\boldsymbol{x}_i^h) \cdot \varphi(\boldsymbol{x}_j^h)\right)
\end{aligned}
\tag{3.11}
$$

for any support vector of class k (i.e., $\boldsymbol{x}_s^k$) with $0 < \alpha_i^{kk} < C_{kk}$ or $0 < \alpha_i^{kh} < C_{kh}$.
Once the centers $\mathbf{a}_k$ and radii R_k have been computed for all classes $k \in [m]$, the corresponding classification regions S_k are defined.

Definition 3.1.1. (*MC-SVDD classification regions*) MC-SVDD training yields m hyperspherical classification regions defined as follows:

$$
S_k \doteq \{\boldsymbol{x} \in \mathbb{R}^N : \|\boldsymbol{x} - \mathbf{a}_k\|^2 \leq R_k^2, \|\boldsymbol{x} - \mathbf{a}_h\|^2 \geq R_h^2; h \in [m]; h \neq k\} \; \forall k \in [m] \tag{3.12}
$$

where $R_k^2, R_h^2, \mathbf{a}_k, \mathbf{a}_h$ represent the radii and the centers of the spheres. Given a dataset $\mathcal{D}$, we indicate with MC-SVDD($\mathcal{D}$) the application of the trained MC-SVDD to dataset $\mathcal{D}$.

To test a new sample $\tilde{\boldsymbol{x}} \in \mathcal{D}_{ts}$, i.e., predict $\tilde{y} = \text{MC-SVDD}(\tilde{\boldsymbol{x}})$ it is necessary to calculate its distance from the center of each hypersphere k, i.e.

$$
\begin{aligned}
d_k \;\doteq\; & \|\tilde{\boldsymbol{x}} - \mathbf{a}_k\|^2 = \big(\varphi(\tilde{\boldsymbol{x}}) \cdot \varphi(\tilde{\boldsymbol{x}})\big) \\
- \; & 2\sum_{i=1}^{n_k} \alpha_i^{kk} \big(\varphi(\tilde{\boldsymbol{x}}) \cdot \varphi(\boldsymbol{x}_i^k)\big) \\
+ \; & 2\sum_{h\neq k} \sum_{i=1}^{n_h} \alpha_i^{kh} \big(\varphi(\tilde{\boldsymbol{x}}) \cdot \varphi(\boldsymbol{x}_i^h)\big) \\
+ \; & \sum_{i,j=1}^{n_k} \alpha_i^{kk} \alpha_j^{kk} \big(\varphi(\boldsymbol{x}_i^k) \cdot \varphi(\boldsymbol{x}_j^k)\big) \\
- \; & 2\sum_{h\neq k} \sum_{i=1}^{n_k} \sum_{j=1}^{n_h} \alpha_i^{kk} \alpha_j^{kh} \big(\varphi(\boldsymbol{x}_i^k) \cdot \varphi(\boldsymbol{x}_j^h)\big) \\
+ \; & \sum_{h\neq k} \sum_{i,j}^{n_h} \alpha_i^{kh} \alpha_j^{kh} \big(\varphi(\boldsymbol{x})_i^h \cdot \varphi(\boldsymbol{x}_j^h)\big)
\end{aligned}
\tag{3.13}
$$

The test object $\tilde{\boldsymbol{x}}$ is then classified according to the following criterion:

1. If $d_k \leq R_k^2$ and $d_h > R_h^2 \; \forall h \neq k$, then $\tilde{\boldsymbol{x}}$ belongs to class k;
2. If $d_k \leq R_k^2$ for several $k \in [m]$, then $\tilde{\boldsymbol{x}}$ belongs to class $k' = argmin_{k\in\mathcal{K}} d_k$, where $\mathcal{K} = \{k \in [m] \mid d_k \leq R_k^2\}$;
3. If $d_k > R_k^2 \; \forall k$ or $\#\{k' \mid k' = argmin_{k\in\mathcal{K}} d_k\} > 1$, then $\tilde{\boldsymbol{x}}$ is unclassified.

That is, the distances between all samples in each class and the center of that class should be smaller than the radius of the corresponding hypersphere, and the distances between all samples in each class and the centers of all the other classes should be larger than the radius of the corresponding hyperspheres. If a given sample belongs to more than one hypersphere, the sample is assigned to the class that lies at a minimum distance. In any other case, the sample is unclassified.

3.1.2 False Positive Rate control for MC-SVDD

This section extends the algorithm to control the classification error of the binary SVDD (originally proposed in [114]) to the multi-class case, with the aim of obtaining well-defined and reliable classification regions, highly representative of each target class.

Disease risk prediction models with high misclassification rates fail to empower, or even mislead clinicians, systematically missing patients in need of a treatment or unnecessarily suggesting treatments for healthy patients. In this perspective, it would be advisable for the

model to not assign a sample to any class if its classification is uncertain and handle it as an outlier, thus abstaining from providing a decision in case of doubtful samples. Controlling the classification error increases the reliability of MC-SVDD, by distinguishing points classified with high confidence from points whose classification is uncertain.

The FPR control approach for MC-SVDD is outlined in Algorithm 1 and consists of the following steps. An OvR strategy is applied to refine each of the m classification regions obtained using MC-SVDD. For each class $k \in [m]$, negative SVDDs (i.e., SVDDs with a single target class as defined in [105]) are iteratively trained to classify points belonging to the target class k (positive samples) against all other classes (negative samples). This iterative process refines the shape of the classification region for class k until the number of negative instances incorrectly classified as positives falls below a predefined threshold η (here set equal to 0.1) or until a maximum number of iterations is reached (e.g., $N_{\max} = 1000$). This refinement sharpens class boundaries, reduces overlap between regions, and improves reliability.

Algorithm 1 MC-SVDD FPR control

Input: $S_1, S_2, \ldots, S_m$ regions from MC-SVDD, threshold on FPR (η), maximum number of iterations ($N_{\max}$).
Output: FPR reduced regions $S_1^*, S_2^*, \ldots, S_m^*$.

for all $k \in [m]$ **do**
  for all $\boldsymbol{x} \in \mathcal{X}$ **do**
    assign

$$\boldsymbol{x} \longrightarrow y \doteq \begin{cases} +1 & \text{if} \quad \boldsymbol{x} \in S_k \\ -1 & \text{otherwise} \end{cases}$$

    and build dataset $\mathcal{D}_{k,0} = \{(\boldsymbol{x}, y) \mid \boldsymbol{x} \in \mathcal{X}, y \in \{-1, +1\}\}$
  end for
  Compute $S_{k,0}^* \doteq \text{SVDD}(\mathcal{D}_{k,0})$
  Set $N = 1$
  while $\text{FPR}(S_{k,N}^*) > \eta$ AND $N \leq N_{\max}$ **do**
    $S_{k,N}^* = \text{SVDD}(S_{k,N-1}^*)$
    $N = N + 1$
  end while
  $S_k^* = S_{k,N}^*$
end for

3.2 MUlti-class Counterfactual explanations via Halton Sampling (MUCH)

This section describes a novel method to leverage the classification regions defined by the MC-SVDD to generate counterfactual explanations in multi-class classification scenarios.

3.2.1 Analytical solution

A dataset $\mathcal{D}$ can be characterized by a subset of modifiable/partially modifiable features $\mathbf{u}$ and a subset of non-modifiable features $\mathbf{z}$. The former can be manipulated, either entirely or partially, through internal or external interventions. One example is clinical biomarkers which are modifiable by therapies or lifestyle changes. The latter are not manipulable or irreversible by nature (e.g., age and diagnosed chronic diseases). As a consequence, an observation $\boldsymbol{x} \in \mathcal{D}$ can be defined as

$$\boldsymbol{x} = (\mathbf{u}, \mathbf{z}) = \left(u^1, u^2, \ldots, u^p, z^1, z^2, \ldots, z^q\right) \in \mathbb{R}^{p+q}$$

Given a trained classifier providing m classification regions (i.e., MC-SVDD$(\cdot)$ as defined in Section 3.1.1) and an observation[2] $\boldsymbol{x}_{f_a} = (\mathbf{u}, \mathbf{z})_{f_a} \in S_a$, called *factual*, the search for its counterfactual explanation consists in determining the minimum joint variation $\Delta\mathbf{u}^*$ of the modifiable and partially modifiable input features necessary to obtain the closest observation

$$\boldsymbol{x}_{f_a}^{cf_b} \doteq (\mathbf{u} + \Delta\mathbf{u}^*, \mathbf{z})_{f_a}^{cf_b}$$

that belongs to the classification region S_b, different from S_a.
Specifically, $\Delta\mathbf{u}^*$ is estimated by solving the following optimization problem:

$$\begin{aligned}
\min_{\Delta\mathbf{u}\in\mathbb{R}^p} \quad & dist\left((\mathbf{u}, \mathbf{z})_{f_a}, (\mathbf{u} + \Delta\mathbf{u}, \mathbf{z})_{f_a}^{cf_b}\right) && \text{(3.14a)}\\
\text{subject to} \quad & \left\|(\mathbf{u} + \Delta\mathbf{u}, \mathbf{z})_{f_a}^{cf_b} - \mathbf{a}_b\right\|^2 \leq R_b^2 && \text{(3.14b)}\\
& \left\|(\mathbf{u} + \Delta\mathbf{u}, \mathbf{z})_{f_a}^{cf_b} - \mathbf{a}_k\right\|^2 \geq R_k^2, && \text{(3.14c)}\\
& \text{with } k \in [m] \text{ and } k \neq b,
\end{aligned}$$

[2]For the sake of clarity, in the multi-class scenario I will use a slightly different notation from the one in Definition 2.1.1 to explicitly specify both the factual and the counterfactual class, here indicated with a and b, respectively.

where $\text{dist}(\cdot,\cdot)$ is the selected distance metrics (e.g., the Euclidean norm), Eq. 3.14b constraints $\boldsymbol{x}_{f_a}^{cf_b}$ to lie inside S_b and Eq. 3.14c constraints $\boldsymbol{x}_{f_a}^{cf_b}$ to lie outside all the regions $S_k \neq S_b$. It is worth noting that, for each factual $\boldsymbol{x}_{f_a} \in S_a$, we can find a set of $m-1$ counterfactual explanations, that is, one for each class b different from a. In other words, for a set of factuals $\mathbf{F}_a \subseteq S_a$ we obtain a set of counterfactual explanations $\mathbf{C}_{\mathbf{F}_a}$ with maximum size equal to $(m-1)|\mathbf{F}_a|$:

$$\mathbf{C}_{\mathbf{F}_a} = \bigcup_{b \in \{1,\ldots,m\} \setminus \{a\}} \mathbf{C}_{\mathbf{F}_a}^b$$

where $\mathbf{C}_{\mathbf{F}_a}^b$ indicates the set of CEs belonging to class b and generated from class a, namely $\mathbf{C}_{\mathbf{F}_a}^b = \{\boldsymbol{x}_{f_a}^{cf_b} \mid \boldsymbol{x}_{f_a} \in S_a \wedge \boldsymbol{x}_{f_a}^{cf_b} \in S_b\}$.

3.2.2 Numerical solution

Since obtaining an exact solution to the optimization problem in Eq. 3.14 is not always feasible due to the presence of non convex constraints, we instead employ an approximate solution. In such numerical approximation CEs are sought within sampled classification regions obtained by applying quasi-random Halton sampling [115], which is a low discrepancy sequence generator. Other generators, such as Sobol, may be applied in this sampling step.

The complete procedure is summarized in Figure 3.1. The preliminary step is *data classification*: the MC-SVDD algorithm is trained on a dataset $\mathcal{D}_{tr}$ and tuned on a dataset $\mathcal{D}_{vl}$, both drawn from the same underlying data distribution. Following hyperparameter tuning, the algorithm yields m optimal classification regions $S_1, S_2, \ldots, S_m$. The main step is *generation of CEs*: the MUCH algorithm (Algorithm 2) is applied to extract CEs from a set of factual observations drawn from a test set $\mathcal{D}_{ts}$ and predicted as belonging to class a, i.e., $\mathbf{F}_a \subseteq S_a$. First, for each classification region $S_i \in [m]$, a region $\tilde{S}_i$ is generated via quasi-random sampling. Then, for each factual $\boldsymbol{x}_{f_a} \in \mathbf{F}_a$, the counterfactual $\boldsymbol{x}_{f_a}^{cf_b}$, $b \neq a$, is identified as the point in $\tilde{S}_b$ that lies at minimum distance from $\boldsymbol{x}_{f_a}$. The choice of the distance function $\text{dist}(\cdot,\cdot)$ plays a key role in the search for CEs, as changing the distance can change the explanation returned. The most natural choice of distance is the distance induced by the classification kernel, that is:

$$dist(\boldsymbol{x}, \mathbf{y}) = K(\boldsymbol{x}, \boldsymbol{x}) - 2K(\boldsymbol{x}, \mathbf{y}) + K(\mathbf{y}, \mathbf{y})$$

This choice is motivated by the fact that the kernel-induced topology during classification influences the relationships between points in the sampled regions. Preserving these distance relationships helps the algorithm identify the most suitable counterfactual explanation.

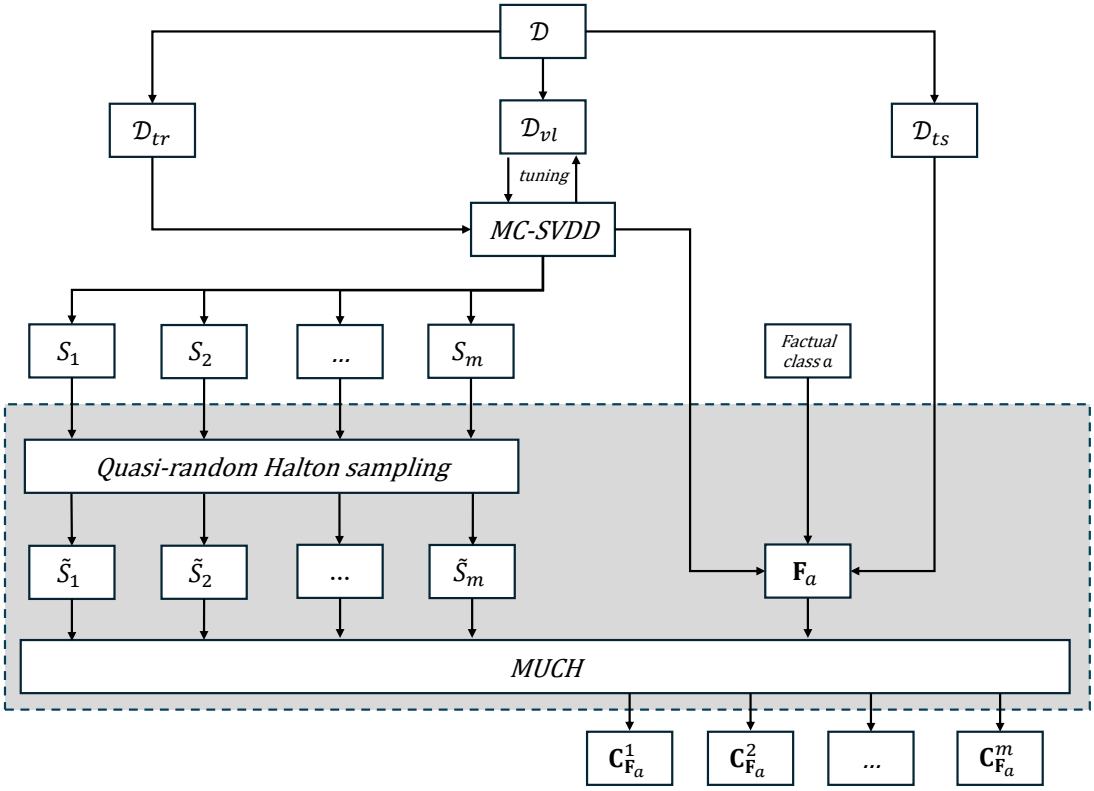


Figure 3.1: Extraction of counterfactual explanations using MUCH.

Algorithm 2 MUCH

Input: $S_1, S_2, \ldots, S_m$ bounded data regions from MC-SVDD, a test set $\mathcal{D}_{ts}$, a factual class of interest (a)

Output: A set $\mathbf{C}_{\mathbf{F}_a}$ of CEs for factual instances of class a

---

$\mathbf{F_a} = \{\boldsymbol{x} \in \mathcal{D}_{ts} \mid MC-SVDD(\boldsymbol{x}) = a\}$
$\mathbf{C_{F_a}} = [\,]$
for $i \in [m]$ **do**
  $\tilde{S}_i = HaltonSample(S_i)$
end for
for $b \in [m], b \neq a$ **do**
  $\mathbf{C}^b_{\mathbf{F}_a} = [\,]$
  for $\boldsymbol{x}_{f_a} = (\mathbf{u}, \mathbf{z})_{f_a} \in \mathbf{F_a}$ **do**
    $\mathbf{d}^b_a = dist\big(\boldsymbol{x}_{f_a}, \tilde{S}_{b_{|\mathbf{z}=\mathbf{z}_{\mathbf{f_a}}}}\big)$
    $\boldsymbol{x}^{cf_b}_{f_a} = \arg\min_{\boldsymbol{x} \in \tilde{S}_b} \mathbf{d}^b_a$
    $\mathbf{C}^b_{\mathbf{F}_a} = \mathbf{C}^b_{\mathbf{F}_a} \cup \{\boldsymbol{x}^{cf_b}_{f_a}\}$
  end for
  $\mathbf{C_{F_a}} = \mathbf{C_{F_a}} \cup \mathbf{C}^b_{\mathbf{F}_a}$
end for

Let n be the number of training data points and m be the number of classes. The total computational cost of the pipeline proposed in Figure 3.1 is given by two contributions: the cost associated with multi-class classification and the cost of the counterfactual explainer. The computational cost associated with MC-SVDD, denoted as $\mathcal{O}$(MC-SVDD), is primarily determined by the two most resource-intensive tasks: solving the quadratic programming problem to compute the Lagrangian multipliers and performing kernelization, which involves both the computation and storage of the kernel matrix. The computational complexity required to solve a quadratic programming typically ranges from $\mathcal{O}(l^3)$ to $\mathcal{O}(l^4)$, with l representing the number of variables, that is, the number of Lagrange multipliers ($\boldsymbol{\alpha}^{kk}$ and $\boldsymbol{\alpha}^{kh}$). The variation in complexity arises from the choice of the optimizer. The number l depends on the number of samples and classes, specifically

- $\boldsymbol{\alpha}^{kk}$: there are $n_1 + n_2 + \cdots + n_m$ Lagrange multipliers for each class k. These are associated with the data points that belong to class k;
- $\boldsymbol{\alpha}^{kh}$: there are $n_1n_2 + n_1n_3 + \cdots + n_1n_m + n_2n_3 + \cdots + n_2n_m + \cdots + n_{(m-1)}n_m$ Lagrange multipliers for each pair of classes (k, h).

Thus, the total number of Lagrange multipliers K can be calculated as follows:

$$l = (n_1 + n_2 + \cdots + n_m) + (n_1n_2 + n_1n_3 + \cdots + n_1n_m + n_2n_3 + \cdots + n_2n_m + \cdots + n_{(m-1)}n_m)$$

The computational cost for kernelization, instead, varies from kernel to kernel [116]. Focusing on the linear kernel, the cost can be estimated in $\mathcal{O}(n^2)$. However, complexity increases when considering a polynomial kernel ($\mathcal{O}(n^{2p})$, where p is the degree of the polynomial) or a Gaussian kernel ($\mathcal{O}(n^{2g})$, where g reflects the complexity due to the exponential function and the Euclidean distance) are considered. Regardless of the kernel choice, the kernelization cost does not overcome the computational cost for solving the quadratic optimization problem. Consequently, the total complexity of the MC-SVDD algorithm can be estimable solely based on the cost to compute the Lagrangian multipliers (i.e. $\mathcal{O}(l^3)$ or $\mathcal{O}(l^4)$).

As discussed in [117], the computational cost of the CEs search depends on several factors: the complexity of the quasi-random sampling, the effort required to compute the distance, the cost associated with solving the minimization problem, and the number of factual observations for which the CEs are to be retrieved. Hence, $\mathcal{O}$(MUCH) can be estimated in $\mathcal{O}\Big(A\Big(\max\Big(\sum_{b\neq a} q_b, |\mathbf{F}_a| \max\Big(\mathcal{C}_{dist}, \sum_{b\neq a} \mathcal{C}_{min,b}\Big)\Big)\Big)\Big)$, where $A \in [m]$ is the number of factual classes of interest, q_b is the computational cost of the random sampling of $\tilde{S}_b$ (references for its estimation can be found in [118]), $\mathcal{C}_{dist}$ is the computational cost for the computation of the distance dist$(\cdot, \cdot)$ [119], and $\mathcal{C}_{min,b}$ is the computational cost of the

minimization of the vector of distances $\mathbf{d}_a^b$ relative to the $b-$th random sampling ($\tilde{S}_b$) [120].

Finally, the total computational cost can be estimated with $\mathcal{O}$ (max(MC-SVDD, MUCH)).

3.2.3 Counterfactual conformity

Providing predictions that can guarantee a sufficiently high level of reliability is crucial in safety-critical areas like healthcare [121]. Likewise, offering reliable explanations is also fundamental to improve trustworthiness. With this aim, this Section introduces a measure of *counterfactual conformity* by adapting and expanding the concept of conformal predictions [103]. The rationale is to quantify the uncertainty of CEs with respect to their ideal properties. According to conformal predictions theory [122], once defined:

- a *calibration set* $\mathcal{D}_{cl} \in \mathcal{X}_{cl} \times \mathcal{Y}_{cl}$ with size n_{cl}[3];
- a desired *error level* $\varepsilon \in (0, 1)$;
- a real-valued *score function* $s : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}$ measuring how much a label y is *conformal* to the sample $\boldsymbol{x}$,

for any $\boldsymbol{x} \in \mathcal{X}$, we can determine a *prediction set* at *level of confidence* $1 - \varepsilon$:

$$\mathcal{C}_\varepsilon(\boldsymbol{x}) = \{\hat{y} \mid s(\boldsymbol{x}, \hat{y}) \leq s_\varepsilon\} \in 2^{\mathcal{Y}}, \tag{3.15}$$

where s_ε is the $\lceil (n_{cl}+1)(1-\varepsilon) \rceil / n_{cl}$ *quantile* of the score values computed on the calibration set. Hence, conformal prediction measures the uncertainty of predictions of machine learning models with a certain confidence level.

In this Section, we take cues from the conformal prediction framework to create a novel metric for assessing the quality of counterfactual explanations. This preliminary approach will not adhere strictly to the formalism of conformal predictions since we are interested in quantifying the uncertainty related to the generation of CEs rather than the prediction of a model. Specifically, given an instance $\boldsymbol{x}$, we substitute the "conformal label" with the "conformal CE", relying on the idea that a CE uniquely belongs to its target class. The definition of counterfactual conformity here introduced assumes that the quality of a CE can be measured considering [63]:

1. the distance between the CE and its factual (i.e., the smaller the distance, the better the CE, considering that the optimal CE should be, by definition, the minimal variation of the input parameters that realizes a change in the prediction label); and

[3]typically, the size of the calibration set is greater than 500 observations [122].

2. the distance between the CE and the corresponding counterfactual class (i.e., the smaller the distance, the more representative the counterfactual explanation is for the class).

The combination of these two requirements leads to a trade-off between the properties of *proximity* (i.e., the counterfactual explanation should be close to the classification boundary) and *plausibility* (i.e., the counterfactual explanation should be representative of the target class) properties defined in Section 2.1.2.

As a measure to assess, concurrently, the two properties, we define a score function as the weighted combination of the distances between the counterfactual explanation $\boldsymbol{x}_{f_a}^{cf_b}$ and its factual $\boldsymbol{x}_{f_a}$ and between the counterfactual explanation $\boldsymbol{x}_{f_a}^{cf_b}$ and the barycenter of the counterfactual class $\mathbf{X}_b^O$ (computed on the training set), respectively:

$$s(\boldsymbol{x}_{f_a}, \boldsymbol{x}_{f_a}^{cf_b}) = \tau \cdot \text{mix\_dist}(\boldsymbol{x}_{f_a}, \boldsymbol{x}_{f_a}^{cf_b}) \ + (1-\tau) \cdot \text{mix\_dist}(\boldsymbol{x}_{f_a}^{cf_j}, \mathbf{X}_b^O) \tag{3.16}$$

where $\tau \in (0,1)$ is a real valued weight and

$$\text{mix\_dist}(x,y) = \left(\frac{\alpha}{\alpha+\beta}\right) \cdot \text{Hamming}(x,y) \ + \left(\frac{\beta}{\alpha+\beta}\right) \cdot \text{Cosine}(x,y) \tag{3.17}$$

is a mixed distance borrowed from [71], with α being the number of categorical input features and β being the number of numerical input features. This distance consists of two components: the normalized Hamming distance for handling categorical features and the cosine distance for managing continuous features. Here, τ is set to 0.5 to give equal importance to the two contributions.

Then, we can define the *conformal counterfactual set* of factual $\boldsymbol{x}_{f_a}$ as the collection of all counterfactual explanations $\boldsymbol{x}_{f_a}^{cf_b}$ where the score value $s(\boldsymbol{x}_{f_a}, \boldsymbol{x}_{f_a}^{cf_j})$ does not exceed the approximate $(1-\varepsilon)-$quantile s_ε, as determined on the calibration set, i.e.

$$\mathcal{C}_\varepsilon(\boldsymbol{x}_{f_a}) = \{\boldsymbol{x}_{f_a}^{cf_b} \mid s(\boldsymbol{x}_{f_a}, \boldsymbol{x}_{f_a}^{cf_b}) \leq s_\varepsilon\}. \tag{3.18}$$

The term "fully conformal CE" applies when the computed counterfactual explanations for a certain factual $\boldsymbol{x}_{f_a}$ (i.e., $\boldsymbol{x}_{f_a}^{cf_b} \ \forall\, b \neq a$), adhere to the aforementioned conformity criterion. Conversely, the term "non-conformal CE" is used when none of the computed counterfactual explanations for a certain factual observation meet the aforementioned criterion. The term "partially conformal CE" is used in any other case. With this interpretation of conformity, an empirical error is made whenever the conformal set does not contain the counterfactual

explanation related to a certain class:

$$\text{err}_{a\to b} = \Pr\{\boldsymbol{x}_{f_a}^{cf_b} \notin \mathcal{C}_\varepsilon(\boldsymbol{x}_{f_a})|\varepsilon\} = \frac{\#\{\boldsymbol{x}_{f_a}^{cf_b} \notin \mathcal{C}_\varepsilon(\boldsymbol{x}_{f_a})|\varepsilon\}}{\#\{\boldsymbol{x}_{f_a}^{cf_b}\}} \tag{3.19}$$

3.2.4 Evaluation on open source benchmark datasets

Before assessing the potential of MUCH on clinical data for chronic disease prevention, the method was applied to frequently referenced multi-class open source datasets, including the FIFA dataset[4], the IRIS dataset[5], and the Stellar Classification Dataset - SDSS17[6]. Experiments were conducted to evaluate the performance of OvR MC-SVDD against other classifiers including OvO MC-SVDD and canonical machine learning methods (i.e., the multi-class generalization of the Support Vector Machine (MC-SVM), Decision Tree, Random Forest and Gradient Boosting) and to evaluate the quality of MUCH counterfactual explanations derived from MC-SVDD classification regions. The OvR MC-SVDD approach yielded higher classification performance on the FIFA dataset with respect to the OvO version, resulting in better plausibility of CEs (Table A.3). In general, the classification performance obtained by training OvR MC-SVDD on the three benchmark datasets was comparable to that of canonical classification algorithms, including the MC-SVM (Tables A.2, A.6, and A.9), Decision Tree, Random Forest, and Gradient Boosting (Tables A.7 and A.10), while presenting advantages in terms of capability to detect outliers, error control and definition of bounded regions of data points. Moreover, MUCH demonstrated satisfactory performance in terms of availability, proximity, discriminative power, and plausibility of the generated counterfactual explanations. These experiments demonstrated that the proposed approach can potentially scale well to tabular datasets of different nature and size, both in terms of the number of records and number of features. Without diverting focus from the application in the clinical context, such experiments are presented in detail in a dedicated appendix (Appendix A).

[4]Retrieved [November 2022] from https://www.kaggle.com/datasets/cashncarry/fifa-23-complete-player-dataset

[5]Retrieved [Dec 2022] from https://www.kaggle.com/datasets/uciml/iris

[6]Retrieved [Dec 2022] from https://www.kaggle.com/fedesoriano/stellar-classification-dataset-sdss17

3.3 Multi-Class Counterfactual Explanations for Chronic Disease Prevention

After introducing (Section 3.1-3.2) and assessing the MUCH method across benchmark datasets of different nature (Appendix A), this section presents a real-world applicative example of the use of CEs in the context of chronic disease prevention, leveraging EMR data from primary care.

Two distinct counterfactual explainers are compared, namely, MUCH (Section 3.2) and the model agnostic version of DiCE [65], briefly described in Section 2.1.2. MUCH and DiCE explainers share some common features, for example: the capability to handle tabular datasets with mixed data (either continuous or categorical), the possibility to specify a set of modifiable and non-modifiable features, and the capability to provide constraint during generation, that is, to provide a range of admissible values for each feature.

The aim of the study presented in this Section and published in [102] is to generate CEs that could serve as the basis for the future creation of personalized risk reduction strategies. The selected case study (*Case Study 1*) focuses on CEs aimed to reduce the risk of developing cardiovascular diseases (CVDs) in a cohort of patients diagnosed with Chronic Obstructive Pulmonary Disease (COPD).

3.3.1 Clinical context

COPD is a high prevalence inflammatory disease, characterized by airflow limitations, predominantly associated with tobacco smoke and other factors like environmental or occupational air pollution [123]. According to a recent systematic review [124], the global prevalence of COPD was estimated around 10.3% in individuals aged 30-79 years in 2019. Regardless of the severity of the disease, COPD patients are often prone to other comorbidities that may contribute to the worsening of their overall health status and increase the likelihood of hospitalization [123]. Examples of these comorbidities include CVDs, which emcompass various clinical conditions affecting the heart and blood vessels such as heart failure, myocardial infarction, stroke, and angina pectoris. Several studies have analyzed the relationship between COPD and CVDs in specific cohorts, highlighting an higher prevalence of CVDs in COPD patients in comparison to non-COPD ones [125]. For example, a meta-analysis including studies published between 1980 and 2015 [126] observed that COPD patients may present a two to five times higher likelihood of CVDs occurrence with respect to non-COPD patients. Moreover, an increased risk of acute CVDs manifestations has been observed after COPD exacerbations [127]. Currently, the presence of CVDs in patients with COPD

is mainly treated following general CVDs guidelines [128, 129]. However, the use of personalized strategies derived specifically in patients with COPD could lead to more effective disease prevention and patient management.

3.3.2 Study dataset

The study dataset was extracted ad hoc from the CPCSSN database (see Appendix B) and contains routinely collected biomarkers from patients diagnosed with COPD receiving primary care services, with the aim of estimating the individual 10-year CVD risk. Inclusion criteria required patients to be over 20 years of age, have a confirmed diagnosis of COPD, and have at least one medical encounter after COPD diagnosis. A 6-month tolerance window was allowed around the recorded diagnosis date to account for the gradual onset and progression typical of chronic diseases such as COPD. The following features were considered: age at COPD onset, sex assigned at birth, body mass index (BMI), systolic and diastolic blood pressure (sBP and dBP, respectively), fasting blood sugar (FBS), low-density lipoprotein (LDL), high-density lipoprotein (HDL), triglycerides (TG), total cholesterol (totChol), smoking (yes, no, ex-smoker), presence of hypertension and/or diabetes, if diagnosed up to 6 months before the diagnosis of COPD. The extracted biomarkers refer to the medical encounter closest to COPD diagnosis date. Values collected up to 6 months before COPD diagnosis were considered to account for possible uncertainty in the diagnosis date. Using these extraction criteria, a sample of 9,613 records (one record per subject) with no missing values was extracted from an initial set of 37,504 subjects diagnosed with COPD originally available in the CPCSSN database.

The output variable considered is the Framingham Risk Score (FRS), a sex-specific multivariable indicator that can be used to estimate the 10-year-risk of developing CVDs. The FRS for each subject was calculated, converted into a percentage risk value, and then grouped into three classes using the Framingham Risk Score Worksheet provided by the Canadian Cardiovascular Society [130]: *low* risk (10-year CVD Risk$<$ 10%, 3,944 records), *moderate* risk (10% $\leq$10-year CVD Risk$<$ 20%, 3,274 records), and *high* risk (10-year CVD Risk$\geq$ 20%, 2,395 records). A summary of the distribution of the dataset features as a function of the output risk class is shown in Table 3.1. Each feature was marked as modifiable, partially modifiable or not modifiable depending on its ability to be manipulated for the sake of risk reduction (e.g., through lifestyle interventions). For each modifiable feature, the maximum acceptable value shown in Table 3.1 is the value used as an upper bound when generating CEs. As it can be observed from the table, these values indicate cut-off values that normally determine a clinically relevant worsening of the patient's health status (e.g., class 2 obesity

or worse for BMI over 35 kg/m^2, hyperlipidemia for total cholesterol above 6.2 mmol/L). Partially modifiable features, like smoking habits, were permitted to vary only in certain directions. For example, an individual who smokes (Smoke="yes") may be able to cease smoking (Smoke = "ex") but cannot transition to the category Smoke = "no" which represents those who have never been smokers. Continuous features were normalized in the $[0, 1]$ range (min-max scaling).

Feature	***Low risk* (N=3,944)**	***Moderate risk* (N=3,274)**	***High risk* (N=2,395)**	**Modifiable**	**Maximum acceptable value**
Age [years]	56 (49-64)	66 (59-73)	73 (67-78)	No	/
Sex at birth	female: 76.32% male: 23.68%	female: 51.1% male: 48.9%	female: 15.03% male: 84.97%	No	/
HTN	no: 80.76% yes: 19.24%	no: 52.05% yes: 47.95%	no: 68.27% yes: 31.73%	No	/
Diabetes	no: 84.18% yes: 15.82%	no: 75.96% yes: 24.04%	no: 68.06% yes: 31.94%	No	/
Smoke	no: 21.78% ex: 27.43% yes: 50.79%	no: 21.65% ex: 33.78% yes: 44.56%	no: 23.51% ex: 40.12% yes: 36.37%	Partial	/
sBP [mmHg]	120 (110-128)	130 (122-140)	140 (130-150)	Yes	140
dBP [mmHg]	73 (68-80)	77 (70-82)	78 (70-84)	Yes	90
BMI [kg/m^2]	27.0 (23.0-32.5)	28.2 (24.3-32.7)	28.0 (25.0-32.0)	Yes	35
FBS [mmol/L]	5.2 (4.8-5.8)	5.5 (5.0-6.1)	5.6 (5.2-6.4)	Yes	7
LDL [mmol/L]	2.65 (2.00-3.33)	2.61 (1.93-3.35)	2.42 (1.78-3.20)	Yes	5
HDL [mmol/L]	1.44 (1.19-1.75)	1.30 (1.06-1.60)	1.34 (1.13-3.66)	Yes	2.5
TG [mmol/L]	1.21 (0.87-1.72)	1.35 (1.92-7.67)	1.39 (1.00-1.99)	Yes	5.7
totChol [mmol/L]	4.76 (4.02-5.50)	4.69 (3.88-5.56)	4.40 (3.56-5.33)	Yes	6.2

Table 3.1: Features distribution as a function of the output class, degree of modifiability, and maximum acceptable value. Numerical features: median (inter-quartile range); categorical features: percentage of samples for each category.

3.3.3 Methodological pipeline

Figure 3.2 provides an overview of the methodological pipeline followed in this Chapter. A custom study dataset, including patients diagnosed with COPD with varying CDV risk, was extracted as described in Section 3.3.2 and partitioned into train and test sets with a 70:30 ratio. After training and optimizing a multi-class classifier for CVD risk classification, CEs were generated from test set observations predicted as high risk of developing CVDs using two counterfactual explainers (MUCH and DiCE). MC-SVDD was used as the underlying classifier because, as shown in Section 3.1, it is flexible and easily controllable, inherently supporting outlier detection. Tuning of the MC-SVDD hyperparameters was performed by 3-fold cross-validation. Moreover, Algorithm 1 (previously formalized in Section 3.1.2) was applied to control the percentage of unclassified points in the MC-SVDD prediction. This method helps to derive CEs that are highly representative of the class they are meant to target, leveraging highly representative classification regions.

While the original DiCE implementation [65] supports non-differentiable classifiers built with common Python frameworks (e.g., scikit-learn, TensorFlow, or PyTorch), MC-SVDD is natively implemented in MATLAB. Consequently, it could not be directly used with DiCE. To minimize discrepancies between the underlying classifier used with MUCH and the one used with DiCE and to enable a fair, direct comparison of the results, a surrogate model that emulates the input/output behavior of the MC-SVDD was trained. The scikit-learn implementation of the MC-SVM classifier was chosen to surrogate the MC-SVDD, given their inherent similarities [131]. A grid search with 3-fold cross-validation was performed to tune the regularization parameter and the kernel of the SVM (best model parameters: C=7, gamma="auto", kernel="rbf"). Besides accuracy, the capacity of the SVM model to surrogate the original MC-SVDD model was assessed using the Cohen's Kappa coefficient [132]. This coefficient assumes values between -1 and 1, with -1 indicating total disagreement, 0 indicating random chance agreement and 1 indicating total agreement between the two models.

Evaluation of CEs was conducted considering availability, discriminative power, proximity, sparsity, plausibility, and diversity (see Section 2.1.2).

Availability and discriminative power were computed as in Section A.1. Availability was defined as the ratio of the number of CEs of class b generated from class a to the total number of factual observations of class a:

$$Availability_{a \to b} = \frac{| \mathbf{C}^{b}_{\mathbf{F}_a} |}{\mathbf{F}_a}\%$$

Discriminative power for a certain set of factual observations was estimated by evaluating

the macro averaged test accuracy of a k-Nearest Neighbor (KNN) classifier ($k = 5$, 5-fold cross-validation) in discriminating factual observations in $\mathbf{F}_a$ from CEs in $\mathbf{C}_{\mathbf{F}_a}$. Sparsity was computed as the average number of modifiable features changed between each available CE of class b generated from class a and its corresponding factual observation:

$$Sparsity_{a\to b} = \frac{1}{\mid \mathbf{C}^b_{\mathbf{F}_a} \mid} \sum_{\boldsymbol{x}_{f_a}\in\mathbf{F}_a:\ \exists \boldsymbol{x}^{cf_b}_{f_a}\in\mathbf{C}^b_{\mathbf{F}_a}} \sum_{var=1}^{|\mathbf{u}|} \Delta(\boldsymbol{x}_{f_a}, \boldsymbol{x}^{cf_b}_{f_a}, var)$$

$$\text{where } \Delta(\boldsymbol{x}_{f_a}, \boldsymbol{x}^{cf_b}_{f_a}, var) = \begin{cases} 1 & \text{if } var \text{ is numeric and } |\boldsymbol{x}_{f_a}(var) - \boldsymbol{x}^{cf_b}_{f_a}(var)| > \text{tol}, \\ 1 & \text{if } var \text{ is categorical and } \boldsymbol{x}_{f_a}(var) \neq \boldsymbol{x}^{cf_b}_{f_a}(var), \\ 0 & \text{otherwise.} \end{cases}$$

A tolerance $tol = 0.1$ was considered.

Proximity, plausibility, and diversity were determined using the mixed distance outlined in Eq. 3.17. Note that, since the distance used in this computation differs from that employed in Section A, the metric values obtained in the two sections are not directly comparable and should be interpreted separately. Proximity was computed as the average mixed distance between each available CE of class b generated from class a and its corresponding factual observation:

$$Proximity_{a\to b} = \frac{1}{\mid \mathbf{C}^b_{\mathbf{F}_a} \mid} \sum_{\boldsymbol{x}_{f_a}\in\mathbf{F}_a:\ \exists \boldsymbol{x}^{cf_b}_{f_a}\in\mathbf{C}^b_{\mathbf{F}_a}} \text{mix\_dist}(\boldsymbol{x}_{f_a}, \boldsymbol{x}^{cf_b}_{f_a})$$

Plausibility was computed as the average mixed distance between each available CE of class b generated from class a and the barycenter of a reference population:

$$Plausibility_{a\to b} = \frac{1}{\mid \mathbf{C}^b_{\mathbf{F}_a} \mid} \sum_{\boldsymbol{x}^{cf_b}_{f_a}\in\mathbf{C}^b_{\mathbf{F}_a}} \text{mix\_dist}(\boldsymbol{x}^{cf_b}_{f_a}, \mathbf{X}^O_b)$$

where $\mathbf{X}^O_b$ is the baricenter of the training set distribution with real output class b.

Diversity was calculated as the average mixed distance between each possible pair of CEs of class b generated from class a:

$$Diversity_{a\to b} = \frac{1}{\frac{|\mathbf{C}^b_{\mathbf{F}_a}|(|\mathbf{C}^b_{\mathbf{F}_a}|-1)}{2}} \sum_{\boldsymbol{x}^{cf_b}_{f_a}} \sum_{\boldsymbol{x}^{cf'_b}_{f_a}} \text{mix\_dist}(\boldsymbol{x}^{cf_b}_{f_a}, \boldsymbol{x}^{cf'_b}_{f_a})$$

where $\boldsymbol{x}^{cf'_b}_{f_a} \neq \boldsymbol{x}^{cf_b}_{f_a}$ and $\boldsymbol{x}^{cf_b}_{f_a}, \boldsymbol{x}^{cf'_b}_{f_a} \in \mathbf{C}^b_{\mathbf{F}_a}$.

Finally, a novel measure for assessing conformity of CEs (Section 3.2.3) was computed to discard explanations that did not reach the desired level of confidence. The score function for counterfactual conformity was calibrated on 80% of the test set and the error was computed on the remaining 20%. Calibration was performed separately for MUCH and DiCE.

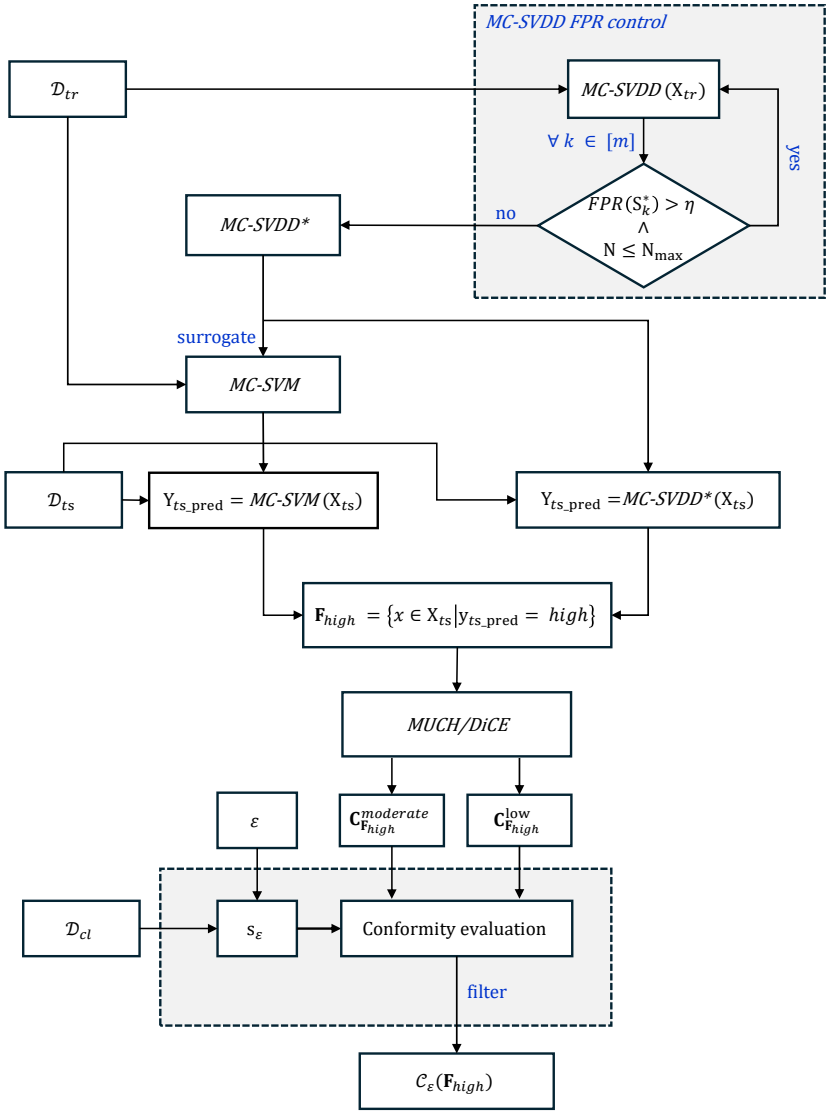


Figure 3.2: Methodological workflow: multi-class classification, generation, and evaluation of CEs for CVDs risk reduction.

The Wilcoxon Signed-Rank Test for paired samples was applied to assess possible statistical differences between CEs and the corresponding factual observations, whereas the Mann-Whitney U test was used to assess possible differences in CEs generated with the two methods. The same test was used to compare CEs generated from subpopulations of patients with/without comorbidities (hypertension, diabetes). A significance level $\alpha = 0.05$

was considered for statistical comparisons and Bonferroni correction was applied to correct for multiple comparisons.

3.3.4 Results

Multi-class classification performance

The MC-SVDD classifier achieved an improved accuracy of 85.6% on the training set after applying FPR control (Algorithm 1) with a 10% of unclassified points, compared to 76.0% accuracy with only 0.07% unclassified points before FPR control. The classification performance on the test set was slightly lower yet satisfactory, achieving 78.6% accuracy with 11.1% of unclassified points, compared to 70.2% accuracy (4.2% of unclassified points) before FPR control. The class-specific sensitivity after FPR control was equal to 88.2% for low risk, 75.0% for moderate risk, and 95.9% for high risk on the training set; 83.3% for low risk, 69.0% for moderate risk, and 83.4% for high risk on the test set. The approach taken by FPR control proved to be more reliable because it opted to not classify data points rather than misclassify them (before control: $\text{FP}_{low} = 49$, $\text{FP}_{moderate} = 212$, $\text{FP}_{high} = 1342$; after FPR control: $\text{FP}_{low} = 41$, $\text{FP}_{moderate} = 125$, $\text{FP}_{high} = 103$). The surrogate MC-SVM model achieved high accuracy in predicting the output of the MC-SVDD (96.9% accuracy on the training set and 92.6% accuracy on the test set). The Cohen's Kappa coefficient was equal to 0.89, suggesting a satisfactory level of agreement between the MC-SVDD and its surrogate model.

Evaluation of counterfactuals explanations

The set of factual observations here considered included only those elements of the test set that were predicted as belonging to the high-risk class by the underlying classifier. This yielded a factual set $\mathbf{F}_{high}$ comprising 682 test samples for MUCH and 690 for DiCE, performed on top of MC-SVDD and its surrogate MC-SVM, respectively. For each factual $\boldsymbol{x}_{f_{high}} \in \mathbf{F}_{high}$, two CEs were generated: one representing a shift from the high- to the moderate-risk class ($\boldsymbol{x}_{f_{high}}^{cf_{moderate}}$), and another from the high- to the low-risk class ($\boldsymbol{x}_{f_{high}}^{cf_{low}}$).

Table 3.2 shows the performance of MUCH and DiCE in terms of availability, discriminative power, proximity, sparsity, plausibility and diversity, defined as in Section 3.3.3. The two methods yielded a high percentage of explanations despite constraints in the generation process, with MUCH having an average availability of 84.6% and DiCE reaching 98.2%. Both methods produced CEs that could be discriminated from points of the factual class with a satisfactory level of accuracy (i.e., discriminative power >77%, with MUCH performing better than DiCE). MUCH was slightly superior than DiCE in terms of plausibility, and

	Availability↑	**Discr Power↑**	**Proximity↓**	**Sparsity↑**	**Plausibility↓**	**Diversity↑**
			MUCH			
$\boldsymbol{x}_{f_{high}}^{cf_{moderate}}$	100.0%	94.6%	0.080	0.590	0.629	0.551
$\boldsymbol{x}_{f_{high}}^{cf_{low}}$	69.1%	98.8%	0.092	0.454	0.643	0.557
			DiCE			
$\boldsymbol{x}_{f_{high}}^{cf_{moderate}}$	98.0 %	77.0 %	0.002	0.792	0.749	0.545
$\boldsymbol{x}_{f_{high}}^{cf_{low}}$	98.4 %	92.3%	0.009	0.658	0.757	0.549

Table 3.2: Quality measures computed on counterfactual explanations generated with MUCH and DiCE methods: full set of explanations. ↑: Higher values indicate better quality; ↓: Lower values indicate better quality.

	Error			**Size**		
	Average error	$\text{err}_{High \to Moderate}$	$\text{err}_{High \to Low}$	Non conformal	Partially conformal	Fully conformal
			MUCH			
$\varepsilon = 0.01$	0.006	0.000	0.011	0.000	0.011	0.989
$\varepsilon = 0.05$	0.050	0.044	0.056	0.022	0.056	0.922
$\varepsilon = 0.10$	0.117	0.122	0.111	0.067	0.100	0.833
			DiCE			
$\varepsilon = 0.01$	0.011	0.015	0.008	0.008	0.008	0.985
$\varepsilon = 0.05$	0.050	0.053	0.046	0.046	0.008	0.947
$\varepsilon = 0.10$	0.141	0.160	0.122	0.084	0.114	0.801

Table 3.3: Error and size of the non-conformal, partially-conformal, and fully conformal sets at varying desired error levels (ε).

diversity. Conversely, DiCE exhibited better proximity and sparsity compared to MUCH. Counterfactual explanations in the moderate risk class had worse discriminative power and slightly worse diversity but better proximity, sparsity and slightly better plausibility than those in the low risk class, for both methods.

Table 3.3 summarizes the error and size of the non-conformal, partially conformal and fully conformal CEs sets obtained with MUCH and DiCE, as a function of ε. From the first column of the table, we can notice that both the algorithms were well calibrated since the average error computed on the evaluation set (i.e., 20% of the test set) was close to the desired error level ε, hence representing a quasi-linear relationship. According to our definition of counterfactual conformity (Section 3.2.3), the higher the number of fully conformal CEs

the more reliable the counterfactual extraction procedure is. Both methods here considered produced a sufficiently high number of fully conformal counterfactuals for small values of ε, meaning that both counterfactual explanations ($\boldsymbol{x}_{f_{high}}^{cf_{low}}$ and $\boldsymbol{x}_{f_{high}}^{cf_{moderate}}$) were representative of the target class while maintaining, by definition, also a minimal distance from the factual. In the following analysis, $\varepsilon = 0.1$ was selected as a compromise between the severity of the conformal check and the number of retained CEs. Furthermore, in healthcare applications such as the one presented here, the use of a higher ε (i.e., a more selective filtering process with respect to counterfactual explanations) might assist in identifying more realistic explanations with regard to the necessary changes in features to determine a change in output class.

In Table 3.4, conformal and non-conformal counterfactual explanations are compared in terms of desired properties. Conformal explanations exhibited superior quality in comparison to non-conformal ones (i.e., lower proximity and plausibility, higher diversity and sparsity). Non-conformal explanations demonstrated higher discriminative power, which can be attributed to their greater distance from the factual points (i.e., poorer proximity), thereby making them more readily distinguishable from the factual points. The comparison between the entire set of retrieved counterfactual explanations (Table 3.2) and the conformal explanations (Table 3.4) shows improved quality after discarding non-conformal explanations, as suggested by the values of proximity, sparsity and plausibility observed, while diversity and discriminative power remained similar.

Regarding the use of CEs for the reduction of CVD risk in patients with COPD, a closer inspection revealed that non-conformal explanations were primarily associated with unrealistically high changes in feature values compared to conformal ones. As an example, Table 3.5 presents one conformal and one non-conformal factual-counterfactual pair (high risk to low risk transition) generated using MUCH. The two factuals shown in Table 3.5 describe male patients who are overweight, are aged between 60-65 years, and are diagnosed with diabetes and chronic hypertension. Notably, the non-conformal CE (Ex2) is associated with higher changes in feature values compared to the conformal one (Ex1), some of which are unrealistic. For example, a lift in BMI from Class 1 obesity to Class 2 obesity and an increase in triglycerides are usually associated with increased CVD risk, and a decrease of about 40 mmHg in systolic blood pressure can be difficult to achieve from a clinical point of view.

Figure 3.3 shows the distributions (median, 25% and 75% percentiles) of the average changes requested by MUCH and DiCE to pass from the high risk class to the moderate risk class (panel 3.3a) and to the low risk class (panel 3.3b). To ensure a fair comparison between the two methods, only common factuals and only fully conformal counterfactuals

	Type	Discriminative Power ↑	Proximity↓	Sparsity↑	Plausibility↓	Diversity↑
			MUCH			
$\boldsymbol{x}_{f_{high}}^{cf_{moderate}}$	Non Conformal	99.49%	0.133	0.559	0.925	0.004
	Fully Conformal	94.34%	0.077	0.591	0.576	0.545
$\boldsymbol{x}_{f_{high}}^{cf_{low}}$	Non Conformal	99.85%	0.143	0.461	0.932	0.002
	Fully Conformal	98.43%	0.080	0.454	0.599	0.558
			DiCE			
$\boldsymbol{x}_{f_{high}}^{cf_{moderate}}$	Non Conformal	97.31%	0.003	0.736	1.151	0.227
	Fully Conformal	77.27%	0.002	0.801	0.692	0.526
$\boldsymbol{x}_{f_{high}}^{cf_{low}}$	Non Conformal	98.47%	0.012	0.613	1.162	0.229
	Fully Conformal	92.65%	0.009	0.664	0.700	0.529

Table 3.4: Quality of counterfactual explanations generated with MUCH and DiCE methods: fully conformal vs non-conformal explanations ($\varepsilon = 0.1$). ↑: Higher values indicate better quality; ↓: Lower values indicate better quality.

		sBP [mmHg]	dBP [mmHg]	BMI [kg/m^2]	FBS [mmol/L]	LDL [mmol/L]	HDL [mmol/L]	TG [mmol/L]	totChol [mmol/L]
Ex1	$\boldsymbol{x}_{f_{high}}$	150	78	31.9	8.0	4.7	0.9	2.7	6.8
	$\boldsymbol{x}_{f_{high}}^{cf_{low}}$	124	72	29.7	6.9	4.5	2.0	2.5	3.2
Ex2	$\boldsymbol{x}_{f_{high}}$	138	72	32.1	8.5	3.0	1.1	1.3	4.8
	$\boldsymbol{x}_{f_{high}}^{cf_{low}}$	97	81	37.7	6.2	1.37	2.1	2.8	1.3

Table 3.5: Examples of conformal (Ex1) and non-conformal (Ex2) CEs generated using MUCH and setting $\varepsilon = 0.1$.

were considered (N=337, $\varepsilon = 0.1$). The Figure shows that CEs generated with MUCH and DiCE differ in terms of changes in modifiable characteristics requested for moving from high to moderate risk and from high to low risk, with MUCH suggesting larger variations than DiCE. Statistically significant changes (i.e., variations in feature values statistically different from 0) were observed in transitions from high to moderate risk in terms of sBP, BMI, LDL, HDL, TRIG, totChol, and FBS for MUCH, and in terms of sBP, BMI, HDL, TRIG, and totChol for DiCE. In transitions from high to low risk, statistically significant changes in terms of all modifiable features, except FBS and LDL were observed using MUCH, and

statistically significant changes in terms of all modifiable features were observed using DiCE.

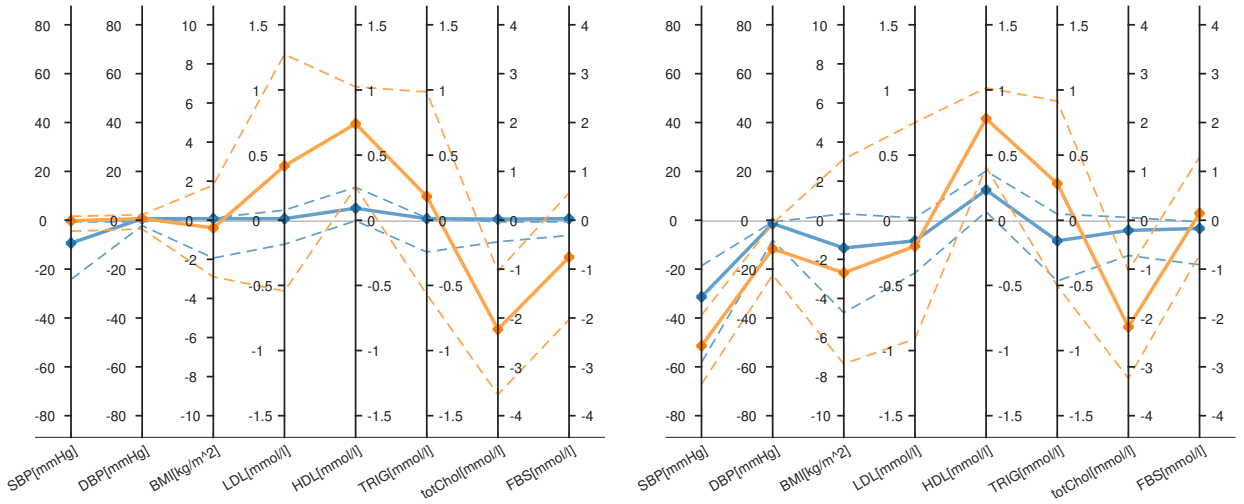
CEs obtained with MUCH were statistically different from those obtained with DiCE for sBP, LDL, HDL, TRIG, totChol, and FBS when analyzing high to moderate risk transitions (Figure 3.3a) whereas all features distributions except for BMI and LDL were statistically different when comparing MUCH and DiCE in terms of high to low transitions (Figure 3.3b). The changes suggested by MUCH to reduce the risk class were associated, on average, with a reduction in sBP and dBP, BMI, and totChol and an increase in HDL. These trends are coherent with a general improvement in the patients' health status and a reduction in cardiovascular risk, i.e., decreased blood pressure, lower weight, and better lipidic profile.

CEs generated by the two methods were also consistent with individual characteristics, for example in relation to comorbidities (diabetes and hypertension). Specifically, by comparing conformal CEs with their factual observations, we observed a greater median decrease in sBP for patients with stage 2 hypertension compared to non-hypertensive ones considering both high to moderate risk transitions (21.00 mmHg higher with DiCE, $p = 3.48 \times 10^{-38}$; 1.91 mmHg higher with MUCH, $p = 1.83 \times 10^{-9}$) and high to low risk transitions (39.00 mmHg higher with DiCE, $p = 2.44 \times 10^{-33}$; 25.77 mmHg higher with MUCH, $p = 1.43 \times 10^{-29}$). Similarly, a significantly higher median decrease in FBS is observed for diabetic patients compared to non diabetic ones considering both high to moderate (0.90 mmol/L higher with DiCE, $p = 7.70 \times 10^{-12}$; 1.17 mmol/l higher with MUCH, $p = 2.50 \times 10^{-7}$) and high to low transitions (1.40 mmol/L higher with DiCE, $p = 1.48 \times 10^{-10}$; 1.09 mmol/l higher with MUCH, $p = 8.89 \times 10^{-11}$).

3.4 Discussion

The work presented in this Chapter builds upon and substantially expands preliminary research conducted during the research fellowship preceding my PhD. More specifically, it aims to formalize a novel pipeline based on multi-class generalization of the SVDD (MC-SVDD, formalized in Section 3.1.1) and a counterfactual explainer (MUCH algorithm, formalized in Section 3.2) extending the previous approach for binary classifiers proposed in [117], which was originally applied to T2D characterization and prediction [133]. Beyond extending the framework from binary to multi-class settings, a more systematic evaluation across four tabular datasets with different nature, size, and degree of complexity has been pursued.

Section 3.3 provides an applicative example of the use of multi-class counterfactual explanations as an original, interpretable data-driven method to support the design of tailored disease prevention strategies. When applied to real-world clinical data, such analyses have practical implications for the end-user as they can assist clinicians in understanding which



(a) Transition from high to moderate risk.

(b) Transition from high to low risk.

Figure 3.3: Distributions of conformal counterfactual explanations ($\varepsilon = 0.1$) simulating transitions from high to moderate (3.3a) and from high to low (3.3b) CVD risk, obtained using MUCH (in orange) and DiCE (in blue), respectively. Solid lines: medians of the distributions; dashed lines: 25% and 75% percentiles.

clinical feature to target to reduce risk of developing chronic disease as shown for the CPC-SSN Case Study 1 dataset (Section 3.3.2). More specifically, the proposed methodological framework was applied to estimate personalized recommendation for reducing the 10-year CVD risk in COPD patients.

The proposed framework consists of two main components: (i) a multi-class classifier, predicting 10-year CVD risk and (ii) a counterfactual explainer.

Although other machine learning methods may be slightly more accurate in the classification task (e.g., Decision Tree=87.0%, Gradient Boosting=90.0%, and MC-SVM=91.0%, compared to 86.5% for the implemented MC-SVDD), the proposed framework benefits from additional properties that are more naturally supported by MC-SVDD classifier. First, MC-SVDD yields bounded decision regions. Each class is represented by a compact hypersphere in the feature space, facilitating the generation of meaningful counterfactual explanations using MUCH and simplifying the optimization problem. In contrast, other machine learning methods (e.g., logistic regression or MC-SVM with linear or polynomial kernels) produce unbounded decision regions, complicating the interpretation of class membership. Second, MC-SVDD inherently supports outlier detection, as samples that fall outside all decision regions are labeled as out-of-distribution instances. Other machine learning methods such as decision tree and MC-SVM assign every sample to a class, often requiring additional mechanisms for outlier handling. Third, MC-SVDD allows explicit control over the FPR by simply adjusting the radius, directly regulating the trade-off between data coverage and

compactness. For such reasons, MC-SVDD was selected as a suitable modeling approach for developing a reliable multi-class disease prevention framework, enabling the identification of well-defined risk regions and providing clinicians with insights into cases with low prediction confidence.

The proposed MC-SVDD served as the underlying classifier for generating CEs using two different counterfactual explainers (MUCH and the model-agnostic version of DiCE), both evaluated in terms of quality of their explanations. Furthermore, a novel counterfactual conformity metric, inspired by the conformal prediction framework, was introduced to increase reliability for the end user by providing a confidence value for each explanation produced and enabling rejection of non-conformal explanations.

The MUCH counterfactual explainer proposed in this Chapter can be categorized as an *exogenous* explainer [63] since it generates explanations that do not necessarily belong to the original dataset, and CEs are retrieved by addressing an optimization problem. The continuous optimization problem is approximated with a quasi-random sampling–based optimization, where the objective function is evaluated on a low-discrepancy set of feasible points sampled within classification regions. Since CEs are sought among a finite set of points, the availability and proximity of each explanation depend on the density of the sampling. However, the higher the number of points in the sampled region, the higher the computational cost (see Section 3.2.2). As a consequence, a trade-off between performance and runtime must be reached. Although applied here to MC-SVDD, the MUCH algorithm is in principle agnostic to the classifier to be used, as long as the classifier can provide m bounded decision regions (e.g., rule based methods, K-Nearest-Neighbors).

Focusing on prevention purposes, CEs were generated to guide transitions from high to moderate or low CVD risk, as estimated by the Framingham risk score. Two types of constraints were imposed during CEs generation: (i) non-modifiability of a subset of features and (ii) rejection of candidate explanations with one or more features exceeding the maximum acceptable values in Table 3.1. Imposing such constraints was essential to promote actionability and feasibility when working with real-world data, and to ensure that the generated explanations aimed to improve, or at least not worsen, the patient's health status. However, such design choice reduced availability with respect to unconstrained generation.

Counterfactual explanations retrieved from a set of common factual observations using MUCH and DiCE showed differences in terms of suggested changes for most of the features. The higher changes and higher discriminative power observed using MUCH (Figure 3.3, Table 3.2) are largely attributable to the reliance of MUCH on the shape of the classification regions provided by MC-SVDD, which are further refined and narrowed by the FPR control algorithm. In general, DiCE had better proximity (i.e., the generated CEs were closer to

the factual) compared to MUCH, and this metric further improved once the non-conformal counterfactuals were filtered out, keeping only the conformal ones (Table 3.4). The better proximity was reflected in the less pronounced variation trend, especially in transitions from high to moderate risk (Figure 3.3a). DiCE provided a higher availability of explanations in the low risk class whereas the two methods had similar availability of explanations in the moderate class. It is worth noting that, despite the high agreement between MC-SVDD and its surrogate MC-SVM (Cohen's Kappa coefficient equal to 0.89), the comparison between MUCH and DICE may be slightly influenced by the differences between the two underlying classifiers.

The use of CEs in clinical applications holds potential as a data-driven method for identifying personalized minimum viable changes to decrease the individual risk [133]. However, there is not a one-size-fits-all solution as there may be differences in the explanations generated using different algorithms that should be evaluated by the physician on a case-by-case basis, based on patient characteristics and clinical feasibility. For example, some CEs might suggest a significant change in biomarkers (e.g., blood pressure, BMI, triglycerides, as shown in the Example E2 in Table 3.5) that could be deemed unrealistic or unfeasible to achieve in practice, even with the help of medications and intensive lifestyle interventions. To facilitate a semi-automatic approach for selecting CEs counterfactual conformity was introduced as a novel quality metric for filtering out explanations that were not compliant with the desired properties. The definition of a score function as in (3.16) combined the measurement of two key properties: proximity and plausibility. Analyzing how the produced explanations were statistically distributed with respect to these properties can help understand the *global* quality of the generated counterfactual explanations. Furthermore, by using counterfactual conformity, each explanation was accompanied by a *local* reliability value. Table 3.4 effectively shows that CEs deemed conformal exhibit better quality compared to non-conformal ones in terms of desired properties. These analyses demonstrate how the newly introduced metric can improve reliability of the proposed AI system, providing physicians with additional information to determine whether to consider or discard the specific output.

Chapter 4

Expert-driven causal learning for T2D prevention based on static observational data

This Chapter focuses on causal learning methods for treatment effect estimation in the context of chronic disease prevention, with a specific application to Type 2 Diabetes (T2D) prevention (*Case Study 2*). This research activity was conducted within the framework of the Horizon Europe project PRAESIIDIUM[1], which seeks to combine AI tools and multiscale mathematical models to forecast transitions between normoglycemia, prediabetes, and T2D, and to determine personalized recommendations to lower individual risk.

The Chapter provides two main contributions, that can be summarized as follows:

- uncover key causal relationships relevant for T2D prevention from a large set of routinely collected primary care EHR data. The causal model is learned directly from data and iteratively refined using prior expert knowledge to unambiguously determine the orientation of relationships between variables [134];
- provide an illustrative example that demonstrates how counterfactual inference, applied to a causal model that incorporates latent factors, can be used to predict the effects of a hypothetical lifestyle intervention (improvement in diet and physical activity, exemplified through a decrease in BMI and FPG) on T2D risk. This framework may ultimately help clinicians in decision making and support the design of tailored prevention programs by simulating "what-if" scenarios that can quantify the predicted

[1]Physics Informed Machine Learning-Based Prediction and Reversion of Impaired Fasting Glucose Management (PRAESIIDIUM), HORIZON-HLTH-2022-STAYHLTH-02, Grant 101095672. Website:https://praesiidium.spindoxlabs.com/

benefits of specific interventions for individual patients.

4.1 Outline on causal models for T2D prevention

Diabetes mellitus is a chronic metabolic disorder characterized by elevated blood glucose levels (hyperglycemia) resulting from impaired insulin secretion or reduced insulin effectiveness [135]. The diagnosis of this condition relies on specific glycemic thresholds, such as those defined by the American Diabetes Association (ADA, [136]). According to the International Diabetes Federation, around 589 million adults aged 20 to 79 were living with diabetes in 2024 [135], with T2D being the most common form, accounting for over 90% of all diabetes cases worldwide.

In T2D, insulin secretion capacity is preserved, but the body's cells do not respond properly to the hormone, leading to a condition known as insulin resistance. This chronic condition can lead to several complications, such as CVD, retinopathy, neuropathy, peripheral artery disease, and foot ulcerations, and ultimately increase the risk of mortality. T2D onset is often completely asymptomatic, a factor that likely contributes to elevated mortality rates and increased healthcare costs. Around the globe, approximately 251.7 million individuals are living with undiagnosed diabetes [135]. In other words, nearly one in every two adults (42.8%) with diabetes are unaware of their condition. Therefore, early diagnosis and treatment are pivotal actions to reduce the risk of related complications [137]. For this reason, large-scale prevention programs have been proposed by national health care services (e.g., [138]). Despite such effort, the planning of large-scale prevention programs, including lifestyle modification programs, often faces organizational and financial challenges, as well as heterogeneity in treatment effectiveness [139].

Leveraging large sets of clinical data to develop AI-based risk prediction models can support the implementation of personalized prevention and treatment strategies [140]. Indeed, in the literature, there are numerous predictive models for T2D that can provide a risk score, as documented is some recent review articles [25, 141, 142]. Besides traditional predictive learning, causal learning methods can be applied to EHRs, to help identify the modifiable risk factors that contribute most prominently to the development of the disease and to predict the response of single individuals—or subgroups of individuals with common characteristics—to a given intervention [31]. Despite the potential usefulness of causal learning techniques in supporting decision-making, their application in the context of T2D prevention remains limited. For example, Wang et al. [143] identified a gap in population-based cohort studies focusing on the identification of causal pathways between known risk factors and T2D onset. To address this gap, they developed a causal model of the occurrence of T2D using prospec-

tive data collected from a Chinese population of 15,934 patients in the Guangzhou Biobank Cohort Study. Shen et al. [37] proposed a causal discovery method to extract causal relationships associated with T2D using a large retrospective EHRs dataset comprising over 57,000 patients from Mayo Clinic. The dataset included information on patient demographics, laboratory results, drug prescriptions, and diagnoses. Kalia et al. [144] applied a marginal structural model to Canadian primary-care EHRs to emulate a randomized trial and estimate the effectiveness of sodium-glucose cotransporter 2 inhibitor medications on reducing glycated hemoglobin (HbA1c) in patients with diagnosed diabetes. Although these studies focused on identifying causal relationships underlying T2D onset, they did not address the possible influence of latent confounders, such as unobserved modifiable risk factors, on the recovered causal model. Moreover, although tailored interventions may improve T2D prevention [139], the use of causal models and counterfactual inference to identify personalized lifestyle recommendations remains largely unexplored.

4.2 Study dataset

The dataset for this study was extracted ad hoc from the CPCSSN database (see Appendix B), using the following procedure. Patients affected by T2D were extracted from an initial cohort of 90,728 patients with a generic diagnosis of diabetes (including both type 1 and type 2 diabetes) using the following inclusion criteria: detection of billing and diagnostic information specifically related to T2D, prescriptions of first- and/or second-line antidiabetic drugs in their EHR, and age at diagnosis greater than 20 years (as defined in [145]). The remaining subjects were labeled as non-diabetic. Patients under 18 years of age at the time of their first primary care encounter and patients with a diagnosis of type 1 diabetes or gestational diabetes were excluded. The following routinely available variables were extracted for each patient:

- general characteristics: patient ID, age, sex assigned at birth, body mass index (BMI);
- smoking status;
- biomarkers: systolic and diastolic blood pressure (sBP and dBP), fasting plasma glucose (FPG), low- and high-density lipoprotein (LDL and HDL), triglycerides (TG), total cholesterol (TC);
- comorbidities: diagnosis of hypertension (HTN), osteoarthritis (OA), chronic obstructive pulmonary disease (COPD), and depression, and related diagnosis date;

- medication prescriptions: antihypertensives, cholesterol lowering medications, corticosteroids, quit-smoking medications, and antidepressants, and related prescription date. Table 4.1 describes the Anatomical Therapeutic Chemical (ATC) code of each extracted group of medications.

Medication group	**ATC code**
Cholesterol lowering medications	C10 (lipid modifying agents)
Antihypertensives	C03 (diuretics), C07 (beta blocking agents), C08 (calcium channel blockers) and C09 (agents acting on the renin-angiotensin system)
Corticosteroids	H02 (corticosteroids for systemic use)
Quit-smoking medications	N07BA (drugs used in nicotine dependence)
Antidepressants	N06AX12 (bupropion)

Table 4.1: Medication groups with their corresponding Anatomical Therapeutic Chemical (ATC) code.

General characteristics, biomarkers, and medications were extracted for each patient in a primary observation window of one year. A patient was included in a specific medication group if two or more prescriptions were found in the primary observation window. Medications with a single prescription were excluded as they could be related to a discontinued therapy and, therefore, unreliable information. For patients with T2D, the primary observation window was defined based on the most recent encounter that occurred up to six months before the diagnosis of T2D, to account for a possible mismatch between the EHR-recorded diagnosis and the actual T2D onset date. The median distance between the most recent encounter date and diagnosis date was 1.59 years (25% percentile = 1.03 years; 75% percentile = 2.68 years). For non-diabetic patients, the last encounter available in the database observation period was used to define the extraction window. The presence of comorbidities was evaluated before the primary observation window to preserve temporal ordering between comorbidity onset and observed biomarkers. The most recent information about smoking status was also considered, when available.

The full data extraction pipeline is reported in Figure 4.1. The training set is summarized in Tables 4.2 and 4.3 and includes 86,618 records with no missing values, related to 84,340 non-diabetic patients (T2D=0) and 2,278 patients with future onset of T2D (T2D=1). Continuous features are represented in terms of median value and interquartile range (IQR), whereas categorical features are represented in terms of percentage of samples for each category. An independent test set of 31,864 records (T2D=0: 30,985 patients; T2D=1: 879 patients) was extracted following the same extraction procedure and used as a basis for the evaluation of the retrieved causal structure and for counterfactual inference. Both the

training and test sets were discretized in the final step of the data curation phase to enable use with the selected causal learning algorithms. Discretization thresholds, presented in the right-most column of Table 4.2, were determined in accordance with physicians and based on clinical reference guidelines to reflect standard and interpretable risk categories (e.g., for "Pressure": normal pressure, elevated, stage I and stage II hypertension). Clinically informed discretization was preferred over other kind of discretization methods (e.g., quantile-based) that could include subjects with different clinical risk within the same level, reducing interpretability of the results.

Feature	**T2D=0 (N= 84,340)**	**T2D=1 (N= 2,278)**	**Levels**
Age	53 (44-64)	57 (49-66)	0: Age $<$ 45 1: 45 $\leq$ Age $<$ 65 2: Age $\geq$ 65
BMI [kg/m^2]	26.9 (23.9-30.6)	31.8 (28.0-36.3)	0 : BMI $<$ 18.5 1: 18.5 $\leq$ BMI $<$ 25 2: 25.0 $\leq$ BMI $<$ 30 3: 30.0 $\leq$ BMI $<$ 35 4: 35.0 $\leq$ BMI $<$ 40 5: BMI $\geq$ 40
Pressure [mmHg]	SBP : 123.0 (113.0 – 132.0) DBP : 76.0 (70.0 – 82 – 0)	SBP : 130.0 (120.0 – 139.0) DBP : 79.8 (73.5 – 84.9)	0: SBP $<$ 120 & DBP $<$ 80 1: 120 $\leq$ SBP $\leq$ 129 & DBP $<$ 80 2: 130 $\leq$ SBP $\leq$ 139 or 80 $\leq$ DBP $\leq$ 89 3 : SBP $\geq$ 130 or DBP $\geq$ 90
FPG [mmol/L]	5.1 (4.7-5.4)	5.9 (5.5-6.4)	0: FPG $<$ 5.6 1: 5.6 $\leq$ FPG $<$ 7.0 2: FPG $\geq$ 7.0
HDL [mmol/L]	1.41 (1.17-1.72)	1.17 (1.00-1.39)	0: HDL $<$ 1 (men) or HDL $<$ 1.3 (women) 1: 1 $\leq$ HDL $<$ 1.6 (men) or 1.3 $\leq$ HDL $<$ 1.6 (women) 2: HDL $\geq$ 1.6
LDL [mmol/L]	2.91 (2.35-3.51)	2.9 (2.27-3.50)	0: LDL $<$ 2.6 1: 2.6 $\leq$ LDL $<$ 3.4 2: 3.4 $\leq$ LDL $<$ 4.2 3: 4.2 $\leq$ LDL $<$ 5.0 4: LDL $\geq$ 5.0
TG [mmol/L]	1.11 (0.79-1.58)	1.66 (1.19-2.32)	0: TG $<$ 1.7 1: 1.7 $\leq$ TG $<$ 2.3 2: 2.3 $\leq$ TG $<$ 5.7 3: TG $\geq$ 5.7
Total Cholesterol [mmol/L]	4.97 (4.33-5.66)	4.93 (4.25-5.67)	0: Tot chol $<$ 5.18 1: 5.18 $\leq$ Tot chol $<$ 6.19 2: Tot chol $\geq$ 6.19

Table 4.2: Distribution of numerical features (median and IQR) in the training set, grouped by non-diabetic (T2D=0) and future T2D patients (T2D=1).

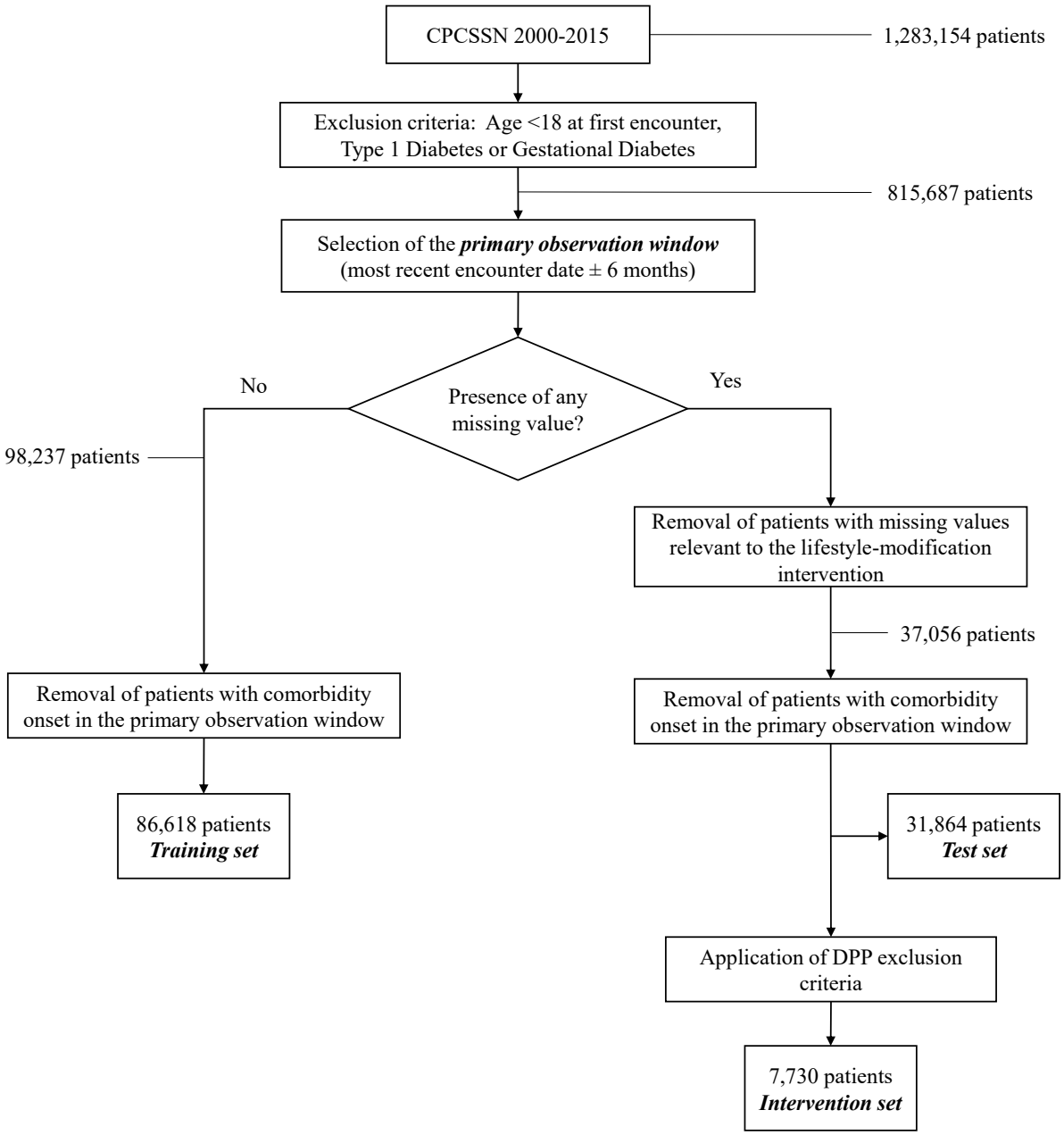


Figure 4.1: Pipeline of extraction of training, test, and intervention datasets.

Feature	**T2D=0 (N= 84,340)**	**T2D=1 (N= 2,278)**	**Levels**
Sex assigned at birth	female: 51,627 male: 32,713	female: 1,172 male: 1,106	0: female 1: male
Smoke	64% never 23% ex 13% current	53 % never 29 % ex 18 % current	0: never smoked 1: ex-smoker 2: current smoker
Antidepressants	no: 99% yes: 1%	no: 98% yes: 2%	0: no 1: yes
Corticosteroids	no: 98% yes: 2%	no: 97% yes: 3%	0: no 1: yes
Antihypertensives	no: 83% yes: 17%	no: 69% yes: 31% %	0: no 1: yes
Cholesterol lowering medications	no: 88% yes: 12%	no: 77% yes: 23%	0: no 1: yes
Quit-smoking medications	no: 99% yes: 1%	no: 98% yes: 2%	0: no 1: yes
Hypertension (HTN)	no: 77% yes: 23%	no: 69% yes: 31%	0: no 1: yes
Chronic Obstructive Pulmonary Disease (COPD)	no: 97% yes: 3%	no: 97% yes: 3%	0: no 1: yes
Depression	no: 88% yes: 12%	no: 87% yes: 13%	0: no 1: yes
Osteoarthritis (OA)	no: 88% yes: 12%	no: 89% yes: 11%	0: no 1: yes

Table 4.3: Distribution of categorical features in the training set, grouped by non-diabetic (T2D=0) and future T2D patients (T2D=1).

4.3 Methodology

The methodological workflow[2] followed in this Chapter consists of two main phases: causal discovery and counterfactual inference, discussed in the following sections. Causal discovery and counterfactual inference were performed using the R package `bnlearn` [146] (R version 4.3.1) and the Java library `Credici` [147], respectively.

4.3.1 Causal discovery

In line with most clinical case studies, the causal structure was not known a priori. Consequently, causal discovery techniques (see Section 2.2.2) were applied to learn the underlying DAG from data. Two complementary structures were retrieved during the causal discovery phase: (i) one involving all observed (endogenous) variables included in the training dataset and listed in Tables 4.2-4.3 (DAG_{all}), and (ii) another over a subset of features, incorporating latent (exogenous) variables representing unobserved factors that may confound the relationships among observed variables (DAG_{sub}).

Causal discovery of DAG_{all}

The causal discovery process was conducted iteratively by alternating between a structure learning step and an expert review step, until the resulting DAG was both data-driven and clinically consistent. The structure learning step of DAG_{all} was carried out using BIC [148] maximization through Hill climbing search [84] on the training data (see Section 2.2.2), employing bootstrapping to assess the variability of the identified relationships as the input data varied. The bootstrap procedure consisted of 100 iterations, each using 90% of the training set and preserving the original proportion of T2D patients. After each learning step, the estimated DAG was reviewed by a physician with expertise on T2D prevention to distinguish, whenever possible, statistical dependencies from clinically plausible cause-and-effect mechanisms. The resulting domain knowledge was then encoded as blacklists (i.e., sets of forbidden edges in the graph) to constrain the next structure learning step [78]. Blacklists were additionally used to preserve temporal ordering according to the data extraction procedure, leveraging the longitudinal nature of EHR data to support the identification of edge orientation, as suggested by [37].

Specifically, blacklists were formulated, together with clinicians, to:

- ensure that Age and Sex assigned at birth were treated as root nodes, i.e., non-modifiable factors, by preventing edges from entering these nodes;

[2]Code available at: https://gitlab-core.supsi.ch/dti-idsia/causal_T2D

- ensure temporal ordering by preventing the T2D node from being the parent of other variables, as the information about T2D was the last available in time;
- prevent medication–medication edges to reduce the risk of misinterpreting associations driven by common causes (e.g., shared clinical characteristics, underlying medical conditions or treatment indications) as direct causal relationships [149];
- ensure that the prescription of a particular group of medications did not lead to the diagnosis of a certain comorbidity or changes in smoking habits, since the presence or absence of comorbidities was verified prior to the medication extraction window;
- ensure that the diagnosis of a comorbidity did not directly influence the diagnosis of another comorbidity;
- ensure that blood biomarkers and pressure values did not influence smoking habits and that they did not influence BMI, assuming that if a causal relationship exists it should be in the opposite direction, as found for example in [37].

The full set of blacklisted relationships is reported in Table 4.4.

As a validation of DAG_{all} in absence of the ground truth structure, its ability to predict T2D onset on the test set was assessed [32]. The estimated DAG was complemented with a Bayesian network parameter estimation, performed by means of a Bayesian posterior distribution with uniform prior. The retrieved classification performance was compared with that of a Tree-Augmented naive Bayes (TAN) [150] classifier, which is another graphical model specifically designed for classification. The TAN model was trained on the training set by selecting T2D onset as output and by estimating the parameters with the same method used to learn parameters of DAG_{all}. To account for class imbalance, the classification threshold was lowered to match the proportion of patients with a future T2D diagnosis in the training set.

Causal discovery of DAG_{sub}

Exogenous latent variables representing information not directly available in the dataset were included in the causal discovery process of DAG_{sub}.

DAG_{sub} was defined on a subset of the available variables. The rationale was to include all the variables that significantly contribute to the occurrence of T2D according to the previously retrieved DAG_{all} and expert opinion, while keeping a reduced computational complexity of the counterfactual inference task. The selected subset of variables comprised exclusively Age, Sex assigned at birth, BMI and laboratory blood sample values. Total

Type	From	To
BL	Sex at birth,Pressure, BMI, LDL, HDL, TG, FPG, Total Cholesterol, Antidepressants, Cholesterol lowering meds, Antihypertensive meds, Corticosteroids, Smoking meds, Smoking, Depression, HTN, OA, COPD, T2D	Age
BL	Age, Pressure, BMI, LDL, HDL, TG, FPG, Total Cholesterol, Antidepressants, Cholesterol lowering meds, Antihypertensive meds, Corticosteroids, Smoking meds, Smoking, Depression, HTN, OA, COPD, T2D	Sex at birth
BL	T2D	Age, Sex at birth, Pressure, BMI, LDL, HDL, TG, FPG, Total Cholesterol, Antidepressants, Cholesterol lowering meds, Antihypertensive meds, Corticosteroids, Smoking meds, Smoking, Depression, HTN, OA, COPD
BL	Antidepressants	Cholesterol lowering meds, Antihypertensive meds, Corticosteroids, Smoking meds, Smoking, Depression, HTN, OA, COPD, T2D
BL	Cholesterol lowering meds	Antidepressants, Antihypertensive meds, Corticosteroids, Smoking meds, Smoking, Depression, HTN, OA, COPD, T2D
BL	Antihypertensive meds	Antidepressants, Cholesterol lowering meds, Corticosteroids, Smoking meds, Smoking, Depression, HTN, OA, COPD, T2D
BL	Corticosteroids	Antidepressants, Cholesterol lowering meds, Antihypertensive meds, Smoking meds, Smoking, Depression, HTN, OA, COPD, T2D
BL	Smoking meds	Antidepressants, Cholesterol lowering meds, Antihypertensive meds, Corticosteroids, Smoking, Depression, HTN, OA, COPD, T2D
BL	LDL, HDL, TG, Total_Cholesterol, FPG, BMI, Pressure	Smoking
BL	Depression	HTN, OA, COPD
BL	HTN	Depression, OA, COPD
BL	OA	Depression, HTN, COPD
BL	COPD	Depression, HTN, OA, COPD
BL	BMI	LDL, HDL
BL	HDL	Antihypertensive meds
BL	HTN	Cholesterol lowering meds
BL	Pressure, LDL, HDL, TG, FBS	BMI
BL	Age, Sex at birth, LDL, HDL, TG, FPG, BMI, Pressure, T2D	Z
BL	Z	Age, Sex at birth
WL	Z	BMI, Pressure, FPG, TG, LDL, HDL, Diabetes

Table 4.4: Full list of blacklisted (BL) and whitelisted (WL) relationships. The last three rows refer to specific constraints for the inclusion of Z as latent variable in DAG_{sub}.

cholesterol was not included in this reduced subset of features because it can be derived from the other lipoproteins in the features list, while comorbidities were excluded due to their chronic nature and medications and smoking were excluded because of their negligible effect on the output of interest according to the estimated DAG_{all}. Additionally, a latent variable Z was included in DAG_{sub} to account for factors not directly observed in the dataset, but inferable from observed variables, that may influence causal relationships between the reduced set of features and T2D. The latent variable modelled as a 4 states discrete variable, represents a broader construct encompassing multiple unmeasured confounders including, but not limited to, lifestyle-related behaviors (e.g., diet, physical activity, and other health behaviors) that may affect the relationships between endogenous nodes.

The structure learning step of DAG_{sub} was carried out using the Structural EM [86] algorithm (see Section 2.2.2), which allowed us to learn a DAG from data in the presence of latent variables. Specifically, the BIC score was maximized, by means of hill-climbing search, and the previously learned DAG_{all} (without latent variable) was used as the starting graph $\mathcal{G}_0$. Structure learning was alternated with expert review, following the same procedure used for DAG_{all}. In addition to the previously defined blacklists, further constraints were added to handle the presence of the latent variable (Table 4.4). Specifically, all edges entering the latent variable were blacklisted to ensure that the latent variable was treated as a root node (i.e., an exogenous node). Moreover, to reflect the theoretical assumption that the latent construct causally influences the observed indicators, all edges connecting the latent variable to endogenous factors (i.e. all nodes of the reduced set except Age and Sex assigned at birth) were included in a whitelist (i.e., set of required relationships). The final structure was then defined as the averaged network over 100 bootstrap runs of Structural EM, by selecting only edges appearing with a frequency higher than 90% to reject potential spurious, unstable, edges.

4.3.2 Counterfactual inference

To illustrate the advantages of employing counterfactual inference in chronic disease prevention, such methodology was used to estimate the impact of a lifestyle modification intervention on T2D onset, using observational data (i.e., the intervention set, shown in Table 4.5), in a setting where lifestyle itself was not directly observed. To model the unknown lifestyle modification, we relied on the well-established impact of lifestyle on BMI and FPG, which were instead observable in the data. Specifically, we first inferred the latent variable Z associated with the observed BMI and FPG variables (factual scenario). Although the latent variable does not explicitly represent lifestyle, it captures underlying unobserved changes

reflected in BMI and FPG. Then, we used this information to refine the probability of T2D onset under a hypothetical lifestyle intervention, modeled as a reduction in the observed BMI and FPG values (counterfactual scenario).

Counterfactual inference using EMCC. The EMCC algorithm ([99], see Section 2.2.3 for a brief description) was applied in its relaxed form, where the unknown structural equations were approximated by CPTs estimated via a frequentist Bayesian-network parameter estimator. In order to reduce complexity of the estimation procedure and ensure computational feasibility, counterfactual inference was conducted on DAG_{sub}, separately for different patient subgroups, based on the values of the nodes Age (under or over 65 years) and Sex assigned at birth. Moreover, all the variables except FPG and BMI were binarized, as shown in Table 4.5. For each patient i, the procedure involved three computational steps:

i) the patient-specific posterior probability of T2D onset in the factual world was computed by marginalizing over the prior distribution on the latent variable, hence:

$$P^i_{\text{T2D}} = P(\text{T2D} = 1|\mathbf{X}_i) = \int P(\mathbf{X}_i, Z)P(Z)dZ,$$

where Z represents the latent variable. This quantity represents the patient-specific probability of T2D onset given the observed features (X_i), averaged over all plausible latent configurations implied by the prior.

ii) for the same patient, the posterior probability distribution of the latent variable, denoted as $P(\text{Z}|\mathbf{X}_i)$, was inferred to capture the latent variable configuration most consistent with the observed factual data.

iii) given the observed features in the factual world ($\mathbf{X}_i$) and the observed reduction in BMI and FPG in the counterfactual world (BMI^*_i and FPG^*_i), the posterior probability of T2D onset in the counterfactual world was computed as

$$P^i_{\text{T2D}^*} = P(\text{T2D}^* = 1|\mathbf{X}_i, \text{BMI}^*_i, \text{FPG}^*_i) = \int P(\text{T2D}^* = 1|\text{BMI}^*_i, \text{FPG}^*_i, Z)P(Z|\mathbf{X}_i)dZ$$

This marginalization combines the patient-specific latent posterior inferred in the factual world with the hypothetical counterfactual conditions.

The relaxed EMCC algorithm was repeated for 200 runs, each initialized with a Dirichlet prior distribution over the latent variable and randomly parameterized CPTs approximating the structural equations. Runs were retained valid only if the prior and CPTs were compatible with the data, i.e., if the marginalized likelihood of the observable data reached its

maximum, as described in [100]. Finally, to capture uncertainty arising from alternative plausible parameterizations of the causal model, the patient-specific lower and upper bounds for the counterfactual probability were estimated as the minimum and maximum counterfactual probabilities obtained across all valid EMCC runs. Hence, the patient-specific counterfactual probability was then expressed as an interval, i.e., $[\underline{P^i_{\text{T2D}^*}}, \overline{P^i_{\text{T2D}^*}}]$.

Feature	**T2D=0 (N= 7177)**	**T2D=1 (N= 553)**	**Updated Discretization levels**
Sex assigned at birth	female: 3654 male: 3523	female: 279 male: 274	0: female 1: male
Age	60 (52-69)	59 (51-67)	0: Age < 65 1: Age ≥ 65
BMI [kg/m^2]	30.0 (27.2-33.9)	32.4 (29.2-36.6)	0 : $BMI < 18.5$ 1: $18.5 \leq$ BMI < 25 2: $25.0 \leq$ BMI < 30 3: $30.0 \leq$ BMI < 35 4: $35.0 \leq$ BMI < 40 5: BMI ≥ 40
Pressure [mmHg]	SBP: 130.0(120.0- 138.0) DBP: 78.0 (72.0-83.3)	SBP: 130 (122-138) DBP: 79.6 (73.0-84.0)	0: Pressure $= \{0, 1\}$ 1: Pressure $= \{2, 3\}$
FPG [mmol/L]	5.6 (5.4-5.9)	6.1 (5.7-6.5)	0: FPG ≤ 5.5 1: $5.6 \leq$ FPG < 7.0 2: FPG ≥ 7.0
HDL [mmol/L]	1.29 (1.08-1.53)	1.16 (0.99-1.40)	0: HDL < 1 (men) or HDL < 1.3 (women) 1: HDL ≥ 1 (men) or HDL ≥ 1.3 (women)
LDL [mmol/L]	2.98 (2.31-3.63)	2.78 (2.10-3.40)	0: LDL < 4.2 1: LDL ≥ 4.2
TG [mmol/L]	1.37 (1.00-1.90)	1.60 (1.21-2.20)	0: TG < 2.3 2: TG ≥ 2.3

Table 4.5: Distribution of features in the intervention set, for non-diabetic (T2D=0) and future T2D patients (T2D=1). Numerical features: median(IQR).

T2D risk labeling. The risk of T2D onset for a certain patient was evaluated by comparing their probabilities of T2D onset in the factual and counterfactual worlds against a discrimination threshold (*thr*). Patients with onset probabilities below the threshold were considered at *low risk* of developing T2D, whereas those with onset probabilities above the threshold were defined as at *high risk*. If the estimated counterfactual probability interval overlapped with the threshold, no risk label could be assigned to the patient. The threshold was pre-established at 7% to align with the proportion of patients with a future T2D diagnosis in the intervention set. This threshold prioritizes sensitivity, potentially leading to more false positives. This choice is justified because the practical cost of labeling errors

is not symmetric, as failing to intervene in case of high T2D risk can have serious consequences, whereas the downsides of recommending lifestyle interventions when unnecessary are relatively minor.

Estimated intervention outcomes. The intervention was deemed effective for a specific patient in case of high T2D risk in the factual world and low T2D risk, even in the worst case scenario, in the counterfactual world, i.e., when $P^i_{\text{T2D}} \geq thr$ and $\overline{P^i_{\text{T2D}^*}} < thr$. The predicted reduction in T2D incidence due to the intervention was then computed as the percentage reduction in the number of subjects at high risk of T2D when applying the hypothetical intervention (counterfactual world) compared to the control case (factual world). Due to non-identifiability, this value belongs to an interval, i.e., $RED\% \in [\text{RED}\%_{low}, \text{RED}\%_{high}]$. The lower bound was calculated as

$$\text{RED}\%_{low} = \frac{n_{\text{I}}}{n_{\text{HR}}},$$

representing the ratio between the number of subjects for whom the intervention was deemed effective (n_{I}) and the initial number of subjects at risk of T2D in the factual world (n_{HR}, i.e., the total number of subjects with $P^i_{\text{T2D}} \geq thr$). The upper bound was obtained by including in the numerator those individuals whose T2D risk in the counterfactual world could not be determined because their outcome probability interval overlapped with thr:

$$\text{RED}\%_{high} = \frac{n_{\text{I}} + n_{\text{overlap}}}{n_{\text{HR}}}$$

Herein, I will report only the lower bound as a conservative estimate of the reduction in T2D incidence.

4.4 Results

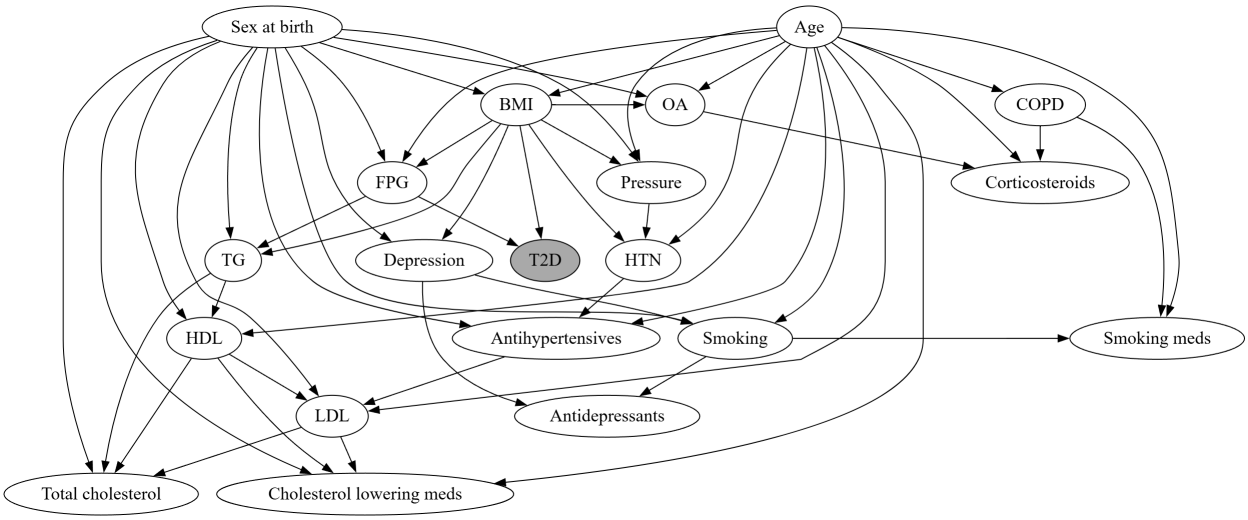
4.4.1 Causal discovery

DAG_{all}, as obtained following the method described in Section 4.3.1, is illustrated in Figure 4.2a. The retrieved causal structure shows many interconnected pathways between nodes, such as the well established positive association between age, smoking and COPD, associations between cholesterol levels (TG, HDL, and LDL) and associations between a diagnosis of a given condition and prescription of relatable medications (e.g., from depression to antidepressants, from HTN to antihypertensives). The reason for the presence of the non-intuitive edge "Smoking→antidepressants" may be due to the prescription of certain medications with

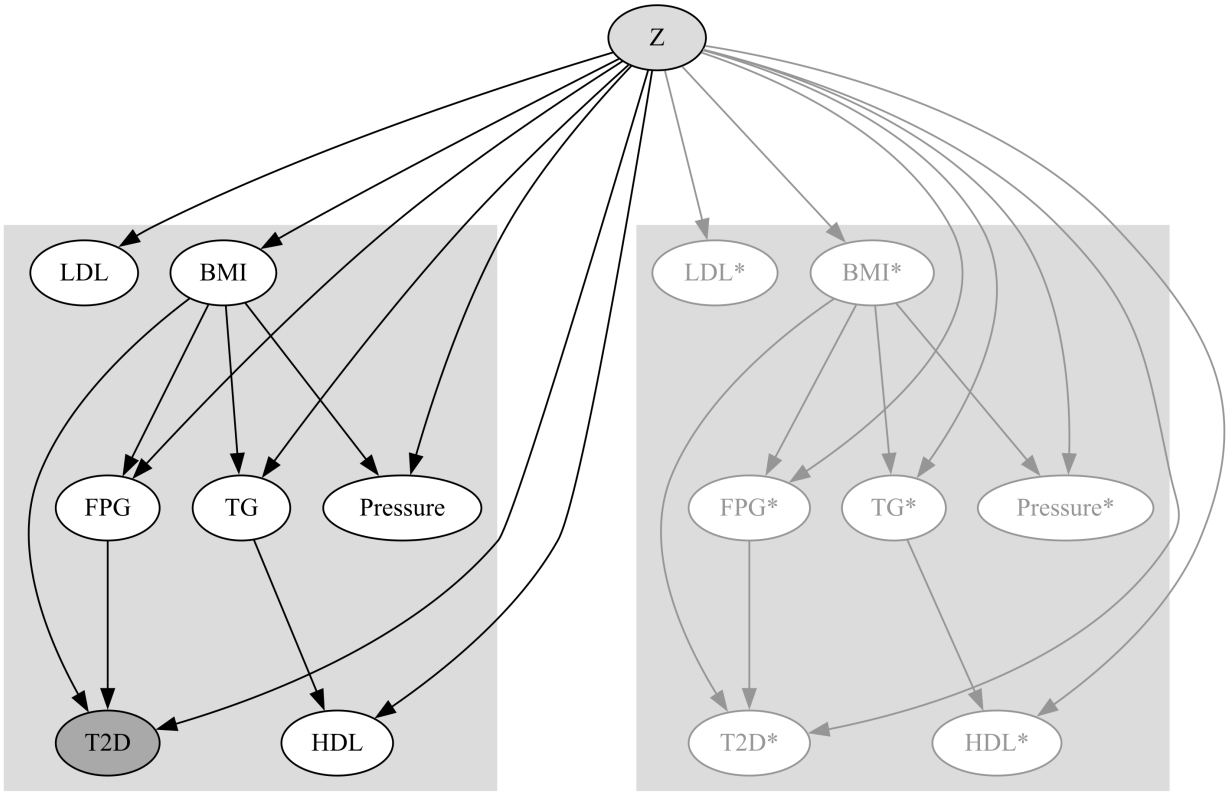
antidepressant properties like bupropion, often prescribed to aid smoking cessation [151]. As a result of the bootstrap procedure, 47 out of 51 edges appeared in more than 90% of the iterations, demonstrating high stability of structure learning. Notably, the edge connecting BMI and OA was the only one showing an ambiguous orientation after structure learning. The latter was determined a posteriori as "BMI → OA" due to the well-established association between high BMI categories and the incidence of osteoarthritis, particularly in weight-bearing joints like the knee [152]. DAG_{all} achieved a test classification accuracy of 75.1%, sensitivity of 76.8%, specificity of 76%, and AUC of 0.81 when predicting future T2D onset. Performance is comparable to that of the TAN classifier trained on the same data, yielding accuracy = 74.7%, sensitivity = 74.5%, specificity 74.8% and AUC = 0.83. Sensitivity and specificity values match those of 12 eligible models highlighted in a recent review on machine learning models for T2D prediction [141] (pooled sensitivity = 0.81, 95% confidence interval = 0.67-0.90; pooled specificity = 0.82, 95% confidence interval= 0.74–0.88), whereas AUC resulted in slightly lower performance with respect to the reported studies (AUC = 0.88, 95% confidence interval= 0.85-0.91). By focusing solely on Bayesian network models, our model performed better than the highest performing model obtained by [143] on a largely overlapping set of features in terms of AUC (0.76) and sensitivity (62.8%), while maintaining comparable accuracy (76.2%) and specificity (77.3%). Given the satisfactory classification performance in predicting T2D, we can conclude that DAG_{all} is properly representing the main relationships underlying T2D onset in the set of routinely collected primary care data here used, offering a plausible and interpretable causal representation in the absence of a known ground truth structure.

As anticipated, incorporating the latent variable into the causal discovery process (DAG_{sub}, shown in the left-hand side of Figure 4.2b) slightly altered the causal relationships among endogenous variables retrieved in DAG_{all} (Figure 4.2a). In particular, some edges originally observed in DAG_{all} are absent in DAG_{sub} (i.e., edges between FPG and TG and between HDL and LDL). These edges likely reflect known relationships such as the positive association between FPG and TG and the negative association between HDL and LDL [153], rather than direct causal effects. In fact, these relationships could be confounded by the presence of unobserved common causes like diet, exercise, metabolic syndrome [154], or genetic factors [155]. The structure of DAG_{sub} demonstrates how the effects of including latent variables can partially compensate for these unobserved confounding factors, reducing the risk of introducing spurious relationships.

In DAG_{sub}, the onset of T2D depends directly on BMI, FPG, and Z, with an indirect contribution from Sex assigned at birth (mediated by BMI) and from Age (mediated by BMI and FPG). BMI can influence the probability of T2D onset in multiple ways, reflecting the



(a) DAG_{all} learned from the complete set of variables listed in Tables 4.2-4.3.



(b) Twin network associated to DAG_{sub}, describing the relationships between observed biomarkers and T2D in the factual world (left-hand side) and counterfactual world (right-hand side, nodes marked with *). Sex at birth and Age nodes, together with the corresponding edges, are not shown to improve the readability of the graph. Sex at birth is a parent for all endogenous nodes except for Age, FPG, and T2D, whereas Age is a parent for all endogenous nodes except for Sex at birth, TG, HDL, and T2D.

Figure 4.2: DAG_{all} (Panel 4.2a) and twin network associated to DAG_{sub} (Panel 4.2b). DAG_{sub} is represented on the left-hand side of Panel 4.2b.

complex interplay between BMI and T2D pathogenesis [156]. More specifically, elevated BMI may act as a driver for adipose tissue inflammation, β-cells disfunction, and insulin resistance leading to T2D development [156]. Accordingly, DAG_{sub} captures two distinct causal paths between BMI and T2D: on the one hand, an increase in BMI can directly cause an increased risk of T2D, and on the other hand, an increase in BMI can lead to elevated baseline FPG, which in turn increases the likelihood of T2D onset. Such pathways are consistent with those identified in a DAG derived using a retrospective EHR dataset of over 57,000 patients from Mayo Clinic [37].

4.4.2 Counterfactual inference

Lifestyle intervention design. As a first approximation, we simulated a lifestyle modification intervention by assuming that each patient in an intervention set (subset of the test set) experienced reductions in BMI and FPG, and assessed the resulting probability of T2D onset. The design of the Diabetes Prevention Program (DPP) trial (1996–2002) [138], which investigated whether a lifestyle intervention could prevent T2D in a population of 3,234 adults from the United States of America, was emulated to showcase the potential of the proposed counterfactual inference approach using realistic and feasible intervention values. The same eligibility criteria used in the DPP trial were applied to the test set (age $\geq$ 25, BMI$\geq$ 24 kg/m^2, and 5.3 $\leq$ FPG $\leq$ 6.9 mmol/L), obtaining an intervention dataset of 7,730 records, distributed as in Table 4.5. Then, reductions in weight and FPG matching those observed, on average, in the lifestyle intervention group of the DPP trial after one and a half years, i.e., a weight loss of 6 kg and a FPG reduction of 0.17 mmol/L [138], were applied to each record of the intervention set. Such values were compatible with a lifestyle modification intervention characterized by healthy low-calorie, low-fat diet coupled with at least 150 minutes per week of moderate intensity physical activity [138].

Given the discretized nature of the intervention dataset, such changes in weight and FPG did not lead to changes in the categorical BMI and FPG values for all patients. Specifically, only 3,217 out of 7,730 patients changed their BMI and/or FPG category after the hypothetical intervention (BMI category only: 2,382 patients; FPG category only: 548 patients; BMI and FPG category: 287 patients), whereas 4,513 patients remained in the same FPG and BMI categories.

T2D risk labeling. SCM theory allows us to compute both factual and counterfactual probabilities of T2D onset and to compare them at the patient-specific level. Figure 4.2b shows the twin network of DAG_{sub}, used for the estimation of both factual and counterfactual probabilities of T2D onset. In the factual world (left-hand side), all features except T2D

were observed. In the counterfactual world (right-hand side), only post-intervention values of BMI and FPG (i.e., BMI_i^* and FPG_i^*, for each patient i in the intervention set) were observed. In the factual world, 4,489 patients were labeled as having a low risk of developing T2D ($P^i_{\text{T2D}} < thr$), whereas 3,241 patients were labeled as at high risk of developing T2D ($P^i_{\text{T2D}} \geq thr$). In the counterfactual world, 3,907 patients were considered at low risk of T2D onset ($\overline{P^i_{\text{T2D}^*}} < thr$) after the intervention, whereas 595 patients were considered at high risk ($\underline{P^i_{\text{T2D}^*}} \geq thr$). The remaining 3,228 patients had counterfactual probability intervals that overlapped with *thr* and therefore they could not be assigned to a definite risk level.

Estimated intervention outcomes. Of the 3,241 subjects initially labeled as at high risk of T2D in the factual world (n_{HR}), only 591 remained at high risk of T2D in the counterfactual world, with their counterfactual probability intervals entirely above the threshold, indicating no improvement attributable to the intervention. The remaining 2,650 patients showed a reduction in the estimated probabilities due to the intervention (with at least the lower bound of their counterfactual probability interval falling below *thr*). Among these, 491 (n_{I}) had probability intervals entirely below the threshold in the counterfactual world, reflecting a clear and measurable reduction in estimated T2D risk attributable to the intervention, while 2,159 patients (n_{overlap}) had intervals overlapping the threshold, indicating uncertain improvement in risk. Consequently, the proposed lifestyle modification intervention was estimated to yield at least to a 15.1% reduction in T2D incidence ($\text{RED}\%_{low}$). As shown in Table 4.6, the intervention appeared to be more effective in men than women, in subjects with BMI lower than 30 kg/m^2 than in those with higher BMI, and in subjects with FPG $\leq$ 5.5mmol/L compared to those with FPG in the prediabetic range. $\text{RED}\%_{low}$ values for different age groups were similar. These trends are consistent with the results reported in [138] (shown as $\text{RED}\%_{\text{DPP}}$ in Table 4.6). The values here estimated are less pronounced compared to the reference values observed in the DPP trial, maintaining nevertheless a certain proportionality. Notably, these values cannot be directly compared due to several factors, which will be discussed further in the Discussion section.

Heterogeneity of intervention efficacy across patients subgroups. Table 4.7 illustrates the variability in predicted intervention efficacy across different patients subgroups, thereby facilitating the identification of patients who are expected to benefit most from the lifestyle intervention. Patients were stratified based on the three features that most strongly influenced the counterfactual probability of T2D onset as determined by XGBoost regression. These features were age, TG, and LDL in female subjects; age, TG, and BMI in male subjects. In female subjects, higher intervention efficacy was expected in those older than 65

years of age compared to younger subjects with the same TG and LDL levels. Conversely, among male subjects, higher intervention efficacy was expected in subjects under 65 years compared to older subjects with the same TG and BMI levels. Specifically, the average RED$\%_{low}$ was 10.8% for females under 65 years, 19.9% for females over 65 years, 19.4% for males under 65 years and 11.3% for males over 65 years.

In female subjects with TG values within the normal range (i.e., TG$<$ 2.3 mmol/L) higher intervention efficacy was observed when LDL levels were outside the normal range (LDL$\geq$ 4.2 mmol/L), independently on age. A reliable value of RED$\%_{low}$ could not be computed for female subgroups with both TG and LDL levels outside the normal range, due to limited sample size (i.e., $n_{\text{HR}} \leq 10$). The highest predicted efficacy for females was observed in subjects over 65 years of age with normal TG levels and LDL outside normal ranges, whereas the subgroup with lower predicted efficacy included subjects with both TG and LDL within normal ranges. In male subjects, higher intervention efficacy was observed when BMI is lower than 30 kg/m^2 compared to higher BMI values. This behavior was observed in all subgroups, independently on age and TG levels. Moreover,

$N_{\textbf{intervention}}$**= 7730**	**Control**	**Intervention**	**Reduction in T2D incidence**	
	n_{HR}	n_{I}	RED$\%_{low}$	RED$\%_{\text{DPP}}$
Sex assigned at birth				
Female	1,513	200	13.2	(40, 64)
Male	1,728	291	16.8	(49, 76)
Age (DPP inclusion: Age $\geq$ 25**)**				
Age $<$ 65*	2,285	348	15.2	(27, 63) (44, 70)
Age $\geq$ 65*	956	143	15.0	(51, 83)
BMI [kg/m^2] (DPP inclusion: BMI $\geq$ 24**)**				
BMI $<$ 30	1,092	266	24.4	(46, 77)
30.0 $\leq$ BMI $<$ 35	1,286	148	11.5	(40, 75)
BMI $\geq$ 35	863	77	8.9	(34, 63)
FPG [mmol/L](DPP inclusion: 5.3 $\leq$ **FPG** $\geq$ 6.9**)**				
FPG $<$ 5.6	229	44	19.2	(38, 68)
FPG $\geq$ 5.6	3,012	447	14.8	(51, 72)
Total	**3,241**	**491**	**15.1**	**(48, 66)**

Table 4.6: Minimum estimated reduction in T2D incidence (RED$\%_{low}$) when performing the hypothetical intervention (counterfactual world) compared to the control case (factual world): total reduction and univariate stratification by sex at birth, Age, BMI, and FPG. RED$\%_{\text{DPP}}$, representing the 95% confidence interval observed in the DPP trial [138], is reported for comparison. (*the following age categories are instead considered in the DPP trial: 25 $\leq$ Age $<$ 45, 45 $\leq$ Age $<$ 60 , and Age $\geq$ 60).

similar values of treatment efficacy were observed in male subgroups with BMI lower than 30 kg/m^2, independently of TG levels. The highest predicted efficacy for males was observed in subgroups of subjects below 65 years and with BMI lower than 30 kg/m^2, whereas the subgroup with lower predicted efficacy included subjects above 65 years, with TG within normal ranges and BMI above 30 kg/m^2.

Sex at birth	**Age**	**TG** [mmol/L]	**LDL** [mmol/L]	**RED%**
females	Age < 65	TG < 2.3	LDL < 4.2	6.8
			LDL ≥ 4.2	27.9
		TG ≥ 2.3	LDL < 4.2	16.2
			LDL ≥ 4.2	n.a.
	Age ≥ 65	TG < 2.3	LDL < 4.2	17.3
			LDL ≥ 4.2	43.5
		TG ≥ 2.3	LDL < 4.2	23.8
			LDL ≥ 4.2	n.a.
Sex at birth	**Age**	**TG** [mmol/L]	**BMI** [kg/m^2]	**RED%**
males	Age < 65	TG < 2.3	BMI < 30	34.0
			$30.0 \leq$ BMI < 35	11.8
			BMI ≥ 35	9.9
		TG ≥ 2.3	BMI < 30	33.7
			$30.0 \leq$ BMI < 35	9.7
			BMI ≥ 35	13.5
	Age ≥ 65	TG < 2.3	BMI < 30	15.8
			$30.0 \leq$ BMI < 35	7.9
			BMI ≥ 35	7.6
		TG ≥ 2.3	BMI < 30	15.8
			$30.0 \leq$ BMI < 35	13.3
			BMI ≥ 35	16.7

Table 4.7: Observed values of RED%$_{low}$ in clusters of subjects: multivariate stratification according to Age, TG, and LDL (female subjects) and Age, TG, and BMI (male subjects).

4.5 Discussion

This Chapter explores how causal learning methods can leverage observational data and medical domain knowledge to support chronic disease prevention, using T2D prevention from EHR data as a practical example. Causal discovery techniques were applied to identify interdependencies among a large set of commonly available EHR features, including blood biomarkers, pressure measurements, diagnoses of chronic diseases and medications, in a portion of the Canadian general population (almost 90,000 patients) captured by the nation-wide CPCSSN database. The causal structure was derived from the whole set of extracted features (DAG_{all}, shown in Figure 4.2a) and refined by expert review by means of constraints

(blacklists and whitelists) that were iteratively defined during the learning process. The retrieved causal structure aligned with established medical knowledge and achieved classification performance comparable to current literature on T2D prevention from observational data [141, 143]. Therefore, it can be used as the starting point to create tools to support medical decisions by summarizing and visualizing relationships among routinely available variables, providing interpretable and causally grounded guidance.

Causal models derived from observational datasets typically rely on the causal sufficiency assumption; however, this assumption is often not satisfied in real-world settings because measuring all factors that significantly influence the observed phenomenon is difficult in practice. For example, lifestyle-related information is often missing, poorly sampled, or poorly coded in EHR data. As a result, such variables are rarely included in existing predictive models of T2D [141], even if improving physical activity and diet is usually a central aspect of diabetes prevention programs and insufficient physical activity and unhealthy diet are well-known risk factors for T2D [138, 157, 158]. To partially compensate for the lack of causal sufficiency in the study dataset, a second causal model accounting for latent factors and represented by DAG_{sub} (Figure 4.2b, left-hand side), was estimated. Specifically, a new variable was included during structure learning to model the effects of unobserved confounders (e.g., lifestyle-related latent factors) on a subset of variables having a direct impact on T2D. This model was used to showcase the potential of applying counterfactual inference to estimate the impact of a hypothetical lifestyle modification intervention on the probability of T2D onset, on an individual basis.

As an illustrative example, a counterfactual scenario leading to a specific reduction in BMI and FPG, consistent with the outcomes reported in [138] (i.e., an average weight loss of 6 kg and a FPG reduction of 0.17 mmol/L), was designed and evaluated. The findings support the conclusion that lifestyle modification causes a reduction in the incidence of T2D compared to no intervention, similarly to the trends observed in the reference DPP trial (Table 4.6). Moreover, such methodology enabled the evaluation of the potential benefits in different subgroups of patients (Table 4.7), showing potential for supporting the identification of patients who are more likely to benefit from lifestyle intervention and emphasizing the heterogeneity in intervention outcomes. Notably, the introduction of an interval estimate for the probability of T2D onset in the counterfactual world allowed us to consider potential bias due to non-identifiability of treatment effects from data [31]. Being exploratory in nature, this applicative example included several design and methodological approximations, which precluded a direct quantitative comparison with the efficacy of lifestyle interventions observed in the DPP trial and other prevention programs (e.g., [159]). First, despite the application of identical eligibility criteria, the intervention dataset extracted from CPCSSN

(Table 4.5) and the original lifestyle intervention group in the DPP trial differed in their underlying data distributions. Specifically, the intervention dataset included subjects with a higher average age (60.1 ± 12.8 vs. 50.6 ± 11.3) and a greater proportion of male subjects (49.1% vs. 32%) compared to the lifestyle intervention group in the DPP trial. Second, the proposed counterfactual inference pipeline estimated intervention efficacy at the individual level, whereas the DPP trial relied on subject from two groups (intervention and placebo) to compute intervention efficacy. Third, as individual-level intervention data were not available from the DPP trial, a uniform reduction in BMI and FPG was applied to all subjects in the intervention set, using the average reductions reported in the DPP as a benchmark. Finally, the proposed methodological approach relied on some design choices aimed at reducing the computational cost of counterfactual inference and ensure tractability (e.g., features discretization and approximation of the unavailable structural equations), leading to approximated estimates.

Given the limitations that prevented direct quantitative comparison with the DPP trial [138], the results in Tables 4.6 and 4.7 should be viewed primarily as illustrative examples rather than definitive findings. Despite all the above factors, the observed trends support the potential of causal learning approaches to formulate hypotheses about the predicted benefits of specific lifestyle interventions based on observational primary care data. In particular, the proposed causal learning approach appears promising for estimating treatment effects from observational data even in the presence of non-identifiable queries—an issue that remains a major challenge in existing studies [31]—and for enabling risk-stratified analyses of treatment effect heterogeneity, thereby facilitating personalization, which is a key factor to improve the effectiveness of prevention programs [139].

Chapter 5

Causal Learning for T2D prevention based on dynamic simulations of disease progression

The previous Chapter illustrated the potential of causal learning for modeling lifestyle modification interventions (e.g., improvement in diet and physical activity, exemplified through a decrease in BMI and FPG) using observational EHR data in static settings, where temporal dynamics were not considered. In contrast, this Chapter presents a complementary analysis on the same case study (*Case Study 2*) that explicitly accounts for the time-dependency of causal relationships, coded through systems of ordinary differential equations (ODEs), enabling a more realistic and nuanced modeling of longitudinal interventions and their effects over time. This research, which focused on the analysis of physical activity personalization for T2D prevention, was conducted within the framework of the Horizon Europe PRAESIIDIUM project, during my three-month visiting period at the Dalle Molle Institute for Artificial Intelligence, DTI – Dipartimento Tecnologie Innovative of SUPSI – Scuola universitaria professionale della Svizzera italiana.

As anticipated in Chapter 4, promoting physical activity is a key strategy in T2D prevention programs [158, 138]. According to the World Health Organization (WHO) 2020 guidelines [160] all adults should engage in regular exercise consisting of at least 150 minutes of moderate-intensity aerobic physical activity, or at least 75 minutes of vigorous-intensity aerobic physical activity per week to slow down or even prevent progression to T2D. Although guidelines were proved to be effective in general, such recommendations may not be adequate for all subjects. Therefore, personalized physical activity plans are required, based on the subject's characteristics and needs [161, 162]. The method presented in this

Chapter and published in [163] offers a promising approach to tailor physical activity based on individual risk profiles from dynamic models of T2D progression.

5.1 Diabetes progression models

Diabetes progression models offer mathematical formulations of blood glucose regulation and its dynamics over time, capturing effects ranging from a minute scale [164, 165] to long-term dynamics spanning years [166, 167, 168, 52]. The minimal model by Bergman et al. [164, 169] forms laid the foundations for many short-term modeling approaches by introducing three state variables: glucose, insulin, and remote insulin. For example, Roy and Parker [165] extended the minimal model to account for the short-term effects of physical activity on glucose-insulin dynamics. Other research works have focused on the development of models that encapsulate metabolic and inflammatory mechanisms on various time scales [170, 171]. For example, Topp et al. [166] proposed a simple model that describes the dynamics of glucose and insulin concentrations, together with the dynamics of β-cells that account for long-term interactions. Subsequent extensions by De Gaetano et al. [172], Ha et al. [167] and De Gaetano & Hardy [173] introduced additional state variables to better capture diabetes progression over several years or even decades. However, these long-term progression models did not incorporate the effects of physical activity on glucose-insulin dynamics.

Recently, De Paola et al. al. [168, 52] filled this gap by introducing a model that captures the long-term effects of regular physical activity on blood glucose regulation. The model, consisting of 12 ODEs and over 50 parameters, can simulate an average adult subject with no assumptions about age, sex assigned at birth, and fitness level. The ODEs system is built on previous models [165, 174, 167] and consists of two different timescales: the *short-term equations* describe how the glucose-insulin regulation mechanism is influenced during a physical activity session on a minute scale, whereas the *long-term equations* describe how glucose regulation behaves over a time span of years, parametrized at a daily scale. However, due to the multi-scale nature of this model, optimization of physical activity parameters is computationally intensive (i.e., $\approx 1\, minute/simulation$ for 20-year simulations). Within PRAESIIDIUM, we derived an approximated version of the aforementioned model that retains the system's macroscopic dynamics while increasing computational efficiency, as reported in [175] and summarized in the following Section. This optimization allowed counterfactual inference to be performed with tractable computational complexity.

5.2 Homogenization of the De Paola model

Homogenization is a powerful approach to reduce the complexity of differential equation models which exhibit dynamics across multiple scales, including high-frequency oscillations over time and distinct spatial patterns [176, 177, 178, 179]. Although traditionally applied to partial differential equations, homogenization techniques can also be effectively applied to ODEs [180]. For the De Paola model [168, 52], homogenization preserves the long-term impact of physical activity without the need to solve the model on a minute scale, by replacing short-scale oscillations with a smoothed or averaged solution that produces a simplified, faster, model capturing macroscopic dynamics.

The De Paola model [168, 52] can be represented by the following system of ODEs:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{y}(t, PA) := \frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}\mathbf{y_1}(t, PA)\\ \mathbf{y_2}(t, \mathbf{y_1})\end{bmatrix} = \begin{bmatrix}\mathbf{f_1}(t, PA, \mathbf{y_1})\\ \mathbf{f_2}(t, \mathbf{y_1}, \mathbf{y_2})\end{bmatrix} =: \mathbf{f}(t, PA, \mathbf{y}) \tag{5.1}$$

for $0 < t \leq t_{\text{end}}$, together with the initial condition $\mathbf{y}(0) = \mathbf{y_0}$. The original units of measurements of the state variables in $\mathbf{y_1}$ and $\mathbf{y_2}$, along with their description, are provided in Table 5.1. A list of all the parameters of $\mathbf{y_1}$ and $\mathbf{y_2}$ is provided in Appendix B of [175].

The control variable is physical activity, $PA(t)$, represented as a periodic continuation of a Heaviside function, defined for $n \in \mathbb{N}$ periods as

$$PA(t) := \begin{cases}\xi & \text{for } k\nu \leq t \leq k\nu + \delta,\\ 0 & \text{for } k\nu + \delta < t < (k+1)\nu,\end{cases} \tag{5.2}$$

where $k = 0, \ldots, n-1$ is an index, ξ is the intensity of exercise above the basal level, defined from 0 to 92%, ν is the period, in days, between two consecutive sessions, and δ is the duration, in minutes, of each training session. $PA(t)$ can be represented compactly using the equivalent value $\overline{PA} = \xi\delta/\nu$ [181].

The vector $\mathbf{y_1}$ contains the *short-term equations* and comprises five state variables $\mathbf{y_1} := [VO_2, G_{\text{pr}}, G_{\text{up}}, I_{\text{e}}, IL6]^T$, that satisfy the following ODEs

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{y_1} := \frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}VO_2\\ G_{\text{pr}}\\ G_{\text{up}}\\ I_{\text{e}}\\ IL6\end{bmatrix} = \begin{bmatrix}\lambda_t\theta(\mathrm{PA(t)} - VO_2)\\ \lambda_t\alpha_2(VO_2 - G_{\text{pr}})\\ \lambda_t\alpha_4(VO_2 - G_{\text{up}})\\ \lambda_t\alpha_6(VO_2 - I_{\text{e}})\\ \lambda_t\kappa_{\text{IL6}}(VO_2 - IL6)\end{bmatrix} =: \mathbf{f_1}(t, PA, \mathbf{y_1}), \tag{5.3}$$

with initial conditions $\mathbf{y_1}(0) = [0,0,0,0,0]^T$ and parameters λ_t (scaling factor), θ, α_2, α_4, α_6, κ_{IL6} as provided in Appendix B of [175]. This set of equations describes how the glucose-insulin regulation mechanism is influenced during an exercise session due to the action of oxygen consumption VO_2, which in turn triggers the other variables. The state variable $IL6$ represents the release of Interleukin-6, a protein produced during exercise that can have anti-inflammatory effects.

The vector $\mathbf{y_2}$ summarizes the *long-term equations* and comprises 7 state variables, $\mathbf{y_2} := [VL, S_{\text{I}}, \Sigma, \Gamma, B, I, G]^T$, that satisfy the following ODEs

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{y_2} := \frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} VL \\ S_{\text{I}} \\ \Gamma \\ \Sigma \\ B \\ I \\ G \end{bmatrix} = \begin{bmatrix} h_{VL}(VL, IL6) \\ h_{S_{\text{I}}}(S_{\text{I}}, VL) \\ h_{\Gamma}(\Gamma, G) \\ h_{\Sigma}(\Sigma, \Gamma, G) \\ h_B(B, VL, \Gamma, \Sigma, G) \\ h_I(I, I_{\text{e}}, \Gamma, \Sigma, B, G) \\ h_G(G, G_{\text{up}}, G_{\text{pr}}, S_{\text{I}}, I) \end{bmatrix} =: \mathbf{f_2}(t, \mathbf{y_1}, \mathbf{y_2}), \tag{5.4}$$

together with the scaled initial conditions $\mathbf{y_2}(0) = [0, 1, \Gamma_{0\lambda}, \Sigma_{0\lambda}, 1, 1, 1]^T$. The right hand side functions are defined as follows:

$$\begin{aligned}
h_{VL}(VL, IL6) &:= \lambda_t \kappa_{\text{s}}(IL6 - VL), \\
h_{S_{\text{I}}}(S_{\text{I}}, VL) &:= d_\lambda(VL)\frac{\theta_{S_{\text{I}}} - \lambda_{S_{\text{I}}} S_{\text{I}}}{\tau_{S_{\text{I}}}}, \\
h_{\Gamma}(\Gamma, G) &:= \frac{g_\lambda(G) - \Gamma}{\tau_\Gamma}, \\
h_{\Sigma}(\Sigma, \Gamma, G) &:= \frac{s_\lambda(\Gamma, \Sigma, G) - \Sigma}{\tau_\Sigma}, \\
h_B(B, VL, \Gamma, \Sigma, G) &:= \frac{p_\lambda(VL, \Gamma, \Sigma, G) - a_\lambda(VL, G)}{\tau_B} B, \\
h_I(I, I_{\text{e}}, \Gamma, \Sigma, B, G) &:= r_\lambda(\Gamma, \Sigma, G)B - \kappa I - \lambda_{I_{\text{e}}I} I_{\text{e}}, \\
h_G(G, G_{\text{up}}, G_{\text{pr}}, S_{\text{I}}, I) &:= \rho_\lambda + \lambda_{tG}\Omega(\lambda_{G_{\text{pr}}} G_{\text{pr}} - \lambda_{G_{\text{up}}} G_{\text{up}}) - (\eta_0 + \lambda_{S_{\text{I}}I} S_{\text{I}} I)G.
\end{aligned}$$

These functions include six auxiliary functions d_λ, g_λ, s_λ, p_λ, a_λ, and r_λ that are described in detail in Appendix A.2 of [175]. The core of the set of long-term equations lies in the variables $[VL, B, I, G]^T$, modeling the glucose-insulin (G and I) negative feedback loop. This feedback mechanism involves the action of β cells (B), responsible for insulin release. The variable VL represents the integral effect of $IL6$ and bridges the two timescales in the model,

Table 5.1: Short description and units of measurement of the state variables in their unscaled form.

Variable	Description	Unit
VO_2	Oxygen consumption during exercise	given in %
G_{pr}	Incremental hepatic glucose production	mg/(kg min)
G_{up}	Increased glucose uptake by working tissues	mg/(kg min)
I_{e}	Incremental insulin removal	μU/ml
$IL6$	Concentration of IL-6 in the muscle	pg/ml
VL	Integral effect of IL-6 released during exercise	(pg/ml) min
S_{I}	Insulin sensitivity	ml/(μU day)
Γ	Shift of the glucose dependence	-
Σ	Insulin secretion capacity	μU/ (μg day)
B	β cells mass	mg
I	Serum insulin concentration	μU/ml
G	Plasma glucose concentration	mg/dl

accounting for the long-term effects of physical activity on β cells and insulin sensitivity (S_{I}) as described by De Paola et al. [168, 52]. Furthermore, the state variables Γ and Σ model mechanisms that link the effects of B on the negative feedback loop between G and I.

The coupling between the short- and long-term variables is one-way: the long-term dynamics depends on the short-term variables, but not vice versa. When numerically solving the system (Eq. 5.1), the short-term equations require the solver to take small time steps, in the order of minutes, to capture the periodic dynamics. Using homogenization these fast fluctuations are replaced with constant values that represent the average contribution of the short-term effects to $\mathbf{y_2}$. This corresponds to substituting the short-term state variables in the full model with constants. This reduction follows the idea of periodic averaging [178]. The averaging procedure can be carried out analytically due to the structure of the control function $PA(t)$ and the form of the short-term dynamics. More specifically, an approximation $\widehat{\mathbf{y}}$ to system (5.1) is introduced:

$$\frac{\mathrm{d}}{\mathrm{d}t}\widehat{\mathbf{y}}(t) = \frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} \widehat{\mathbf{y_1}}(t) \\ \widehat{\mathbf{y_2}}(t, \widehat{\mathbf{y_1}}) \end{bmatrix} = \begin{bmatrix} \widehat{\mathbf{f_1}}(t, \widehat{\mathbf{y_1}}) \\ \mathbf{f_2}(t, \widehat{\mathbf{y_1}}, \widehat{\mathbf{y_2}}) \end{bmatrix} = \widehat{\mathbf{f}}(t, \widehat{\mathbf{y}}), \tag{5.5}$$

for $0 < t \leq t_{\text{end}}$. For $\widehat{\mathbf{y_1}}$, we define a system

$$\frac{\mathrm{d}}{\mathrm{d}t}\widehat{\mathbf{y_1}}(t) = \frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}\widehat{VO_2}\\ \widehat{G_{\text{pr}}}\\ \widehat{G_{\text{up}}}\\ \widehat{I_{\text{e}}}\\ \widehat{IL6}\end{bmatrix} = \begin{bmatrix}0\\0\\0\\0\\0\end{bmatrix} = \widehat{\mathbf{f}}_{\mathbf{1}}(t, \widehat{\mathbf{y_1}}),$$

with the initial conditions $\widehat{\mathbf{y_1}}(0) := \mu$, where μ is used to approximate the oscillating behavior of the short-term effects described by $\mathbf{y_1}$. The right hand side for $\widehat{\mathbf{y_2}}$ remains as written out in System (5.4), but now gets the inputs from $\widehat{\mathbf{y_1}}$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\widehat{\mathbf{y_2}} = \frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}\widehat{VL}\\ \widehat{S}_{\text{I}}\\ \widehat{\Gamma}\\ \widehat{\Sigma}\\ \widehat{B}\\ \widehat{I}\\ \widehat{G}\end{bmatrix} = \begin{bmatrix}h_{VL}(\widehat{VL}, \widehat{IL6})\\ h_{S_{\text{I}}}(\widehat{S}_{\text{I}}, \widehat{VL})\\ h_{\Gamma}(\widehat{\Gamma}, \widehat{G})\\ h_{\Sigma}(\widehat{\Sigma}, \widehat{\Gamma}, \widehat{G})\\ h_{B}(\widehat{B}, \widehat{VL}, \widehat{\Gamma}, \widehat{\Sigma}, \widehat{G})\\ h_{I}(\widehat{I}, \widehat{I}_{\text{e}}, \widehat{\Gamma}, \widehat{\Sigma}, \widehat{B}, \widehat{G})\\ h_{G}(\widehat{G}, \widehat{G_{\text{up}}}, \widehat{G_{\text{pr}}}, \widehat{S}_{\text{I}}, \widehat{I})\end{bmatrix} = \mathbf{f_2}(t, \widehat{\mathbf{y_1}}, \widehat{\mathbf{y_2}}), \tag{5.6}$$

with the previously given initial conditions.

With respect to the original formulation [168, 52], the two timescales were aligned and the state variables were scaled to improve numerical stability [182]. The technical steps to transform the original system into the scaled version here described are provided in Appendix A of [175].

5.3 Methodology

This Section describes a counterfactual inference framework that exploited dynamic cause-and effect relationships derived from prior physics-informed knowledge, formalized through ODEs. A large dataset of simulations generated with the homogenized De Paola model was used to analyze the benefits of personalized physical activity plans on T2D risk over a 5-year time window. Particular attention was given to the glucose trajectories of individuals whose progression to T2D could have been prevented through tailored physical activity interventions and to evaluate variability in their response to exercise. Moreover, inter-individual variability in glucose responses to WHO-compliant physical activities (i.e., at least 75 min/week of vigorous-intensity aerobic physical activity) was evaluated in prediabetic individuals.

5.3.1 Structural Causal Model

The homogenized De Paola model in Eq. 5.5 provides deterministic relationships between variables and allows the derivation of a dynamic causal model M that satisfies the SCM theory, as demonstrated in [183].

In the examined case, M is a parametric SCM (see Section 2.2.3) represented by a triplet $\langle \mathbf{U}, \mathbf{V}, \mathcal{F} \rangle$ that consists of:

- $\mathbf{U} = \{G_{\text{obs}}, I_0, B_0, \Omega, \theta_{S_\text{I}}, \tau_{S_\text{I}}, v\}$ denoting patient-specific conditions as described in Table 5.2 or observations like $G_{\text{obs}} = [G_0, \ldots, G_{n_{\text{obs}}-1}]$, which comprises glucose measurements at instant $t_0, \ldots, t_{n_{\text{obs}}-1}$;
- $\mathbf{V} = \{\widehat{G}, \widehat{I}, \widehat{B}, \widehat{\Sigma}, \widehat{\Gamma}, \widehat{S}_\text{I}, \widehat{VL}, PA\}$ representing dynamic processes that are determined by other variables in $\mathbf{U} \cup \mathbf{V}$;
- $\mathcal{F} = \{f_1, ..., f_m\}$, functions derived from ODEs, mapping $\mathbf{U} \times (\mathbf{V} \backslash V_j)$ to a specific endogenous variable V_j, $j = 1, \ldots, m$.

The dynamic SCM can be visually represented with a graph $\mathcal{G}(M)$, as shown in Figure 5.1.

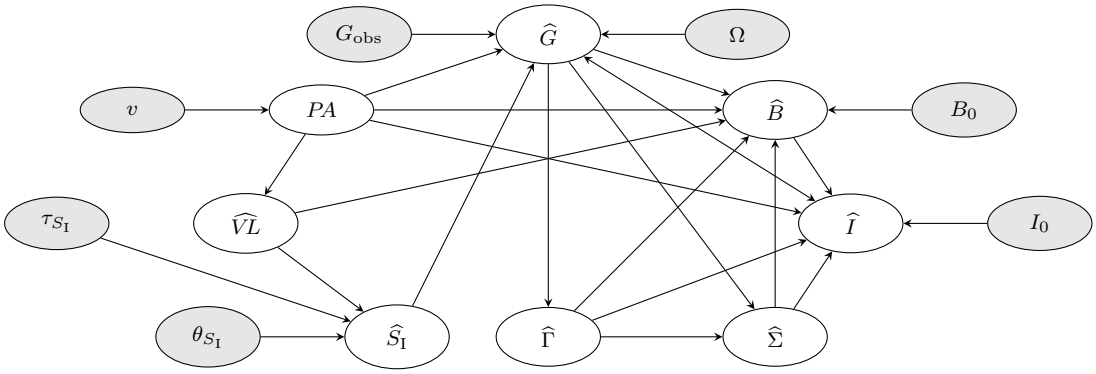


Figure 5.1: Causal graph $\mathcal{G}(M)$ of the homogenized De Paola model. Gray nodes: exogenous variables; white nodes: endogenous variables, indicating both the state variable and its first-order derivative.

The parametric nature of the model allows the representation of subjects with different propensities for T2D onset, by sampling variables in $\mathbf{U}$ within clinically acceptable ranges, using predefined step sizes (Table 5.2). In such a setup, if $\mathcal{U}$ denotes the (finite) set of possible values for $\mathbf{U}$, an instantiation $\mathbf{u} \in \mathcal{U}$ of the exogenous variables deterministically induces the values of all the endogenous variables and therefore represents specific individuals.

However, our knowledge of the system may be incomplete as some of the exogenous variables in $\mathbf{U}$ may not be determined in real-world settings, for instance because they

cannot be easily measured (e.g., B_0, θ_{S_I}, τ_{S_I}) or because they represent latent factors (e.g., v which indicates the willingness of a subject to follow the prescribed physical activity plan). As a result, we may only have partial knowledge of an individual. If $\mathbf{E}$ denotes the set of observable variables (i.e., $\mathbf{E} = \{G_{\mathrm{obs}}, \Omega, PA\}$) then, for each partial evidence $\mathbf{E} = \mathbf{e}$, $\mathcal{U}_e \subseteq \mathcal{U}$ represents the set of exogenous values compatible with $\mathbf{e}$, while $\mathbf{u_e}$ represents one of its elements [184]. In other words, the exact value $\mathbf{u}$ is unknown, but we can identify a set of compatible values $\mathcal{U}_e$, where each element $\mathbf{u_e}$ represents a possible configuration consistent with the observed evidence.

Variable	Unit	Description	Range	Step
G_0	mg/dl	Plasma glucose concentration at t=0	70-120	10
I_0	μU/ml	Serum insulin concentration at t=0	5-20	5
B_0	mg	β-cells mass at t=0	800-1300	100
Ω	kg	Weight	50-150	20
τ_{S_I}	days	Insulin sensitivity time constant	90-330	60
θ_{S_I}	ml/μU/days	Insulin sensitivity target value	0.18-0.68	0.10

Table 5.2: Description of variables in $\mathbf{U}$, including parameter ranges and step sizes used to define the parametric SCM.

5.3.2 Counterfactual Inference

As shown in Section 2.2.3, counterfactual inference examines what-if scenarios to understand the impact of hypothetical actions on an outcome of interest, e.g., the effect of a physical activity plan on blood glucose. When only partial evidence is available, as in the current analysis, the counterfactual query $\mathbf{Y_x}(\mathbf{u})$ is non-identifiable, and must be evaluated for each $\mathbf{u_e} \in \mathcal{U}_e$, thus yielding a set of compatible solutions [99].

Counterfactual query definition. Let $X = PA$, $Y = \widehat{G}$, and $\mathcal{PA}$ be the set of possible physical activity plans, parametrized as shown in Table 5.3 (minimum: 1 times/week, 30 minutes/session, 10% intensity; maximum: 3 times/week, 90 minutes/session, 70% intensity). Assume that, for a given individual, the factual physical activity $pa \in \mathcal{PA}$ results in glucose levels at $t_{\mathrm{end}} = 5$ years exceeding the diabetic threshold, representing the person's actual progression to diabetes, i.e., $\widehat{G}|_{t=t_{\mathrm{end}}} \geq thr_{\mathrm{T2D}}$, with $thr_{\mathrm{T2D}} = 125$ mg/dl.

Our objective is to evaluate the counterfactual query $\widehat{G}_{pa^*}(\mathbf{u_e})\,|_{t=t_{\mathrm{end}}}$, that is, *the value that* $\widehat{G}$ *would have obtained at* $t = t_{\mathrm{end}}$ *had pa been different, namely* $pa^* \in \mathcal{PA}\backslash\{pa\}$, and to identify all physical activity plans that could have prevented the subject from progressing from normoglycemia or prediabetes to T2D at $t = t_{\mathrm{end}}$. To compute the counterfactual query,

Variable	Unit	Description	Range	Step
ξ	%	Intensity of physical activity	10-70	10
$1/\nu$	times/week	Weekly frequency of physical activity	1-3	1
δ	minutes/session	Duration of a physical activity session	30-90	30

Table 5.3: Intensity, weekly frequency and duration of physical activity plans in $\mathcal{PA}$.

we suppose to observe $\mathbf{e} = \{\boldsymbol{g}_{\mathbf{obs}}, \omega, pa\}$, where $\boldsymbol{g}_{\mathbf{obs}} = [g_0, \ldots, g_{n_{\text{obs}}-1}]$ comprises one or more glucose measurements at instant $t_0, \ldots, t_{n_{\text{obs}}-1}$, such that g_0 is the glucose value at $t = 0$ and $t_{n_{\text{obs}}-1} < t_{\text{end}}$ with $t_{\text{end}} = 5$ years, whereas ω and pa are fixed during the observation period. Then, we follow these steps:

1. *Abduction.* Determine compatible sets of values for the exogenous nodes, given the evidence. Observing $\mathbf{e}$ and $\widehat{G}|_{t=t_{\text{end}}} \geq thr_{\text{T2D}}$ reduces the set of possible values of $\mathbf{U}$ from $\mathcal{U}$ to $\mathcal{U}_e \subseteq \mathcal{U}$, with $\mathcal{U}_e = \{u \in \mathcal{U} \mid \mathbf{E} = \mathbf{e}, \widehat{G}|_{t=t_{\text{end}}} \geq thr_{\text{T2D}}\}$.

2. For each compatible configuration of the exogenous nodes $\mathbf{u_e} \in \mathcal{U}_e$ and for each hypothetical physical activity plan $pa^* \in \mathcal{PA}\backslash\{pa\}$:

   - *Action:* perform the intervention $do(PA = pa^*)$.
   - *Prediction:* compute the potential response of $\widehat{G}$ to action, i.e., $\widehat{G}_{pa^*}(\mathbf{u_e})$.

   Once we have found the glucose trajectories that align with the observed evidence for each physical activity plan $pa^* \in \mathcal{PA}\backslash\{pa\}$, we determine the set of counterfactual physical activity plans as the hypothetical values of PA that could have prevented T2D at t_{end}, i.e.,

   $$\mathcal{PA}^*_{\mathbf{u_e}} = \left\{ pa^* \in \mathcal{PA}\backslash\{pa\} \mid \widehat{G}_{pa^*}(\mathbf{u_e})|_{t=t_{\text{end}}} < thr_{\text{T2D}} \right\}, \tag{5.7}$$

   and we determine the least demanding plan $pa^{*\min}_{\mathbf{u_e}}$, characterized by the equivalent value

   $$\overline{pa}^{*\min}_{\mathbf{u_e}} = \min_{pa^* \in \mathcal{PA}^*_{\mathbf{u_e}}} \overline{pa}^*. \tag{5.8}$$

3. Lastly, we consider the most conservative physical activity plan with a preventive factor $pa^{*\max}_{\mathbf{e}}$, characterized by the equivalent value

   $$\overline{pa}^{*\max}_{\mathbf{e}} = \arg\max_{\mathbf{u_e} \in \mathcal{U}_e} \overline{pa}^{*\min}_{\mathbf{u_e}}. \tag{5.9}$$

The optimization in Eqs. (5.8) and (5.9) might be extremely challenging: besides exploring a multidimensional domain, it requires an analytical solution of the ODEs model in closed form. Since this solution is not available in the system under study, the counterfactual query is here approximated using a numerical solution[1]. First, a dataset $\mathcal{D}$ that covers k simulation scenarios, considering the conditions in Table 5.2 is created. Then, counterfactual inference is performed as described in Algorithm 3. The computational complexity of the generation phase is $\mathcal{O}(k)$, whereas the complexity of Algorithm 3, for each $\mathbf{e}$, is $\mathcal{O}(|\mathcal{U}_e|)$ when $n_{\text{obs}} = 1$ and $\mathcal{O}(|\mathcal{U}_e| \cdot |\mathcal{PA}\backslash\{pa\}|) \approx \mathcal{O}(|\mathcal{U}_e| \cdot |\mathcal{PA}|)$ otherwise. With N observations of e, the overall complexity is $\mathcal{O}(k + N \cdot |\mathcal{U}_e| \cdot |\mathcal{PA}|)$. Hence, the feasibility of the approach mainly depends on the number of simulations covered by $\mathcal{D}$ and the time required to simulate these scenarios.

Algorithm 3 Counterfactual Inference

Input: A dataset $\mathcal{D}$ of model simulations, a partial evidence $\mathbf{e}$
Output: The value $\overline{pa}_{\mathbf{e}}^{*\max}$

$\mathcal{D}_{\mathbf{e}} \leftarrow$ all rows in $\mathcal{D}$ where $\mathbf{E} = \mathbf{e}$ and $\widehat{G}|_{t=t_{\text{end}}} \geq thr_{\text{T2D}}$ ▷ factual simulations
$\mathcal{U}_{\text{e}} \leftarrow$ all combinations $\mathbf{u_e} \in \mathcal{D}_{\mathbf{e}}$
for all $\mathbf{u_e} \in \mathcal{U}_e$ **do**
    if $n_{\text{obs}} = 1$ **then**
        $\mathcal{PA}^*_{\mathbf{u_e}} \leftarrow$ all $pa^* \in \mathcal{PA}\backslash\{pa\}$ in $\mathcal{D}\backslash\mathcal{D}_{\mathbf{e}}$ compatible with $\mathbf{u_e}$, $\mathbf{e}$, and with $\widehat{G}|_{t=t_{\text{end}}} < thr_{\text{T2D}}$, as defined by Eq. (5.7)
    else
        New initial conditions $IC' \leftarrow$ current output of the ODEs model at time $t_{n_{\text{obs}}-1}$
        for all $pa^* \in \mathcal{PA}\backslash\{pa\}$ **do**
            sim $\leftarrow$ `RunModelSimulation`(IC', $\mathbf{u_e}$, $t_{\text{end}} - t_{n_{\text{obs}}-1}$)
            if sim has $\widehat{G}|_{t=t_{\text{end}}} < thr_{\text{T2D}}$ **then**
                Save pa^* in $\mathcal{PA}^*_{\mathbf{u_e}}$
            end if
        end for
        $\overline{pa}_{\mathbf{u_e}}^{*\min} \leftarrow$ minimum value within $\mathcal{PA}^*_{\mathbf{u_e}}$, defined by Eq. (5.8)
    end if
end for
$\overline{pa}_{\mathbf{e}}^{*\max} \leftarrow$ maximum value of $\overline{pa}_{\mathbf{u_e}}^{*\min}$ across all $\mathbf{u_e}$ by Eq. (5.9)

[1]Code available at: https://gitlab-core.supsi.ch/dti-idsia/Causal_ODEs_T2D

5.4 Results

5.4.1 Simulation Phase

The original (Eq. 5.1) and the homogenized (Eq. 5.5) ODE systems were compared on a set of 19,683 simulations, obtained by varying six model parameters (Ω, τ_{S_I}, θ_{S_I}, ξ, $1/\nu$, δ) and three initial conditions (G_0, I_0 and B_0). The original system was solved implicitly in Python, applying a fifth-order implicit Runge-Kutta method with a maximum time step of 1 hour. This configuration enabled the simulation of a 5-year period of the original system in approximately 35 seconds on a LENOVO workstation equipped with an Intel Core i7-10700K CPU (8 cores, 16 threads, 2.9 GHz) and 32 GB of RAM. By removing the step-size constraint, the homogenized system achieved a substantial computational speed-up, completing the same 5-year simulation in approximately 0.025 seconds, representing a reduction in computational time by a factor of $1/\lambda_t$, which reflects the relationship between the two timescales. Figure 5.2 compares the long-term behavior of the original (Eq. 5.1) and homogenized (Eq. 5.5) systems by evaluating their solutions at discrete time instants marking the start of physical activity, i.e., at $t_k := k\nu$ for $k = 1, \ldots, n$ with $t_n = t_{\mathrm{end}} = 5$ years. The two systems exhibited very similar behavior over time, indicating that the error introduced by the homogenization procedure remained reasonably small and was bounded over time. Hence, through homogenization, the long-term impact of physical activity on blood glucose regulation is preserved while avoiding minute-scale computations, thereby increasing computational efficiency and allowing extensive exploration of multiple simulation scenarios.

A dataset $\mathcal{D}$ composed of $k = 1,814,400$ five-year simulations of the homogenized system was generated to model diverse subjects, accounting for all variable combinations in Table 5.2 and all physical activity plans in Table 5.3. In such dataset, 6.85% of the subjects developed T2D during the simulation period. Of these, 94.1% engaged in physical activity that is insufficient according to the WHO guidelines [160]. Normoglycemic or prediabetic subjects at $t = 0$ were never predicted to progress to T2D in five years if their initial β-cell mass exceeded 1200 mg or if $\theta_{S_\mathrm{I}} \geq 0.28$ ml/μU/d.

Sobol sensitivity analysis [185] was conducted using the SALib Python library [186] to measure the influence of each variable on the output of interest, i.e., $\widehat{G}|_{t=t_{\mathrm{end}}}$. A Random Forest regressor was used as a surrogate of the homogenized model (min-max scaling, 70-30% training-test split, mean absolute error on the test set of 0.004) and input samples for sensitivity analysis were sampled using the Saltelli sampler (N=2048). The variables that most significantly influenced the output according to Sobol sensitivity analysis were

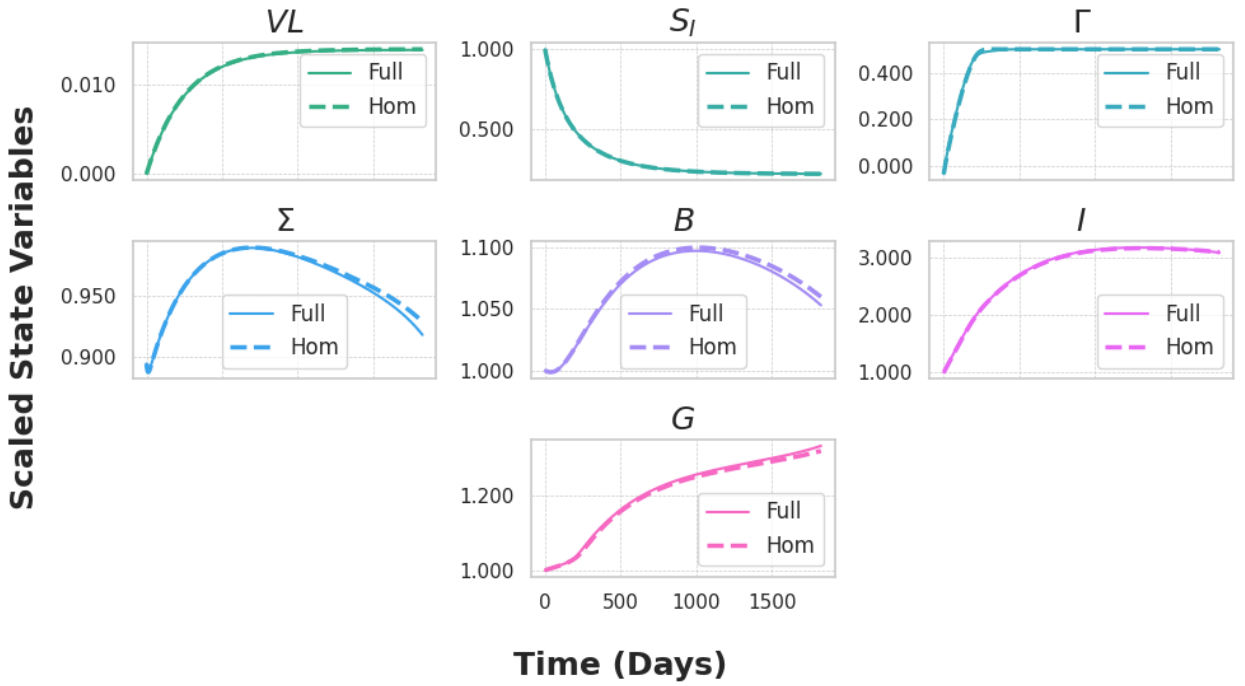


Figure 5.2: Evolution of long-term state variables in the original model (full system, solid lines) and its computationally efficient version (homogenized system, dashed lines) for $t_{\text{end}} =$ 5 years.

θ_{S_I} (0.91), B_0 (0.70), $\overline{PA}$ (0.44), and τ_{S_I} (0.33). The weight Ω presented a lower impact (2.33×10^{-3}), while G_0 and I_0 had almost no impact on the output (total sensitivity $< 1 \times 10^{-6}$). Only a small part of the total variability (e.g., 0.18 for θ_{S_I}) was due to changes in a single input parameter, while most of the total variability was due to interaction terms.

5.4.2 Counterfactual Inference Phase

During the inference phase, we hypothesized a set of potential observations $\mathbf{e}$ ($n_{\text{obs}} = 1$) and ranges of $\widehat{G}|_{t=t_{\text{end}}}$, resulting from all combinations of the following values (432 combinations): $g_0 \in \{70, 120\}$, $\omega \in \{50, 70, 90, 110, 130, 150\}$, and 9 fixed physical activity plans with $1/\nu \in \{1, 2, 3\}$, $\xi \in \{30, 50, 70\}$, $\delta = 60$ minutes and $\widehat{G}|_{t=t_{\text{end}}} \in \{(125, 150), (150, 175), (175, 200), (200, 600)\}$.

Figure 5.3 summarizes counterfactual inference outputs across six dimensions: ω, g_0, $\widehat{G}|_{t=t_{\text{end}}}$, the number of factual simulations ($|\mathcal{D}_{\mathbf{e}}|$), the number of factual simulations with at least one counterfactual ($|\mathcal{D}_{\mathbf{e}}| : \exists\ \mathcal{PA}^*_{\mathbf{u_e}}$), and the median number of counterfactual physical activities for each factual simulation ($|\mathcal{PA}^*_{\mathbf{u_e}}|(median)$). Lines are colored by factual pa, with ribbon thickness reflecting the number of observations along each path. A total of 280 out of 432 observations e yielded valid factual simulations in the specified range of $\widehat{G}|_{t=t_{\text{end}}}$, with a

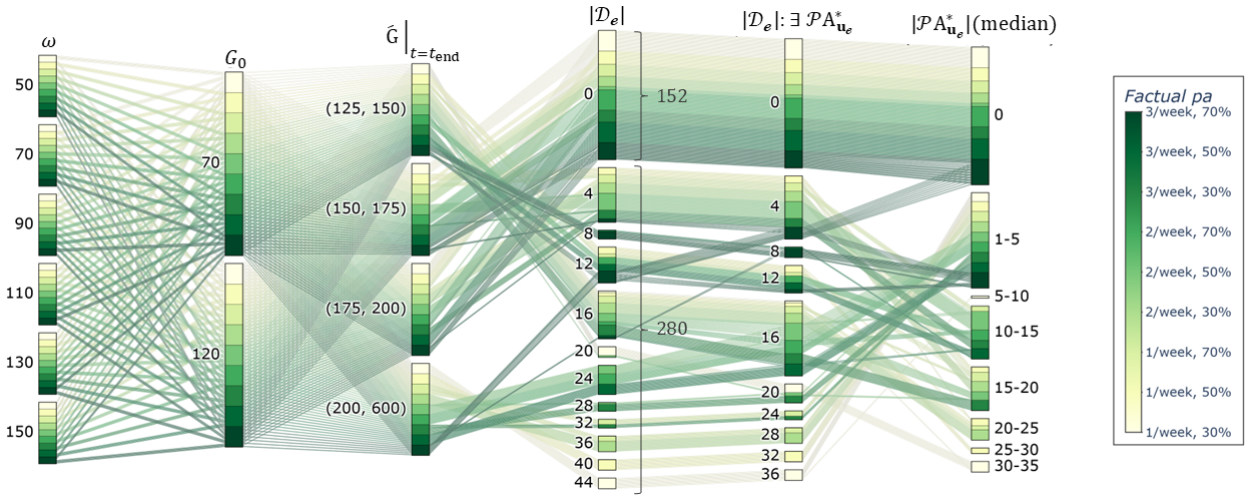


Figure 5.3: Parallel categories plot that summarizes counterfactual inference across multiple dimensions, as a function of the factual *pa*.

median of 16 factuals each (min=4, max=44). These simulations, in turn, yielded a median of 11 counterfactual physical activities each (min=0; max=51). Additionally, 108 out of 280 observations with valid factuals presented at least one simulation without a corresponding counterfactual physical activity plan. These factual simulations represented subjects whose progression to T2D could not be prevented by modifying physical activity alone, as their glucose metabolism was already compromised. They were characterized by a rapid decay of insulin sensitivity ($\tau_{S_\mathrm{I}} = 90$ days, $\theta_{S_\mathrm{I}} = 0.18$ ml/μU/d) and a reduced β-cells mass ($B_0 = 800$ or 900 mg), resulting in hyperglycemia at the end of the observation period ($\widehat{G}|_{t=t_{\mathrm{end}}}$ between 200 and 600 mg/dl).

Table 5.4 describes the effect of *pa*, B_0 and τ_{S_I} on the least demanding counterfactual plan, $\overline{pa}^{*\min}_{\mathbf{u_e}}$, across the set of factual simulations. Both frequency and intensity of the factual *pa* influence $\overline{pa}^{*\min}_{\mathbf{u_e}}$, with a greater demand when the observed intensity and frequency are higher. Moreover, a lower β-cells mass and lower values of τ_{S_I}, indicating a faster decrease in insulin sensitivity, require increasing physical activity to prevent T2D onset. The analysis of $\overline{pa}^{*\max}_{\mathbf{e}}$, that is, the average value associated to the plan ensuring T2D prevention across all compatible exogenous node values $\mathbf{u_e}$, shows that $\overline{pa}^{*\max}_{\mathbf{e}}$ heavily depends on the factual *pa* and on $\widehat{G}|_{t=t_{\mathrm{end}}}$. However, the calculation of $\overline{pa}^{*\max}_{\mathbf{e}}$ as in Eq. (5.9) uses a conservative approach which can lead to an over-pessimistic estimate, i.e., suggesting the most intense physical activity if e is not sufficiently informative. More measures of glucose, i.e., $n_{\mathrm{obs}} > 1$, allow a better estimation of exogenous parameters and $\overline{pa}^{*\max}_{\mathbf{e}}$, as illustrated in the followings.

Table 5.4: Effect of factual pa, B_0 and τ_{S_I} on $\overline{pa}^{*\min}_{\mathbf{u_e}}$ (median and IQR).

Factual *pa*			**B_0**		**τ_{S_I}**	
2/week 50%	**3/week 50%**	**3/week 70%**	**800 mg**	**1100 mg**	**90 days**	**330 days**
1.04 (0.63)	1.36 (0.59)	1.63 (0.44)	1.36 (0.86)	0.81 (0.10)	1.63 (1.07)	0.71 (0.30)

Example: Individual response to WHO-compliant physical activity. The response of different individuals to the same physical activity plan may vary depending on genetic and acquired factors that could affect their degree of insulin sensitivity and β-cells functionality. Indeed, an individual who is prediabetic and follows the WHO guidelines [160] could still progress to T2D according to the model. As an example, we simulated individuals with a weight of 70 kg who engaged in vigorous-intensity aerobic physical activity for 90 minutes per week, with observed prediabetic measurements at six months ($n_{\mathrm{obs}} = 2$ with $t_1 = 0.5$ years, left-hand side) and at one year ($n_{\mathrm{obs}} = 3$ with $t_1 = 0.5$ and $t_2 = 1$ year, right-hand side). Figure 5.4 shows the set of possible glucose trajectories of such individuals. Some individuals may not develop T2D within the next five years, thus not requiring intervention (Group A in Figure 5.4, corresponding to simulations with $B_0 \geq 1100$ mg and $\tau_{S_\mathrm{I}} \geq 150$ days, $B_0 = 1000$ mg and $\tau_{S_\mathrm{I}} \geq 210$ days, or $B_0 = 900$ mg and $\tau_{S_\mathrm{I}} \geq 330$ days). For most of the simulated individuals who develop T2D disease progression could have been prevented (Group B). However, in a subset of subjects, modifying physical activity alone would have been insufficient to reverse T2D progression, mainly due to impaired insulin sensitivity and β-cells dysfunction (Group C, corresponding to simulations with $B_0 \leq 1000$ mg and $\tau_{S_\mathrm{I}} = 90$ days or $B_0 = 800$ mg and $\tau_{S_\mathrm{I}} \leq 150$ days).

Figure 5.5 illustrates preventable factual glucose trajectories selected from Group B of

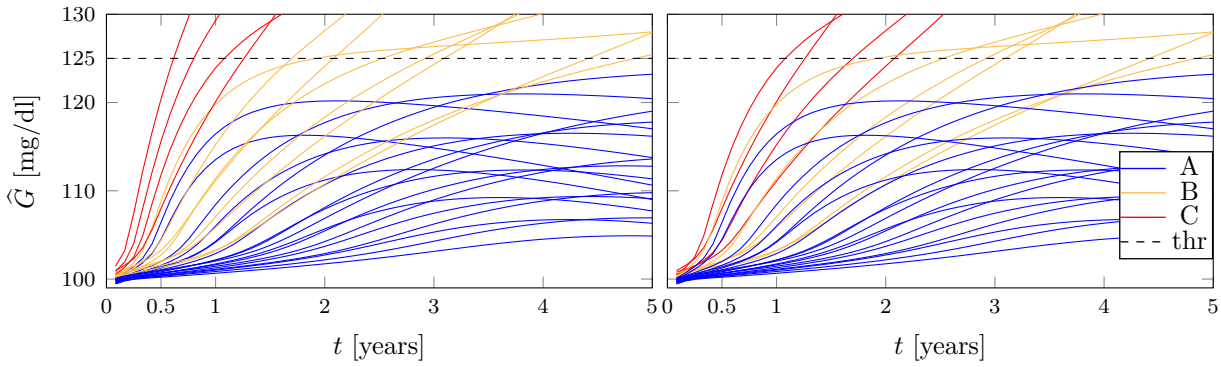


Figure 5.4: Factual glucose trajectories as a function of insulin sensitivity and β-cells mass representing prediabetic trajectories (Group A), reversible progression to T2D (Group B) and non-reversible progression to T2D (Group C) over a span of five years.

Figure 5.4 (yellow lines) and the corresponding counterfactual trajectories (gray lines). Given the same evidence, the progression towards T2D is considerably influenced by the degree of insulin sensitivity impairment and β-cells dysfunction, particularly in terms of defects in β-cells mass due to increased apoptosis. The faster a patient is expected to progress to T2D, the more limited is the range of potential counterfactual plans, making it more challenging to prevent the disease through exercise. For example, a subject with $B_0 = 900$ mg and $\tau_{S_\mathrm{I}} = 270$ days is predicted to develop T2D at t=5 years, as shown in the bottom-left panel of Figure 5.5c. At t=0.5 years, the progression can be prevented with a minimum of 90 minutes of exercise twice per week at 40% intensity. For the same B_0 but a lower τ_{S_I} (210 days, center-left panel of Figure 5.5c), the subject is expected to develop T2D in approximately three years. At t=0.5 years, prevention is still possible, but it requires more intense physical activity (e.g., 90 minutes of exercise, two times per week, at 70%). For an even lower τ_{S_I} (150 days, right-left panel of Figure 5.5c), the subject is predicted to develop T2D in less than two years. In this case, halting progression would require maximum effort (e.g., 90 minutes of exercise, three times per week, at 70%).

Lastly, it is evident that additional measures of the glucose trajectory (e.g., $n_{\mathrm{obs}} = 2$ or $n_{\mathrm{obs}} = 3$) lead to less conservative estimates, facilitating the customization of physical activity plans to individual needs. Indeed, two additional glucose measurements at six months and one year (right panels) reduce the number of compatible factual trajectories (Figure 5.5) and the number of potential counterfactual plans compared to a single additional measurement at six months (left panels). In turn, the identification of more specific counterfactual physical activity plans avoids creating exercise overload for the individual.

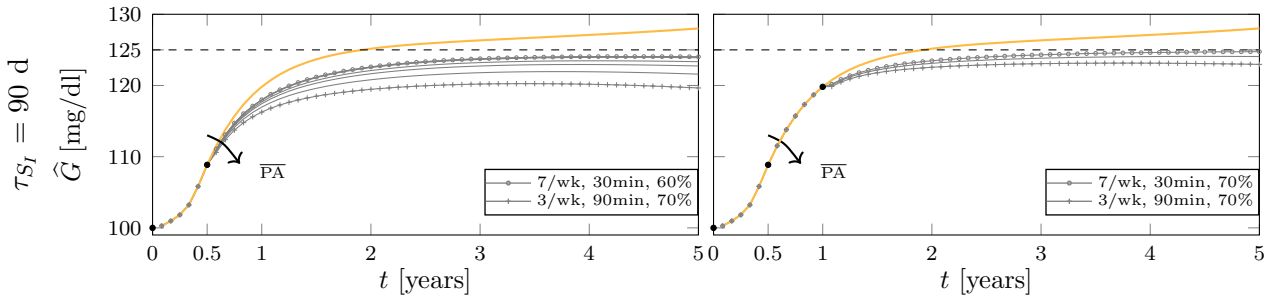
5.5 Discussion

This Chapter presents a numerical counterfactual inference approach based on a large set of simulations and grounded on parametric SCMs, that explicitly accounts for causal relationships in longitudinal settings. The proposed approach was applied to a diabetes progression model [52] that provides a mechanistic description of blood-glucose regulation through ODEs. The objective of counterfactual inference was to examine the impact of tailored physical activity recommendations (i.e., fixed duration, intensity and frequency of exercise) on T2D progression in individuals at risk, e.g., those with prediabetes. The analysis used patient-specific parametrization of the ODEs model (Tables 5.2-5.3) to examine variability arising from differences in pancreatic β-cells functionality and insulin sensitivity. Impairments in these mechanisms, both genetic and acquired (e.g., inflammation, oxidative stress, free fatty acid levels, gluco- and lipo-toxicity), can affect insulin secretion and cells responsiveness to

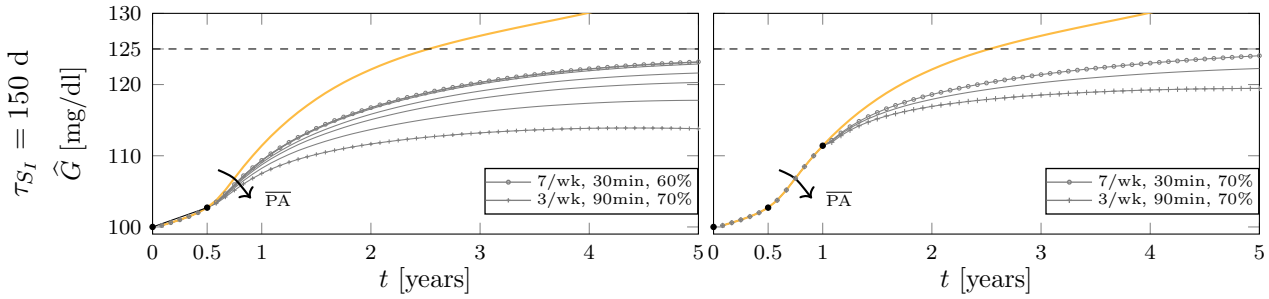
insulin, significantly contributing to the pathogenesis of T2D [187]. A homogenization approach was applied to approximate the disease progression model (see Section 5.2), enabling accurate and fast simulation of the long-term effects of physical activity on glucose regulation without resolving minute-scale dynamics. This substantially reduced the computational cost (i.e., from 35 to 0.025 seconds/simulation), making it feasible to explore a wide range of simulation scenarios.

Three groups of five-year glucose trajectories were analyzed (Figure 5.4). Particular attention was given to Group B, composed of individuals whose progression to T2D could have been prevented via customized physical activity plans. Within this group, the availability of longitudinal measures allowed for a more precise definition of counterfactual physical activity plans (Figure 5.5). Ultimately, this analysis evaluated the effectiveness of a physical activity plan a posteriori, offering insights for tailored suggestions in similar future subjects. Such kind of personalization could better inform clinical decisions, quantifying heterogeneity in treatment effect compared to current guidelines and optimizing patient benefits, as highlighted by recent dose-response studies (e.g., [188]).

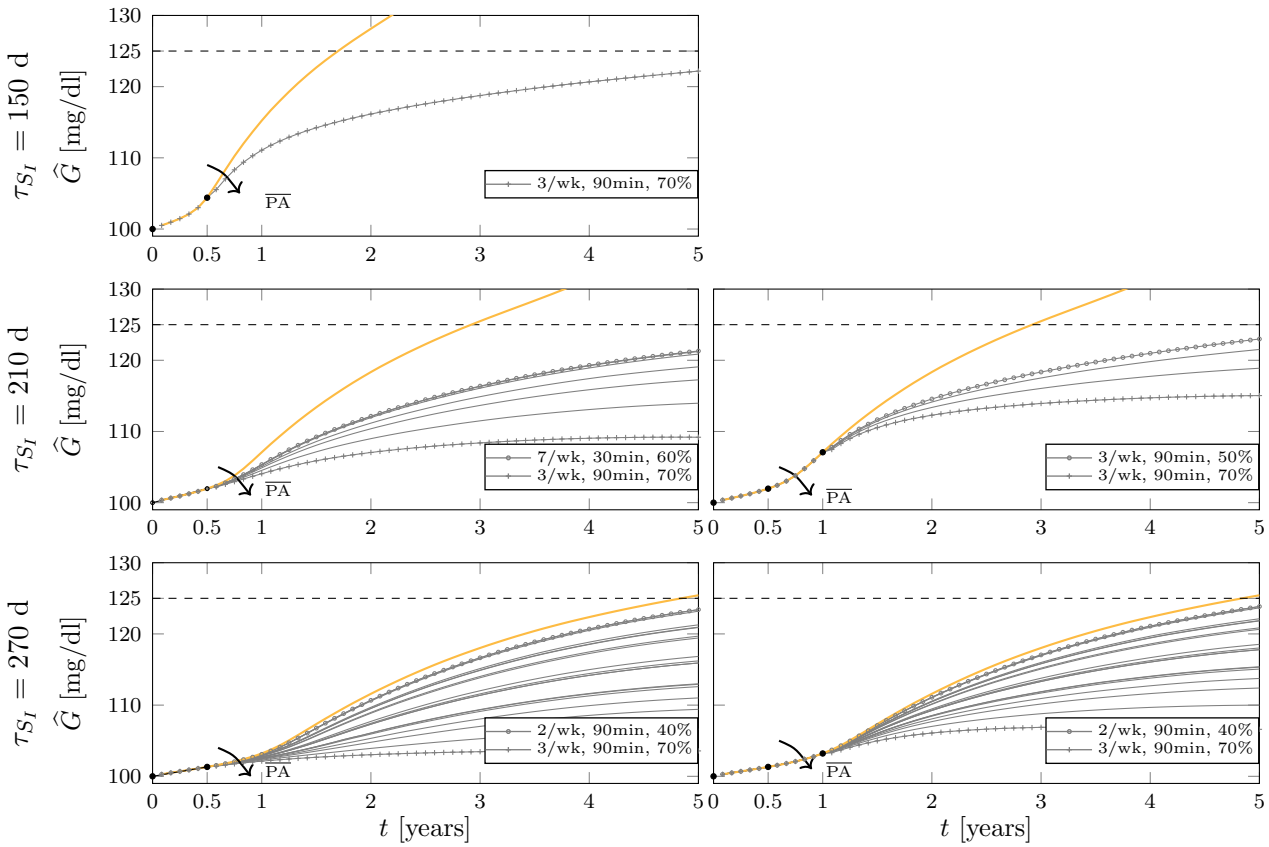
The study presents some limitations. First, a restricted set of initial conditions and model parameters was varied over a limited grid to reduce computational complexity. Hence, the results rely on an implicit assumption of uniformity across the sampled points, which is likely unrealistic. While the current parametric SCM approach allowed systematic exploration of the feature space, performance could improve if the underlying probability distributions of the parameters were estimated from data. Future studies could also consider a broader range of initial conditions (e.g., insulin sensitivity at $t = 0$), and using a finer sampling grid to extend the explored feature space. Additionally, the proposed approach was subject to the same limitations of the underlying diabetes progression model, e.g., its inability to differentiate between subjects with different age, sex assigned at birth or diet. To address these limitations, additional features need to be incorporated and evaluated. Lastly, both the model and the retrieved counterfactuals should be validated with real data based on clinical trials. A validation study could involve subjects at risk of developing T2D by t_{end} according to the ODEs model. Identified subjects would be recommended a counterfactual physical activity plan and their progress would be monitored to assess the plan's preventive effectiveness. Implementing personalized treatments may face challenges like patients' motivation for long-term exercise and socioeconomic obstacles. Thus, it would be crucial to involve physical therapists in the design of plans that meet patients' needs.



(a) $B_0 = 1100$mg

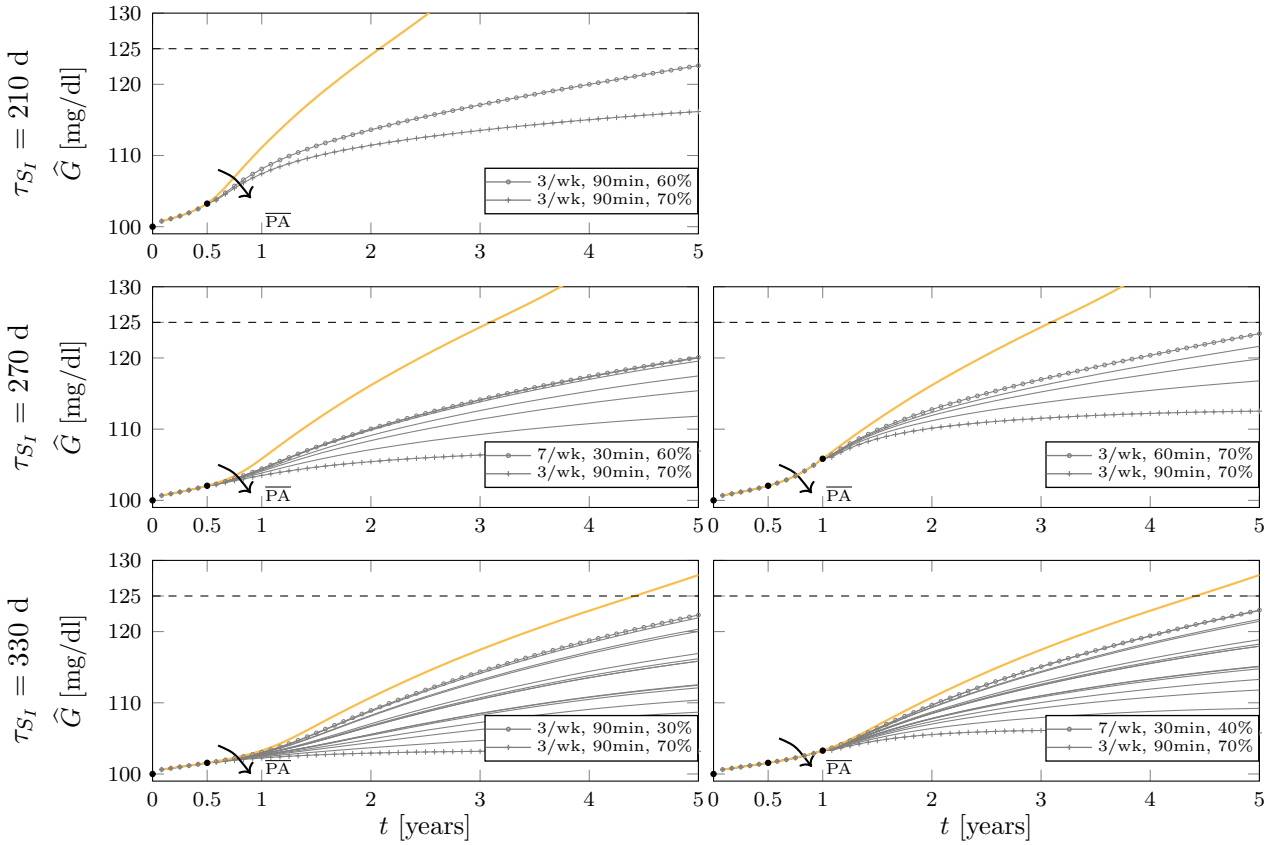


(b) $B_0 = 1000$mg



(c) $B_0 = 900$mg

Figure 5.5: Glucose trajectories for individuals with $\omega = 70$ kg and *pa* of 90 min, 1/wk at 70% intensity, starting from $g_0 = 100$ mg/dl and showing prediabetic measurements at six months ($n_{\text{obs}} = 2$ with $t_1 = 0.5$ years, left panels) and at one year ($n_{\text{obs}} = 3$ with $t_1 = 0.5$ and $t_2 = 1$ year, right panels). Examples of factual glucose trajectories of individuals progressing to T2D over a span of five years (yellow lines) with their counterfactual trajectories, corresponding to the effect of all alternative physical activities reverting to prediabetes (gray lines).



(d) $B_0 = 800$mg

Figure 5.5: (Continued) Glucose trajectories for individuals with $\omega = 70$ kg and *pa* of 90 min, 1/wk at 70% intensity, starting from $g_0 = 100$ mg/dl and showing prediabetic measurements at six months ($n_{\text{obs}} = 2$ with $t_1 = 0.5$ years, left panels) and at one year ($n_{\text{obs}} = 3$ with $t_1 = 0.5$ and $t_2 = 1$ year, right panels). Examples of factual glucose trajectories of individuals progressing to T2D over a span of five years (yellow lines) with their counterfactual trajectories, corresponding to the effect of all alternative physical activities reverting to prediabetes (gray lines).

Chapter 6

Structural Counterfactual Explanations for T2D prevention

6.1 Bridging XAI and causality

Even though they were born and evolved separately, XAI and causal learning are regarded as deeply interconnected fields. Both strive to make AI systems more trustworthy, either by explaining their decisions in a human-comprehensible way, or by modeling the underlying causal mechanisms of the modeled phenomenon through structural assumptions grounded in domain or data-acquired knowledge (see Chapters 1 and 2). As summarized by a recent review by Carloni et al. [12], literature investigates the interplay between these two fields from different perspectives, which can be summarized into three main viewpoints:

- **lack of causality as a major limitation of XAI approaches**. A set of publications focus on the fact that most of the current XAI techniques rely on extracting associative relationships from data, lacking a foundation in the theory of causality, leading to potentially unfeasible or erroneous decisions in practice [48]. Indeed, there is no guarantee that the explanations provided reflect true causal relationships between input data and suggested decisions. Even a highly accurate input-output mapping learned from data may rely on spurious correlations between features and outcomes. Achieving specific performance metrics does not necessarily imply that an XAI system has captured the underlying mechanisms of a phenomenon [189], and apparent associations can be misleading (e.g., as demonstrated by the Simpson's paradox[190, 73]), hampering fairness and transparency [191];

- **causality as a means to get to XAI**. This second perspective is driven by the idea that causality is propaedeutic to XAI. Indeed, some studies suggest that causal learning

tools such as SCM, do-operator, and causal metrics (e.g., probability of necessity and probability of sufficiency) can be leveraged to revisit and empower existing XAI methods (e.g.,[192, 49, 69, 72]). Others, instead, argue that causal models are inherently interpretable, as estimating the causal model underlying a certain phenomenon and represent it through a DAG provides an intrinsic explanation of its inner functioning (e.g., [193]);

- **XAI as a means to get to causality**. A third, albeit limited, body of work identified by Carloni et al. [12] is based on the idea that XAI can promote causal understanding of a phenomenon. More specifically, XAI techniques could be used to generate hypotheses about potential causal relationships from data, to be futher investigated by domain experts [18, 194].

Purely data-driven XAI methods, such as counterfactual explainers (e.g., MUCH as presented in Chapter 3), may fail to capture the complex causal chains underlying an intervention, as they typically focus on isolated feature perturbations which may not translate into feasible actions in real-world settings. Such approaches implicitly assume that factors act independently, whereas real-world variables are tightly coupled through causal dependencies [195]. Consequently, incorporating causal knowledge into the generation of Counterfactual Explanations (CEs), for example by constraining admissible feature changes according to a SCM, can improve their practical feasibility and represent a step toward AI-driven, actionable recommendations.

Building on the potential of causal learning tools to improve the feasibility of XAI techniques, this Chapter presents an exploratory study aimed at integrating causal constraints into the generation of purely data-driven CEs for chronic disease prevention. As in Chapters 4-5, T2D prevention models will be examined as an application scenario (*Case Study 2*). In particular, an observational dataset capturing transitions between glycemic states (normoglycemia-NG, prediabetes-PD, and T2D) will be analyzed, extending the work presented in [196]. To support prevention strategies, CEs will be specifically computed for factual observations that represent a transition from prediabetes to T2D, with the aim of identifying modifiable factors that would need to change for a patient to regress to normoglycemia rather than follow the observed progression to T2D.

6.2 Clinical context

Prediabetes is a reversible clinical condition characterized by impaired fasting glucose or impaired glucose tolerance, which often precedes the onset of T2D [135]. Although prediabetes

does not always progress to T2D, it should be regarded as a serious health condition, closely linked to metabolic syndrome and associated with an elevated risk of cardiovascular disease, stroke, and mortality [197]. Hence, the timely detection of prediabetes and the implementation of interventions aimed at promoting reversion to normoglycemia are of paramount importance.

While there are abundant studies exploring T2DM and its management strategies (see Section 4.1), the attention given to prediabetes remains relatively low [198, 199]. The prediabetic population is often underrepresented in predictive modeling efforts due to the subtlety of its symptoms. A further challenge contributing to this research gap is the limited availability of high-quality, comprehensive prediabetes datasets. This scarcity of data poses a significant barrier to the development and validation of robust predictive models specifically designed for prediabetes prognosis and intervention planning. Given the scarcity of datasets specifically focused on prediabetes, more general-purpose datasets can be extracted from primary care EHRs and leveraged to develop predictive models, providing a sufficiently broad population and rich clinical information.

6.3 Study dataset

The study dataset was extracted ad hoc from the CPCSSN database (see Appendix B), with the aim of analyzing transitions between the three glycemic states (NG, PD, T2D) using retrospective data, as described in the published study reported in [196]. The dataset included patients above 18 years of age who had at least two distinct measurements in time of FPG, glycated hemoglobin (HbA1c), or 2-hour Plasma Glucose (2hPG) and it was created as follows:

- *Step 1: Glycemic states labeling.* Each measure of FBS, HbA1c, or 2hPG was given a label (NG, PD, or T2D). The measure was labeled as T2D if it occurred less than six months before the date of T2D diagnosis, the measure was labeled as NG if the biomarker was in the normoglycemic range as defined by the American Diabetes Association guidelines [136] (i.e., $FPG < 5.6$ mmol/L, $HbA1c < 5.7\%$ or $2hPG < 7.8$ mmol/L)[1], otherwise it was labeled PD. Encounters occurring within a one-week interval were merged into a single glycemic measure, with the assigned label reflecting the most severe state among the combined encounters (T2D > PD > NG).

[1]There is no universally accepted diagnostic criterion for prediabetes, as diagnostic tests and threshold values vary among organizations. For the purposes of this Thesis, the ADA definition was adopted, in light of the sociodemographic and lifestyle similarities between the Canadian population considered and the American population.

- *Step 3: Stable states definition.* For each patient, two or more consecutive measures with the same label identified a "stable" glycemic state.

- *Step 4: Fluctuations removal.* Criteria based on the time distance between stable glycemic states were introduced to determine if an encounter between two states should be discarded. The rationale was to disregard single measurements that were likely to be associated with glycemic fluctuations. After this step a dataset where each record contained a glycemic state was computed, as shown in Table 6.1. Each record represented a reversible, stable, glycemic state that went from *StartDate* to *EndDate*. *CurrentState* contains the glycemic condition in between the two dates (i.e., NG or PD), while *FutureState* contained the glycemic condition of the same patient in the next stable state (i.e., NG, PD, or T2D).

- *Step 5: Features extraction.* For each *CurrentState* (i.e., NG or PD), a set of routinely measured clinical features (see features list in Section 4.2) was extracted in the corresponding time window. All the data falling within the start and end dates (i.e., *StartDate* and *EndDate* columns, respectively) of that particular state was considered. Entries which fell outside of the acceptable ranges specified in Table B.2, defined as in [200], were removed. In cases where more than one value of a given feature was found in the same time window, the values were averaged. The age of a patient was calculated as the difference in years between the central date of the time window and the patient's birth date.

This Chapter focused on a subset of the dataset consisting of 6329 subjects who are currently in a prediabetic state (*CurrentState*=PD) and hence at elevated risk of developing the chronic condition (T2D), yet still within a reversible, prevention window. Individuals in this condition may either progress to type 2 diabetes (*FutureState*=T2D) or revert to normoglycemia (*FutureState*=NG), as shown in Table 6.1. The resulting dataset is summarized in Tables 6.2 (numerical features) and 6.3 (categorical features). Numerical features are represented in terms of median value and interquartile range (IQR), whereas categorical features are represented in terms of percentage of samples for each category.

Patient_ID	**CurrentState**	**StartDate**	**EndDate**	**FutureState**
1001000000001377	PD	05/05/2005	01/31/2007	T2DM
1001000000001178	PD	07/08/2008	09/27/2010	NG

Table 6.1: Illustration of the dataset structure capturing all stable glycemic state transitions, generated after completion of Step 4.

Feature	NG (N= 4862)	T2D (N= 1467)	Clinically plausible range limits
Age	61.38 (53.66-69.71)	59.93 (51.98-68.13)	-
BMI [kg/m^2]	28.30 (25.29-31.80)	31.75 (28.20-36.25)	15.00-35.00
Pressure [mmHg]	SBP: 128.64 (120.00-137.05) DBP: 77.20 (71.33-82.00)	SBP: 130.82 (123.00-139.50) DBP: 79.38 (74.00-84.25)	SBP: 70.00-160.00 DBP: 60.00-100.00
FPG [mmol/L]	5.73 (5.60-5.90)	6.12 (5.87-6.42)	3.20-9.00
HDL [mmol/L]	1.33 (1.12-1.60)	1.17 (1.00-1.38)	1.00-2.00
LDL [mmol/L]	3.05 (2.44-3.64)	2.89 (2.26-3.45)	1.50-5.20
TG [mmol/L]	1.35 (0.98-1.87)	1.69 (1.24-2.29)	0.50-3.50
Total Cholesterol [mmol/L]	5.12 (4.46-5.82)	4.94 (4.30-5.60)	4.20-6.20

Table 6.2: Distribution of numerical features (median and IQR), grouped by future glycemic state, and clinically plausible range limits used to constraint counterfactual explanations.

Feature	NG (N= 4862)	T2D (N= 1467)
Sex assigned at birth	female: 55% male: 45%	female: 54% male: 46%
Smoke	60% never 28% ex 12% current	55% never 29% ex 16% current
Antidepressants	no: 95% yes: 5%	no: 96% yes: 4%
Corticosteroids	no: 93% yes: 7%	no: 91% yes: 9%
Antihypertensives	no: 61% yes: 39%	no: 50% yes: 50%
Cholesterol lowering medications	no: 67% yes: 33%	no: 56% yes: 44%
Quit-smoking medications	no: 96% yes: 4%	no: 95% yes: 5%
Hypertension (HTN)	no: 57% yes: 43%	no: 55% yes: 45%
Chronic Obstructive Pulmonary Disease (COPD)	no: 95% yes: 5%	no: 94% yes: 6%
Depression	no: 83% yes: 17%	no: 83% yes: 17%
Osteoarthritis (OA)	no: 79%% yes: 21%	no: 81% yes: 19%

Table 6.3: Distribution of categorical features, grouped by future glycemic state.

6.4 Methodology

The dataset set was randomly splitted in training ($\mathcal{D}_{tr}$, 80%), and test set ($\mathcal{D}_{ts}$, 20%), with stratification. Continuous features were normalized in the $[0, 1]$ range (min-max scaling), whereas categorical features were one-hot encoded.

Causal knowledge was integrated in the CEs generation process by means of the CEILS method proposed by Crupi et al. [72]. This methodology pursues increased practical feasibility of CEs by leveraging cause-effect relationships encoded by an SCM under the causal sufficiency assumption, i.e., assuming that the given causal graph is complete and fully informative. Its main advantage is that it can be applied on top of any existing counterfactual explainer, remaining agnostic to the method used to generate explanations. The CEILS methodology consists of two main steps [72]: (i) construction of a causal pipeline, guided by an SCM, that maps each input variable $X_i \in \mathbf{X}$ to output Y in a latent space; (ii) generation of CEs as shifts in the latent space, capturing actions that change the predicted output while respecting the underlying causal flow. These actions act as soft intervention on certain variables, superimposed on the changes induced by the variable's parent nodes. Finally, the latent representation of CEs is converted back to the original feature space.

Step (i). First, we built a SCM from a set of causal relationships among input variables $\mathbf{X}$, encoded in a DAG, and a set of structural equations measuring the strength of such relationships.

Since the set of variables under consideration and their temporal ordering was identical to that used in Chapter 4, the same causal relationships encoded by DAG_{all} (Figure 4.2a) were used, while allowing the structural equations to differ. Specifically, the set of structural equations of the SCM was inferred from data in $\mathcal{D}_{tr}$ through Deep Neural Networks (DNNs), following [72]. DNNs were trained to estimate the value of each feature X_i from its causal parents $Pa(X_i)$. Then, each exogenous variable U_i, modeling unobserved conditions for X_i, was approximated with the residuals $\hat{U}_i = X_i - \hat{X}_i$, with $\hat{X}_i = DNN_i(Pa(X_i))$. Once all DNNs were learned in a topological order following causal dependencies encoded by the DAG (from root nodes to children nodes), the set of approximated structural equations ($\hat{\mathcal{F}} : \hat{\mathbf{U}} \rightarrow \hat{\mathbf{X}}$) was obtained. Hence, under the ANM assumption (see Section 2.2.3), the value of each node i was recursively estimated from the residuals through the approximated structural equation $\hat{f}_i \in \hat{\mathcal{F}}$, namely $\hat{X}_i = \hat{f}_i(U)$, with

$$\hat{f}_i(U) := \begin{cases} \hat{U}_i & \text{if } i \, root \, node \\ DNN_i(\{\hat{f}_p(U)\}_{p \in Pa(X_i)}) + \hat{U}_i & \text{if } i \, non-root \end{cases} \tag{6.1}$$

Compared to the original formulation proposed in [72], which modeled each structural equation through regression using a simple two-layer fully connected DNN trained with mean-squared error (MSE) loss, assuming all variables were continuous, this analysis explicitly handled mixed-type data. Here, continuous and categorical variables were treated in a separate way: continuous variables were still modeled with a regression network trained using MSE, while categorical variables were modeled through a softmax classifier made of a two-layer fully connected DNN trained using sparse categorical cross-entropy loss. Additionally, each per-node model included batch normalization and dropout for regularization (dropout rate= 0.1), allowing for more robust and stable training. All models were trained using Adam optimizer with a learning rate of 1e−4 and ReLU activation function. Also, residuals were explicitly aligned prior to concatenation to ensure correct ordering.

Subsequently, a classifier C (risk prediction model, $C : \mathbf{X} \rightarrow Y$) was modeled as a fully connected feed-forward DNN with four hidden layers of decreasing width (128, 64, 32, and 16 neurons, respectively), using ReLU activations, and a softmax output layer. Batch normalization, and dropout (dropout rate=0.3) layers were added between hidden layers for regularization. The model was trained using the Adam optimizer (learning rate = 0.001) and sparse categorical cross-entropy loss. Compared to the original model in [72], which had only two hidden layers and no regularization, the architecture here used was deeper and more robust to outliers. Lastly, models $\hat{\mathcal{F}}$ and C were composed to yield a model C_u, $(C_u : \mathbf{U} \rightarrow Y)$, that was able to predict the classification output (i.e., *FutureState* from the residuals), by following the causal flow of the underlying DAG.

Step (ii). Following the completion of step (i), step (ii) was carried out as follows. Model C_u was used to compute causally constrained CEs in the residual latent space for a certain observation in the test set ($\boldsymbol{x} \in \mathcal{D}_{ts}$) whose predicted $FutureState$ is T2D. The CEILS generation procedure is structured in three steps [72], in accordance with Pearl's counterfactual computation [96]:

- estimation of latent variables $\mathbf{u}$ from the original sample $\boldsymbol{x}$, i.e., $\mathbf{u} = \hat{\mathcal{F}}^{-1}(\boldsymbol{x})$ (abduction);
- generation of a CE ($\mathbf{u}^{cf} = \mathbf{u} + \delta$), computed as the minimal shift δ in the latent space able to change the prediction of the underlying classifier C_u from $FutureState = T2D$ to $FutureState = NG$ using a certain counterfactual explainer (i.e., DiCE), (intervention);
- computation of the corresponding CE in the original feature space, i.e., $\boldsymbol{x}^{cf} = \hat{\mathcal{F}}(\mathbf{u}^{cf})$ (prediction).

Experimental settings. Experiments were performed using `DiCE` [65] as the underlying counterfactual explainer, adopting its two model agnostic variants (random search and genetic search) as well as its gradient-based variant (see Section 2.1.2).

Two types of constraints were considered when generating CEs. The first type enforces *compliance with the underlying causal structure*. In this regard, CEs were generated either with causal constraints by applying the CEILS method on top of DiCE ($\texttt{CEILS}_{random}$, $\texttt{CEILS}_{genetic}$, and $\texttt{CEILS}_{gradient}$) or without causal constraints using the original DiCE formulation ($\texttt{DiCE}_{random}$, $\texttt{DiCE}_{genetic}$, $\texttt{DiCE}_{gradient}$). The second type governs *feature mutability and the permitted range of variation*, reflecting domain knowledge about clinically plausible value ranges and granting the actionability property. In this regard three configurations were considered: no constraints (*NC*); domain-specific feature constraints without explicit range limits, where the permitted range of variation was bounded only by the training set values (*FC-TR*), domain-specific features constraints, including clinically plausible range limits (*FC-CLIN*).

For *FC-TR* and *FC-CLIN*, the following domain-specific feature constraints were considered: age and sex assigned at birth were considered as immutable features, diagnosed medical conditions were considered as chronic (e.g., an individual can develop a given condition but they cannot revert from a diseased state to a healthy state regarding such condition), smoking habits were considered as partially-modifiable and were permitted to vary only in certain directions (i.e., an individual who smokes may be able to quit smoking and become ex smoker but they cannot transition to a "never" label regarding smoking status). For *FC-CLIN*, the clinically plausible range limits of each variable are reported in Table 6.2, reflecting values associated with a potential deterioration in health.

Evaluation. In summary, 18 configurations were evaluated, reflecting all combinations of three search strategies (random vs genetic vs gradient-based search), the absence or presence of constraints enforcing compliance with the underlying causal structure (DiCE vs CEILS+DiCE), and three configurations reflecting different feature mutability and permitted range of variation (*NC* vs *FC-TR* vs *FC-CLIN*).

As anticipated, the analysis focused on test set observations predicted to progress from PD to T2D (i.e., records with $CurrentState = PD$ and $FutureState = T2D$), concentrating specifically on CEs that could improve the patient's health status for prevention purposes, therefore reverting the output class from T2D to NG (i.e., CEs with $CurrentState = PD$ and $FutureState = NG$). To ensure a fair comparison, we included only factual samples for which all the 18 configurations produced a CE, resulting in a set of N = 56 common valid samples. To quantify the magnitude of change induced by CEs across different configurations

for a given search strategy, a global mixed distance between each factual instance and its corresponding CE was computed by aggregating changes across numerical and categorical features. For numerical features, the distance was computed as the L1 distance in the normalized feature space. For categorical features, the distance was defined as the number of feature-wise mismatches between factual and counterfactual instances. The final mixed distance was obtained as a weighted sum of numerical and categorical components, ensuring normalization to the $[0, 1]$ range.

All statistical analyses followed a within-subject (paired samples) design. Normality of the distributions was assessed independently for each condition using the Shapiro–Wilk test. Since the normality assumption was violated, the Friedman test, i.e., a non-parametric repeated-measures test, was applied as a global test. When the global test indicated statistical significance, post-hoc pairwise comparisons were conducted using Wilcoxon signed-rank tests. A significance level $\alpha = 0.05$ was considered for statistical comparisons and Holm-Bonferroni correction was applied to correct for multiple comparisons.

The generated CEs were evaluated in terms of availability, discriminative power, proximity, sparsity, and causal feasibility. Availability, discriminative power, and sparsity were computed as in Section 3.3.3. Proximity was computed separately for numerical and categorical features, as described in [72]. Causal feasibility, defined as the extent to which CEs respect the underlying causal structure, was computed as in [69] (Constraint Feasibility Score) using the percentage of CEs satisfying monotonic constraints defined in a reference SCM. Notably, the reference SCM used for evaluating causal feasibility is different from the one used by CEILS to generate explanations. Specifically, CEILS estimates its SCM using DNNs (Step (i)), whereas the reference SCM is a predefined, discretized Bayesian network that serves as ground truth. Features in the reference SCM were discretized via quantile binning (n=50 bins).

6.5 Results

Most DNNs in the set of approximated structural equations $\hat{\mathcal{F}}$ achieved low train and test MSEs, despite sharing the same architecture, with test errors ranging from 0.00 to 0.25 for all equations except $\hat{f}_{\text{HDL}}(U)$ ($\approx$0.64), $\hat{f}_{\text{smoking}}(U)$ ($\approx$0.95), and $\hat{f}_{\text{Pressure}}(U)$ ($\approx$2.34). Train and test error values are closely matched, indicating minimal overfitting, while the few higher-error nodes reflected more complex or noisy relationships, which may require specialized architectures or additional hyperparameter tuning. The risk prediction model C achieved a test accuracy of 79% in predicting the *FutureState*, with a weighted average precision of 80% and a weighted average recall of 79%, demonstrating moderate predictive performance,

as reflected by a Matthews correlation coefficient of 0.43.

The analysis of CEs capable of changing the predicted output from $FutureState = T2D$ to $FutureState = NG$ is reported below. As described in Section 6.4, CEs were obtained using 18 different configurations, reflecting all combinations of search strategies (random, genetic, and gradient-based), the absence (DiCE) or presence (CEILS+DiCE) of constraints enforcing compliance with the underlying causal structure, and three configurations representing different feature mutability settings and permitted ranges of variation (*NC*, *FC-TR*, and *FC-CLIN*).

Table 6.4 reports the distributions of the average changes required by each tested configuration to change the predicted output class. These changes were calculated as the difference between counterfactual and factual values (N=56). Hence, negative values indicate that, on average, improving the future patient's condition (from $CurrentState = PD$ to $FutureState = NG$, rather than progressing to $FutureState = T2D$) would require a reduction in the corresponding biomarker, according to the model. A more detailed analysis examining the percentage of changes in feature values between the factual observations and the corresponding CEs, as well as the direction of these changes (decrease, increase, or negligible variation), is summarized in Tables 6.5-6.7 for the random, genetic, and gradient-based search methods, respectively. Results indicated that CEILS (second row of Tables 6.5-6.7) frequently modified both numerical and categorical features to achieve a change in predicted output, resulting in a higher overall percentage of changes in feature values compared to DiCE (first row of Tables 6.5-6.7). For instance, CEILS altered 100% of the records when considering modifiable continuous features (i.e., BMI, FPG, HDL, LDL, TG, and Total Cholesterol), whereas DiCE primarily varied features such as BMI, FPG, and HDL. Specifically, for BMI, 80.4% (of which 64% corresponded to an improvement while 35.6% to a negligible variation) of test samples showed a change in the counterfactual relative to the factual when using $\texttt{DiCE}_{random}$ with *FC-CLIN* feature-constraints (average change: -9.528 $\pm$11.022 [kg/m^2]). This pattern occurred because CEILS not only captured the independent relationships in the dataset but also respected the interactions among variables encoded in the estimated SCM. It is important to note that these changes also reflected residual estimation errors; for example, a small change (i.e., -0.018$\pm$0.014) was observed in age even when the counterfactual value of age was constrained to match the factual value (*FC-TR* and *FC-CLIN* constraints).

When no constraints on feature mutability were applied (i.e., *NC* column of Tables 6.5-6.7) all features were free to vary, leading to unrealistic changes in non-modifiable features like age and sex assigned at birth. For example, in $\texttt{DiCE}_{genetic}$, age changed in up to 14.29% of the samples, with an increase occurring 87.5% of the time (average change: 2.566 $\pm$10.784

years), while sex assigned at birth changed in 5.4% of the samples. Such behavior was attenuated when applying causal constraints using CEILS, indicating that incorporation of causal constraints mitigated the risk of extracting unrealistic CEs, particularly when the genetic and gradient-based search methods were used.

The Friedman test indicated a significant difference in median mixed distances between counterfactual-factual pairs across different constraints methods (random search: $p=7.16 \times 10^{-62}$; genetic search: $p=1.19 \times 10^{-61}$; gradient based search: $p=1.09 \times 10^{-61}$). Post-hoc pairwise comparisons (Wilcoxon signed-rank tests with Holm-Bonferroni correction) revealed that $\texttt{DiCE}_{random}$ and $\texttt{CEILS}_{random}$ produced significantly different explanations across all matched conditions: *NC* ($p=1.13 \times 10^{-9}$), *FC-TR* ($p=1.03 \times 10^{-9}$), and *FC-CLIN* ($p=1.58 \times 10^{-9}$). Similarly, $\texttt{DiCE}_{genetic}$ and $\texttt{CEILS}_{genetic}$ produced significantly different explanations across all matched conditions: *NC* ($p=1.12 \times 10^{-9}$), *FC-TR* ($p=1.89 \times 10^{-9}$), and *FC-CLIN* ($p=1.43 \times 10^{-9}$). Additionally, $\texttt{DiCE}_{genetic}$ without constraints on feature mutability (*NC*) was significantly different from $\texttt{DiCE}_{genetic}$ with *FC-TR* feature-constraints ($p=0.026$). Finally, $\texttt{DiCE}_{gradient}$ and $\texttt{CEILS}_{gradient}$ produced significantly different explanations across all matched conditions: *NC* ($p=1.89 \times 10^{-9}$), *FC-TR* ($p=1.28 \times 10^{-9}$), and *FC-CLIN* ($p=1.81 \times 10^{-9}$).

Table 6.8 compares the sets of CEs obtained with the 18 configurations here tested in terms of desired properties (availability, discriminative power, proximity, sparsity, and causal feasibility). All methods achieved a very high availability of CEs (minimum availability: 96.99%) despite the imposed generation constraints. Availability decreased under stricter settings, particularly when $FC-CLIN$ constraints were applied, while the highest availability (100.00%) was observed in the absence of both causal and feature-mutability constraints (i.e., DiCE + NC). All methods produced CEs that could be discriminated from points of the factual class with a satisfactory level of accuracy (i.e., discriminative power >96.00% for all methods), with explanations generated using CEILS performing better than the original DiCE formulation in terms of discriminative power. CEILS achieved better (i.e., lower) continuous proximity but worse (i.e., higher) categorical proximity compared to DiCE when no constraints on feature mutability were applied (*NC*), as well as when $FC-TR$ constraints were considered. This trend was consistent with the observation that CEILS frequently modified both numerical and categorical features to achieve a change in predicted outcome (as shown in Tables 6.5-6.7), thereby assigning greater influence to categorical variables. In contrast, DiCE relied predominantly on changes in numerical features, which consequently require larger average shifts in continuous variables. This behavior was also reflected in the sparsity metric. On average, CEILS modified a larger number of features (more than eight) to change the outcome class while respecting the underlying causal flow, whereas the original

DiCE formulation typically focused on minimal changes, altering only one to two features at a time, disregarding causal relationships. Finally, as expected, CEILS achieved better causal feasibility (high average value with reduced standard deviation) compared to DiCE, owing to its ability to generate explanations that adhere to a given causal structure, whereas DiCE relies solely on data-driven changes.

	Features	*NC*	*FC-TR*	*FC-CLIN*
DiCE$_{random}$	Age	3.164 ±9.700	0.000 ±0.000	0.000 ±0.000
	BMI [kg/m^2]	-6.402 ±11.224	-4.738 ±10.348	-9.528 ±11.022
	FPG [mmol/L]	-0.432 ±0.870	-0.509 ±0.867	-0.612 ±0.979
	HDL [mmol/L]	0.383 ±0.684	0.495 ±0.654	0.032 ±0.178
	LDL [mmol/L]	0.084 ±0.676	0.152 ±0.800	0.145 ±0.540
	TG [mmol/L]	0.408 ±1.921	0.363 ±1.634	-0.051 ±0.296
	Total Cholesterol [mmol/L]	0.026 ±0.192	0.109 ±0.815	-0.040 ±0.298
CEILS$_{random}$	Age	0.392 ±7.595	-0.018 ±0.014	-0.018 ±0.014
	BMI [kg/m^2]	-11.257 ±9.984	-11.626 ±9.811	-12.205 ±10.236
	FPG [mmol/L]	0.094 ±0.826	0.066 ±0.810	0.095 ±0.770
	HDL [mmol/L]	0.105 ±0.450	0.098 ±0.449	0.096 ±0.470
	LDL [mmol/L]	-0.130 ±0.936	-0.109 ±0.921	-0.143 ±0.917
	TG [mmol/L]	0.152 ±1.371	0.314 ±1.796	0.406 ±1.943
	Total Cholesterol [mmol/L]	0.020 ±1.036	0.056 ±1.033	0.031 ±1.032
DiCE$_{genetic}$	Age	2.566 ±10.784	0.000 ±0.000	0.000 ±0.000
	BMI [kg/m^2]	-3.914 ±8.598	-3.228 ±10.873	-7.508 ±10.322
	FPG [mmol/L]	-0.590 ±0.941	-0.601 ±0.853	-0.682 ±0.924
	HDL [mmol/L]	0.402 ±0.685	0.495 ±0.715	0.090 ±0.253
	LDL [mmol/L]	0.300 ±0.988	0.120 ±0.634	0.051 ±0.356
	TG [mmol/L]	0.424 ±1.765	0.547 ±1.834	-0.024 ±0.249
	Total Cholesterol [mmol/L]	-0.003 ±0.646	0.018 ±0.136	0.043 ±0.546
CEILS$_{genetic}$	Age	-0.018 ±0.014	-0.018 ±0.014	-0.018 ±0.014
	BMI [kg/m^2]	-11.395 ±9.846	-11.550 ±9.822	-11.460 ±10.088
	FPG [mmol/L]	-0.021 ±0.811	-0.002 ±0.735	-0.099 ±0.751
	HDL [mmol/L]	0.074 ±0.434	0.093 ±0.459	0.124 ±0.442
	LDL [mmol/L]	-0.115 ±0.931	-0.115 ±0.906	-0.132 ±0.929
	TG [mmol/L]	0.239 ±1.630	0.154 ±1.489	0.343 ±1.873
	Total Cholesterol [mmol/L]	0.043 ±1.057	0.022 ±1.000	0.055 ±1.058
DiCE$_{gradient}$	Age	1.735±9.008	0.000±0.000	0.000±0.000
	BMI [kg/m^2]	-4.965 ±11.070	-3.862 ±8.726	-6.193 ±9.937
	FPG [mmol/L]	-0.403 ±0.860	-0.517 ±0.852	-0.794 ±0.961
	HDL [mmol/L]	0.577 ±0.740	0.619 ±0.741	0.062 ±0.214
	LDL [mmol/L]	0.025 ±0.185	0.066 ±0.348	0.058 ±0.241
	TG [mmol/L]	0.305 ±1.590	0.130 ±0.733	-0.025 ±0.147
	Total Cholesterol [mmol/L]	0.152 ±0.811	0.184 ±0.844	0.061 ±0.344
CEILS$_{gradient}$	Age	-0.018 ±0.014	-0.018 ±0.014	-0.018 ±0.014
	BMI [kg/m^2]	-11.545 ±9.841	-11.033 ±10.912	-12.076 ±10.286
	FPG [mmol/L]	-0.096 ±0.855	0.090 ±0.825	0.049 ±0.800
	HDL [mmol/L]	0.095 ±0.453	0.110 ±0.478	0.084 ±0.444
	LDL [mmol/L]	-0.113 ±0.910	-0.116 ±0.909	-0.142 ±0.929
	TG [mmol/L]	0.323 ±1.885	0.012 ±1.104	0.206 ±1.527
	Total Cholesterol [mmol/L]	0.002 ±1.019	0.028 ±1.045	0.058 ±1.038

Table 6.4: Changes (average +-SD) in feature values between factual observation and suggested CE, as a function of different constraints during generation.

	Features	*NC*	*FC-TR*	*FC-CLIN*
DiCE$_{random}$	Age	12.5%(↓0, ∼14.3, ↑85.7)	0%	0%
	BMI [kg/m^2]	73.2%(↓46.3, ∼51.3, ↑2.4)	64.3%(↓30.6, ∼66.7, ↑2.8)	80.4%(↓64.4, ∼35.6, ↑0)
	FPG [mmol/L]	42.9%(↓54.2, ∼45.8, ↑0)	46.4%(↓65.4, ∼34.6, ↑0)	53.6%(↓67.7, ∼33.3, ↑0)
	HDL [mmol/L]	39.3%(↑68.2, ∼31.8, ↓0)	46.4%(↑84.6, ∼15.4, ↓0)	19.6%(↑18.2, ∼72.7, ↓9.1)
	LDL [mmol/L]	21.4%(↓8.3, ∼75, ↑16.7)	26.8%(↓6.7, ∼73.3, ↑20.0)	35.7%(↓10.0, ∼60.0, ↑30.0)
	TG [mmol/L]	26.8%(↓20.0, ∼53.3, ↑26.7)	26.8%(↓6.7, ∼73.3, ↑20.0)	23.2%(↓30.8, ∼61.5, ↑7.7)
	Total Cholesterol [mmol/L]	25.0%(↓0, ∼92.9, ↑7.1)	28.6%(↓6.2, ∼68.7, ↑25.0)	25.0%(↓7.1, ∼92.9, ↑0)
	Sex assigned at birth	0%	0%	0%
	Pressure [mmHg]	1.8%	3.6%	7.1%
	Smoke	1.8%	0%	0%
	Antidepressants	0%	0%	0%
	Corticosteroids	0%	3.6%	7.1%
	Antihypertensives	1.8%	1.8%	3.5%
	Cholesterol lowering medications	0%	0%	0%
	Quit-smoking medications	0%	0%	1.8%
	HTN	0%	5.4%	1.8%
	COPD	0%	1.8%	0%
	Depression	5.4%	1.8%	0%
	OA	1.8%	0%	3.6%
CEILS$_{random}$	Age	10.7%(↓33.3, ∼16.7, ↑50.0)	1.8%(↓0, ∼100, ↑0)	1.8%(↓0, ∼100, ↑0)
	BMI [kg/m^2]	100% (↓82.2, ∼8.9, ↑8.9)	100%(↓85.7, ∼5.4, ↑8.9)	100% (↓89.3, ∼5.4, ↑5.3)
	FPG [mmol/L]	100% (↓32.2, ∼33.9, ↑33.9)	100% (↓32.1, ∼37.5, ↑30.4)	100% (↓28.6, ∼37.5, ↑33.9)
	HDL [mmol/L]	100% (↑50.0, ∼8.9, ↓41.1)	100% (↑50.0, ∼10.7, ↓39.3)	100% (↑48.2, ∼8.9, ↓42.9)
	LDL [mmol/L]	100% (↓48.2, ∼5.4, ↑46.4)	100%(↓44.6, ∼8.9, ↑46.4)	100%(↓46.4, ∼7.2, ↑46.4)
	TG [mmol/L]	100% (↓62.5, ∼7.1, ↑30.4)	100% (↓62.5, ∼5.4, ↑32.1)	100% (↓62.5, ∼8.9, ↑28.6)
	Total Cholesterol [mmol/L]	100% (↓39.3, ∼17.9, ↑42.9)	100%(↓37.5, ∼19.6, ↑42.9)	100%(↓44.6, ∼16.1, ↑39.3)
	Sex assigned at birth	0%	0%	0%
	Pressure [mmHg]	75.0%	75.0%	75.0%
	Smoke	64.3%	62.5%	64.3%
	Antidepressants	3.6%	5.4%	8.9%
	Corticosteroids	30.4%	28.6%	30.4%
	Antihypertensives	51.8%	51.8%	51.8%
	Cholesterol lowering medications	60.7%	60.7%	66.1%
	Quit-smoking medications	25.0%	23.2 %	21.4%
	HTN	17.9%	19.6%	19.6%
	COPD	73.2%	60.7%	66.1%
	Depression	37.5%	37.5%	37.5%
	OA	5.4%	3.6 %	5.4%

Table 6.5: Percentage of changes in feature values between factual and suggested counterfactual, as a function of different constraints during generation (search method: *random search*). The symbols ↑ and ↓ represent upward and downward shifts with respect to the factual value, respectively. Green arrows indicate changes associated with an improvement in the feature value, whereas red arrows indicate a worsening. The symbol ∼ denotes counterfactual explanations whose values lie within a 5% relative variation of the factual value.

	Features	*NC*	*FC-TR*	*FC-CLIN*
DiCE$_{genetic}$	Age	14.29%(↓12.5, ∼0, ↑87.5)	0	0
	BMI [kg/m^2]	66.1%(↓29.7, ∼70.3, ↑0)	64.3%(↓33.3, ∼58.4, ↑8.3)	75.0% (↓52.4, ∼47.6, ↑0)
	FPG [mmol/L]	48.2%(↓70.4, ∼29.6, ↑0.0)	60.7%(↓61.8, ∼38.2, ↑0.0)	57.14%(↓81.3, ∼18.7, ↑0.0)
	HDL [mmol/L]	41.1%(↑78.3, ∼17.4, ↓4.3)	44.6%(↑76.0, ∼24.0, ↓0.0)	25.0% (↑50.0, ∼50.0, ↓0)
	LDL [mmol/L]	25.0%(↓0, ∼57.1, ↑42.9)	28.6%(↓12.5, ∼62.5, ↑25.0)	23.2%(↓0, ∼84.6, ↑15.4)
	TG [mmol/L]	30.4%(↓11.8, ∼58.8, ↑29.4)	26.8%(↓0.0, ∼66.7, ↑33.3)	26.8%(↓20.0, ∼66.7, ↑13.3)
	Total Cholesterol [mmol/L]	26.8%(↓6.7, ∼86.6, ↑6.7)	25.0%(↓0.0, ∼92.9, ↑7.1)	26.8%(↓13.4, ∼73.3, ↑13.3)
	Sex assigned at birth	5.4%	0%	0%
	Pressure [mmHg]	3.6%	5.4%	7.1%
	Smoke	1.8%	0%	0%
	Antidepressants	3.6%	3.6%	3.6%
	Corticosteroids	5.4%	1.8%	3.6%
	Antihypertensives	0%	5.4%	3.6%
	Cholesterol lowering medications	3.6%	1.8%	1.8%
	Quit-smoking medications	0%	7.1%	1.8%
	HTN	0%	5.4%	0%
	COPD	1.8%	1.8%	1.8%
	Depression	0%	1.8%	1.8%
	OA	1.8%	3.6%	3.6%
CEILS$_{genetic}$	Age	1.8%(↓0.0, ∼100, ↑0.0)	1.8%(↓0.0, ∼100, ↑0.0)	1.8%(↓0.0, ∼100, ↑0.0)
	BMI [kg/m^2]	100.0%(↓83.9, ∼7.2, ↑8.9)	100.0%(↓83.9, ∼8.9, ↑7.2)	100.0%(↓82.2, ∼8.9, ↑8.9)
	FPG [mmol/L]	100.0%(↓39.3, ∼30.4, ↑30.4)	100.0% (↓35.7, ∼37.5, ↑26.8)	100.0%(↓42.9, ∼30.3, ↑26.8)
	HDL [mmol/L]	100.0%(↑48.2, ∼8.9, ↓42.9)	100.0%(↑48.2, ∼8.9, ↓42.9)	100.0% (↑51.8, ∼8.9, ↓39.3)
	LDL [mmol/L]	100.0%(↓46.5, ∼8.9, ↑44.6)	100.0%(↓46.5, ∼8.9, ↑44.6)	100.0% (↓48.2, ∼3.6, ↑48.2)
	TG [mmol/L]	100.0%(↓62.5, ∼7.1, ↑30.7)	100.0%(↓62.5, ∼7.1, ↑30.7)	100.0%(↓60.8, ∼7.1, ↑32.1)
	Total Cholesterol [mmol/L]	100.0%(↓37.5, ∼17.9, ↑44.6)	100.0%(↓39.3, ∼17.8, ↑42.9)	100.0%(↓37.5, ∼19.6, ↑42.9)
	Sex assigned at birth	0%	0%	0%
	Pressure [mmHg]	75.0%	76.8%	75.0%
	Smoke	64.3%	64.3%	64.3%
	Antidepressants	3.6%	3.6%	5.4%
	Corticosteroids	25.0%	26.8%	30.4%
	Antihypertensives	51.8%	50.0%	51.8 %
	Cholesterol lowering medications	53.6%	53.6 %	55.4%
	Quit-smoking medications	21.4%	25.0%	25.0%
	HTN	17.9%	19.6%	17.9%
	COPD	67.9%	62.5%	62.5%
	Depression	37.5%	37.5%	35.7%
	OA	3.6%	1.8%	3.6%

Table 6.6: Percentage of changes in feature values between factual and suggested counterfactual, as a function of different constraints during generation (search method: *genetic search*). The symbols ↑ and ↓ represent upward and downward shifts with respect to the factual value, respectively. Green arrows indicate changes associated with an improvement in the feature value, whereas red arrows indicate a worsening. The symbol ∼ denotes counterfactual explanations whose values lie within a 5% relative variation of the factual value.

	Features	*NC*	*FC-TR*	*FC-CLIN*
DiCE$_{gradient}$	Age	10.7%(↓0, ∼16.7, ↑83.3)	0%	0%
	BMI [kg/m^2]	71.4%(↓32.5, ∼65, ↑2.5)	62.5%(↓31.4, ∼68.6, ↑0)	67.9%(↓50, ∼50, ↑0)
	FPG [mmol/L]	39.3%(↓45.5, ∼54.5, ↑0)	46.4%(↓69.2, ∼30.8, ↑0)	60.7%(↓79.4, ∼20.6, ↑0)
	HDL [mmol/L]	42.9%(↑91.7, ∼8.3, ↓0)	53.6%(↑83.3, ∼16.7, ↓0)	26.8%(40↑, 46.7∼, ↓13.3)
	LDL [mmol/L]	21.4%(↓0, ∼91.7, ↑8.3)	23.2%(↓0, ∼84.6, ↑15.4)	28.6%(↓0, ∼68.8, ↑31.2)
	TG [mmol/L]	23.2%(↓0, ∼84.6, ↑15.4)	23.2%(↓0, ∼84.6, ↑15.4)	21.4%(↓16.7, ∼83.3, ↑0)
	Total Cholesterol [mmol/L]	32.1%(↓11.1, ∼67.7, ↑22.2)	35.7%(↓10, ∼60, ↑30)	26.8%(↓0, ∼86.7, ↑13.3)
	Sex assigned at birth	1.8%	0%	0%
	Pressure [mmHg]	5.4%	1.8%	7.1%
	Smoke	1.8%	0%	0%
	Antidepressants	3.6%	1.8%	3.6%
	Corticosteroids	3.6%	5.4%	1.8%
	Antihypertensives	0%	5.4%	5.4%
	Cholesterol lowering medications	1.8%	0%	1.8%
	Quit-smoking medications	5.4%	1.8%	0%
	HTN	0%	1.8%	3.6%
	COPD	1.8%	1.8%	0%
	Depression	5.4%	1.8%	0%
	OA	1.8%	1.8%	8.9%
CEILS$_{gradient}$	Age	1.8%(↓0, ∼100, ↑0)	1.8%(↓0, ∼100, ↑0)	1.8%(↓0, ∼100, ↑0)
	BMI [kg/m^2]	100%(↓83.9, ∼8.9, ↑7.1)	100%(↓82.2, ∼8.9, ↑8.9)	100%(↓82.2, ∼8.9, ↑8.9)
	FPG [mmol/L]	100%(↓42.9, ∼30.4, ↑26.7)	100%(↓30.4, ∼7.1, ↑30.4)	100% (↓33.9, ∼33.9, ↑32.2)
	HDL [mmol/L]	100%(↑50.0, ∼8.9, ↓41.1)	100%(↑50.0, ∼8.9, ↓41.1)	100%(↑50.0, ∼8.9, ↓41.1)
	LDL [mmol/L]	100% (↓44.6, ∼9.0, ↑46.4)	100% (↓48.2, ∼7.2, ↑44.6)	100% (↓48.2, ∼7.2, ↑44.6)
	TG [mmol/L]	100% (↓62.5, ∼7.1, ↑30.4)	100% (↓62.5, ∼7.1, ↑30.4)	100%(↓60.7, ∼7.1, ↑32.1)
	Total Cholesterol [mmol/L]	100%(↓39.3, ∼16.1, ↑44.6)	100%(↓39.3, ∼16.1, ↑44.6)	100%(↓35.7, ∼17.9, ↑46.4)
	Sex assigned at birth	0%	0%	0%
	Pressure [mmHg]	75.0%	75.0%	75.0%
	Smoke	66.0%	66.0%	64.3%
	Antidepressants	5.4%	7.1%	3.6%
	Corticosteroids	30.4%	28.6%	26.8%
	Antihypertensives	51.8%	50.0%	51.8%
	Cholesterol lowering medications	55.4%	58.9%	58.9%
	Quit-smoking medications	23.2%	25.0%	25.0%
	HTN	19.6%	17.9%	17.9%
	COPD	51.8%	66.1%	71.4%
	Depression	37.5%	39.3%	35.7%
	OA	1.8%	3.6%	1.8%

Table 6.7: Percentage of changes in feature values between factual and suggested counterfactual, as a function of different constraints during generation (search method: *gradient-based optimization*). The symbols ↑ and ↓ represent upward and downward shifts with respect to the factual value, respectively. Green arrows indicate changes associated with an improvement in the feature value, whereas red arrows indicate a worsening. The symbol ∼ denotes counterfactual explanations whose values lie within a 5% relative variation of the factual value.

	Property	*NC*	*FC-TR*	*FC-CLIN*
DiCE$_{random}$	Availability	100.00%	100.00%	97.15%
	Discriminative power	96.00%	96.46%	96.00%
	Proximity (continuous)	2.17±1.48	2.13 ±1.48	0.88 ±0.47
	Proximity (categorical)	0.23 ±0.43	0.20 ±0.41	0.23 ±0.43
	Sparsity	1.25±0.51	1.32 ±0.51	1.32 ±0.54
	Causal feasibility	71.94% ±13.57%	71.79% ±13.2 %	72.18% ±14.29%
CEILS$_{random}$	Availability	99.05%	98.50%	98.73%
	Discriminative power	98.18%	97.55%	98.00%
	Proximity (continuous)	0.93±0.73	0.97±0.78	0.99±0.77
	Proximity (categorical)	0.28±0.50	0.28±0.48	0.28 ±0.48
	Sparsity	9.07 ±2.04	8.77±2.05	8.96 ±1.99
	Causal feasibility	91.22% ±5.46%	91.20% ±5.08%	91.35% ±5.02%
DiCE$_{genetic}$	Availability	100.00%	100.00%	96.99%
	Discriminative power	96.10%	96.09%	96.00%
	Proximity (continuous)	2.03±1.42	2.09±1.44	0.85±0.47
	Proximity (categorical)	0.22±0.43	0.24±0.45	0.22 ±0.43
	Sparsity	1.50 ±0.54	1.52 ±0.50	1.36 ±0.58
	Causal feasibility	72.14% ±13.74%	70.92% ±13.24%	72.88% ±14.22%
CEILS$_{genetic}$	Availability	98.73%	98.42%	98.73%
	Discriminative power	97.73%	97.64%	97.46%
	Proximity (continuous)	0.94 ±0.76	0.97±0.80	0.94±0.76
	Proximity (categorical)	0.28±0.49	0.28±0.48	0.26 ±0.48
	Sparsity	8.75 ±2.04	8.64±2.07	8.78 ±2.08
	Causal feasibility	91.78% ±4.71%	91.59% ±5.09%	91.25% ±5.33%
DiCE$_{gradient}$	Availability	100.00%	100.00%	97.78%
	Discriminative power	96.46%	96.64%	96.18%
	Proximity (continuous)	2.09±1.44	2.17 ±1.46	0.88±0.47
	Proximity (categorical)	0.21 ±0.41	0.20±0.41	0.25±0.46
	Sparsity	1.34 ±0.51	1.41 ±0.49	1.39 ±0.56
	Causal feasibility	71.84% ±13.29%	71.91% ±13.46%	72.27% ±14.22%
CEILS$_{gradient}$	Availability	98.89%	99.21%	99.21%
	Discriminative power	97.64%	98.18%	98.09%
	Proximity (continuous)	0.97±0.79	0.94±0.72	0.94±0.76
	Proximity (categorical)	0.27 ±0.48	0.28 ±0.48	0.31±0.51
	Sparsity	8.75±2.04	8.89 ±2.09	8.40 ±2.10
	Causal feasibility	91.30% ±4.91%	91.27% ±5.12%	91.50% ±4.64%

Table 6.8: Quality of counterfactual explanations generated with DiCE and CEILS methods under different domain-specific constraints with random, genetic and greadient-based search strategies.

6.6 Discussion and possible clinical translation

This chapter investigates the suitability of CEs for identifying key variables, among those routinely monitored in primary care, whose modification could allow a prediabetic patient to reduce the risk of progression to T2D and restore a normoglycemic state, expanding the work presented in [196]. To this end, CEs were generated under multiple configurations, comparing approaches that explicitly incorporated causal knowledge (i.e., CEILS+DiCE)

with purely data-driven methods (original DiCE formulation) using different search strategies, and further varying the degree of feature-level modifiability (feature mutability and the permitted range of variation) to reflect increasingly realistic clinical assumptions by applying progressively stricter constraints during generation.

The methodological framework presented in this Chapter comprises three components: (i) a risk prediction model that forecasts future glycemic states (NG, PD, or T2D) based on current features; (ii) a counterfactual explainer aimed at reverting predicted outcomes from T2D to NG; and (iii) an SCM that models causal relationships used to constrain the generated explanations, enhancing their feasibility.

Although the underlying risk prediction model relied on general-purpose biomarkers commonly available in EHRs rather than condition-specific indicators (e.g., HbA1c, diet), it still achieved moderate performance in predicting the future glycemic state (79% test accuracy). These results indicate that some prognostic signals for prediabetes to T2D progression can already be captured from routinely collected clinical data, achieving classification performance consistent with findings reported in the literature [198, 199]. Nevertheless, further refinement and the inclusion of additional features including specific biomarkers, family history, lifestyle-related factors (exercise, diet, alcohol consumption) and socio-economic status may be helpful to improve predictive accuracy and therefore the quality of the generated explanations.

Overall, the counterfactual explanations here extracted suggested changes that were coherent with improvement in biomarkers toward a better health status (e.g., decrease in BMI, increase in HDL), as shown in the trends reported in Tables 6.5-6.7. The seemingly counterintuitive shifts toward higher LDL (i.e., worse) values were consistent with a trend that is present in such kind of data, as noted also in [201]. In particular, the widespread use of LDL-lowering medications, such as statins, among individuals with dyslipidemia and those progressing towards T2D resulted in lower observed LDL levels in subjects with $FutureState = T2D$ (see Tables 6.2 and 6.3).

In general, CEILS produced CEs that were statistically different from those generated by DiCE in terms of magnitude of change with respect to the factuals. Compared to DiCE, CEILS tended to modify a larger number of features, resulting in better sparsity, as an effect of explicitly accounting for causal dependencies among variables (Table 6.8). In contrast, DiCE changed fewer features, primarily altering numerical variables by a larger amount, corresponding to worse (higher) continuous proximity. No significant differences were observed in the generated CEs when varying feature mutability and the permitted range of variation.

As expected, adding constraints during the generation process reduced the search space, resulting in lower availability. Also, the experimental results clearly show that the original

purely data-driven DiCE formulation recommends explanations that are less feasible with respect to the underlying causal structure (lower causal feasibility).

Although the results presented in this Chapter and in Chapter 3 are still preliminary, they illustrate how the analysis of CEs can leverage data to identify the changes necessary to shift from a high-risk condition to a lower-risk one. By highlighting minimum viable changes needed to improve the patient's health condition, CEs can in principle provide valuable insights, especially in cases where the condition is still reversible (e.g., prediabetes). Furthermore, this Chapter demonstrated how purely data-driven explainers can be enhanced by incorporating clinical-domain knowledge through causal and feature mutability constraints, without compromising quality metrics and while improving actionability and feasibility. Future effort should be placed on further model tuning and more extensive validation of each component of the proposed framework (i.e., risk prediction model, counterfactual explainer and underlying SCM), including evaluation of potential bias and generalizability on external independent cohorts representing diverse geographical populations (e.g., the English Longitudinal Study of Ageing [202]). Moreover, a more accurate estimation of structural equations for the SCM (improving Step(i) in Section 6.4) should be pursued, as all nodes are currently modeled using the same DNN architecture, one for numerical and one for categorical variables.

Once rigorously tested and validated, the framework could be integrated into routine clinical workflows, as illustrated in Figure 6.1 and described below:

1. *Features Collection.* During primary care visits or routine health assessments, patients' relevant biomarkers are measured and recorded in the EHR. If the clinician identifies a patient as presenting a prediabetic condition ($CurrentState = PD$), they can activate the decision support system;

2. *Risk Estimation.* The risk prediction model analyzes the patient's current features profile and estimates the future glycemic state (NG, PD, or T2D). This risk assessment provides an evidence-based foundation for preventive intervention. The proposed risk can be compared against other available risk indicators[2];

3. *Generation of Counterfactual Explanations.* If the subject is predicted to progress to the chronic disease condition ($FutureState = T2D$), the counterfactual explainer is used to provide personalized, minimum-viable changes in collected features associated with a healthier metabolic state ($FutureState = NG$). These targets represent actionable changes that could revert the patient's predicted trajectory;

[2]For example: prediabetes risk test provided by the American Diabetes Association and the Centers for Disease Control and Prevention: https://www.cdc.gov/prediabetes/risktest/index.html

4. *Clinical Interface for Decision Support.* The counterfactual explanations are presented to the clinician through a user-friendly interface. Clinicians can use these insights to develop a tailored prevention plan, which may include lifestyle intervention and/or pharmacologic strategies;

5. *Patient Monitoring.* By comparing the patient's evolving biomarker profile to the proposed counterfactual targets over time, the clinicians can assess whether the proposed prevention plan is successfully moving the patient toward a normoglycemic trajectory and adjust it if insufficient improvement is observed.

It is worth noting that the proposed clinical workflow follows a human-in-command paradigm, whereby the clinician remains the final decision-maker, retaining full oversight and control over prevention planning and patient management, in line with current regulatory recommendations [17, 15].

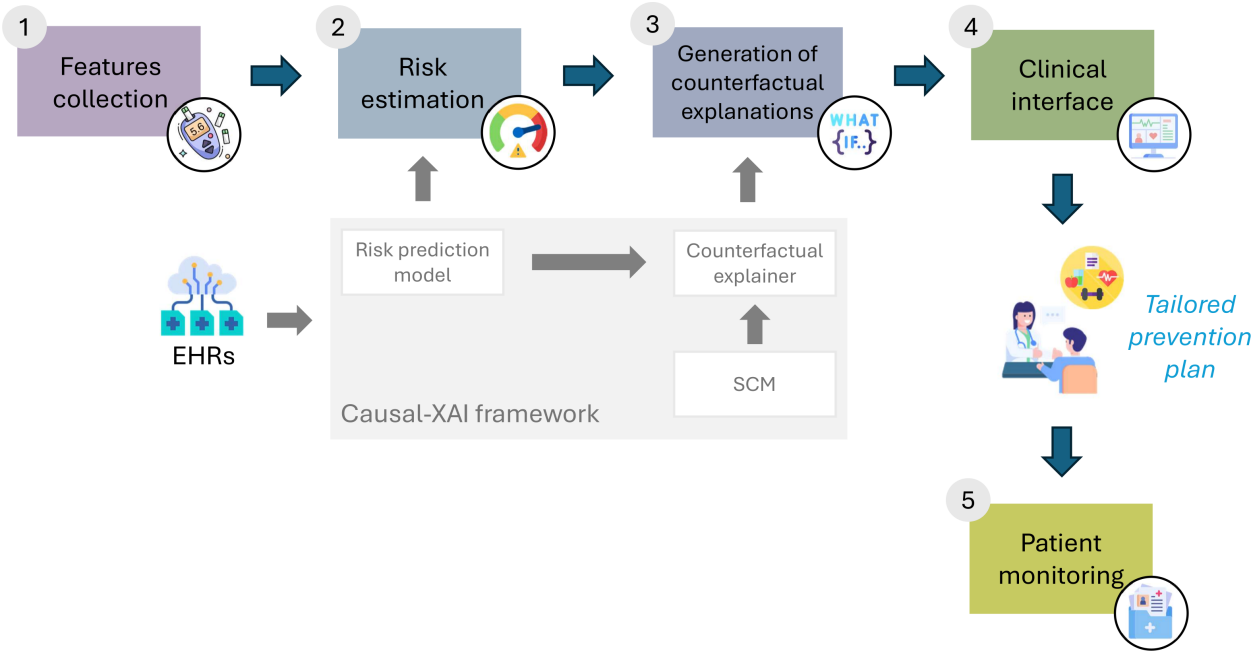


Figure 6.1: Example of clinical workflow embedding counterfactual explanations for chronic disease prevention. Icons made by Freepik, smashingstocks, Karyative and Vectorsclub from www.flaticon.com.

Chapter 7

Conclusions and future works

The present thesis work investigated advanced methodological frameworks for the development of data-driven predictive models targeting chronic diseases prevention.

A major aspect of the work concerned the interpretability of the models' decision making process. To approximate forms of reasoning that mimic human cognitive processes, particular emphasis was placed on counterfactual analysis, meaning the retrospective evaluation of hypothetical ("what-if") scenarios. In this work, counterfactual analysis, leveraging both purely associative XAI methods and causal learning techniques, was sistematically applied to chronic disease prevention case studies to demonstrate its practical utility in identifying how targeted modifications of modifiable risk factors could be leveraged to improve patient outcomes. The analysis of these case studies demonstrates that counterfactual analysis can substantially enhance model transparency while generating clinically meaningful insights capable of guiding preventive strategies and informing therapeutic decision pathways.

What follows is a summary of the main contributions provided throughout the Thesis, its limitations, and possible future developments.

Purely data-driven counterfactual explanations. Chapter 3 focuses on the XAI-oriented form of counterfactuals, namely counterfactual explanations (CEs), as defined by Wachter et al. [46]. The Chapter describes a novel methodological framework for the generation of CEs based on a multi-class classifier relying on hyperspherical classification regions (MC-SVDD) and a counterfactual explainer addressing an optimization problem on bounded classification regions (MUCH) [101, 102]. Experiments on three benchmark open source datasets (Appendix A) and one clinical dataset purposely extracted from primary care EHRs (Section 3.3) demonstrate that MC-SVDD is accurate in enclosing different classes of data points, with a negligible percentage of unclassified instances. Furthermore, the analysis outlines how classification regions derived from MC-SVDD can be effectively utilized to

extract CEs that hold potential for guiding personalized preventive actions, e.g., by reducing the individual 10-year-risk of CVDs in patients with COPD (*Case Study 1*, Section 3.3). Counterfactual explanations generated using the MUCH algorithm are evaluated from a computational perspective, using a range of quality measures and a newly introduced conformity measure.

Although the framework presented in Chapter 3 is inherently based on associative relationships and does not explicitly incorporate causal constraints, several data extraction and model design choices were introduced to improve actionability and reliability of the generated explanations. On one hand, constraints on feature mutability are imposed during the generation of CEs, for example, restricting modifications only to features that are clinically modifiable, enforcing predetermined directions of change (partial modifiability), or limiting modifications to clinically meaningful ranges. On the other hand, the combination of MC-SVDD with MUCH enhances reliability on multiple fronts. MC-SVDD naturally defines minimum-volume hyperspherical regions that exclude anomalous data points, while also enabling refinement of the decision boundary by controlling classification error (FPR control, Section 3.1.2). This yields regions in which the search for CEs is more representative of the target class. Furthermore, the introduction of the counterfactual conformity measure (Section 3.2.3) allows us to filter out CEs that fail to meet a required level of compliance with target properties. Together, these mechanisms help avoid potentially uncertain predictions and provide the end-user with an associated confidence level for each generated explanation, therefore increasing reliability. This, in turn, supports clinical decision-making while preserving the clinician's autonomy to accept or reject the proposed explanation.

The approach proposed in Chapter 3, supported by both global and local quality metrics, provides a valuable framework to assess alternative counterfactual explainers based on specific requirements. While further validation would strengthen the current findings, the analysis also opens several promising directions for future research. Future studies should aim to exploit the model-agnostic nature of the MUCH algorithm, and to investigate the effects of different sampling techniques and densities. Moreover, in multi-class settings where the output can be interpreted as an ordered sequence of states (e.g., varying levels of disease severity), it would be highly informative to investigate how a single factual instance transitions gradually from one label to the next one. By doing so, one could define a series of intermediate CEs that reflect incremental, clinically feasible changes rather than abrupt jumps between extreme states. This approach would not only enhance the interpretability and practical applicability of CEs but also provide insights into the progressive nature of the condition, offering clinicians a more nuanced understanding of potential interventions and their predicted impact over time.

Additionally, to ensure effective application of this approach in practice, it would be essential to develop methodological tools aimed at enhancing the clinical plausibility and feasibility of the proposed interventions. Preliminary findings shown in Chapter 3 indicate that conformal CEs may be more realistically applicable than non-conformal ones, suggesting that the counterfactual conformity measure here introduced is a step towards a more precise methodology for assessing the quality of counterfactual explanations. However, an optimal value of ε has not yet been defined. In the future, it will be necessary to establish criteria for selecting ε by balancing the trade-off between the number of discarded points and desired characteristics, defined by a combination of quality metrics and expert knowledge. Moreover, further research should include a deeper investigation of counterfactual conformity on a wider range of datasets and in relation to various measures of counterfactual quality. To fully achieve the goal of estimating viable recommendations for disease prevention, further research should focus on incorporating medical knowledge into the counterfactual generation process, for example by defining expert-driven dynamic bounds that indicate a plausible range of acceptable changes for each subject.

Causal learning and counterfactual inference. Chapters 4 and 5 describe how causal learning methods can be used to exploit counterfactual reasoning by following a chain of cause-and-effect encoded by a SCM [91]. These methods can be extremely useful in guiding the design of personalized prevention strategies [203] by using the known or inferred causal structure to predict how a patient's future outcome might change if different hypothetical interventions are performed. Lifestyle modification interventions for T2D prevention were considered as a case study (*Case Study 2*), leveraging two different sources of data: large sets of EHR data of primary care extracted from the CPCSSN database and 5-year simulations derived from a dynamical diabetes progression model that incorporates the long-term benefits of physical activity [52].

More specifically, Chapter 4 presented one of the first attempts to apply expert-driven causal discovery and inference methods for T2D prevention using EHR data, also accounting for latent confounders. As an illustrative example, counterfactual inference with latent confounders was used to simulate a lifestyle intervention exemplified by realistic reductions in FBS and BMI based on the average effects reported in the DPP trial [138]. The findings indicate a beneficial effect of the hypothetical intervention on predicted T2D incidence (minimum estimated reduction in T2D incidence of 15.1%), alongside heterogeneity in intervention effects across multivariate stratifications of participants. Possible uncertainties arising from the unidentifiability of the counterfactual query were addressed by providing a probability interval estimation. Nevertheless, the reliability of the estimated intervention

effects heavily depended on the quality and generalizability of the underlying dataset, the completeness of the recovered causal model and the methodological choices made to reduce computational complexity and ensure feasibility of the estimation procedure. Therefore, future clinical studies are needed to collect individual-level intervention data and validate the estimated effects. External validation of both the causal model and the estimated intervention efficacy on different datasets will be essential for assessing the generalizability of the results. The approximations introduced to reduce computational complexity of the estimation procedure (e.g., features discretization and the use of the relaxed EMCC algorithm) could be improved in the future by incorporating additional knowledge about known relationships between variables (e.g., the monotonicity of some structural equations).

While Chapter 4 focuses on static cause-effect relationships captured from observational data, Chapter 5 utilizes causal learning in dynamic settings, explicitly modeling temporal dependencies. Indeed, Chapter 5 presents an example of dynamic SCM to examine the impact of tailored physical activity recommendations (in terms of fixed frequency, duration and intensity of exercise) on the prevention of T2D onset, in individuals at elevated risk, for example those with prediabetes, on a 5-years basis [163]. The approach was first explicitly formalized in its analytical solution, adapted to the case of parametric, dynamic, SCMs, and then implemented in its numerical version (Algorithm 3), using a large number of simulations generated via a computationally efficient approximation of the T2D progression model obtained through ODE homogenization [175]. The analysis of inter-individual variability in glucose responses to WHO-compliant physical activities (Figures 5.4 and 5.5), highlighted the need to tailor exercise to individual physiological profiles (e.g., insulin sensitivity and β-cells functionality). Although applied here in the context of T2D progression, the proposed methodology can be adapted to any numerically solvable ODEs model with one or more observations of the over time, allowing to estimate treatment outcomes in different physiological systems. The proposed method leverages dynamical cause-and-effect relationships derived from prior physics-informed knowledge described in terms of ODEs, rather than purely associative relations [183], thus reducing issues typically related to observational data like selection bias and data sparsity [144]. Noticeably, the reliability of the estimated intervention effects heavily depends on the quality of the underlying ODE model and its parametrization. Therefore, future research should focus on improving the representativeness of diabetes progression models (e.g., modeling patient-specific characteristics like age, sex assigned at birth or diet) and their parameters estimation (e.g., through clinical trials and/or physics-informed AI methods, preliminary results available in [204]).

From a strictly methodological perspective, Chapter 5 analyzed treatment effect estimation in longitudinal setting within Pearl's SCM framework under the assumption of time-

invariant confounders and treatments (i.e., lifestyle interventions), where both factual and counterfactual physical activity plans were assumed constant over time. While this setting allows for tractable counterfactual analysis, it does not capture the complexities of real-world treatment regimes. Future works could extend the analysis to time-varying settings, for example using methods grounded in Rubin's POF framework [90]. One possibility could be the application of inverse probability of treatment weighting in Marginal Structural Models [205] that provides a principled approach for estimating causal effects in the presence of time-varying treatments (e.g., incremental physical activity plans) and time-varying confounders that are affected by prior treatment, enabling unbiased estimation of causal effects under the assumption of no unobserved confounders (sequential ignorability). The ignorability assumption can be relaxed in several ways, including the use of proxy variables within SCMs and the adoption of deconfounding frameworks. The former approach was employed in Chapter 4 for static settings. In particular, the presence of unobserved confounders in static settings has been addressed using the EMCC algorithm, introducing latent variables as root nodes in partially specified SCMs to enable counterfactual inference. EMCC was used to estimate the posterior distribution over latent variables consistent with the observed data and structural assumptions. This approach assumed the existence of proxy variables for latent confounders, thereby enabling partial identification and uncertainty quantification through interval estimates for counterfactual outcomes. In future research, deconfounding methods for longitudinal settings could be investigated to enable unbiased estimation of treatment effects over time in the presence of unobserved confounders. One promising direction is the use of deep learning–based approaches, such as Time Series Deconfounder [206], a factor model over time modeled as a recurrent neural network with multitask output, or the more recent Lipschitz-bounded neural stochastic controlled differential equations (LipSCDE) network [207] which is able to manage sparse, or irregular time series, being particularly suited for longitudinal observational data with irregular samples like EHRs.

Moreover, sensitivity analysis methods for unmeasured confounding [208, 209] could be incorporated to evaluate the robustness of causal conclusions. Such analysis would provide a systematic assessment of how relaxation of the ignorability assumption may impact estimated treatment effects and the validity of causal claims.

Structural causal explanations. The concept of counterfactual has been explored in both XAI and causality literature. Reflecting this dual perspective, this Thesis first examined counterfactual analysis within the XAI framework (CEs) in Chapter 3, and then addressed counterfactual reasoning from a causal inference perspective in Chapters 4 and 5. These two perspectives, though seemingly distinct, can be unified under what Carloni et

al. defined as *structural causal explanations* [12]. This research area focuses on generating CEs that are not only interpretable but also actionable and feasible in practice, providing algorithmic recourse which tries precisely to fill the gap between explanations and recommendations [210, 49]. Purely data-driven CEs are limited in this regard: they inform users about potential outcomes changing the model prediction but do not indicate how to achieve them in practice [49], often providing outcomes that do not respect natural laws and interactions between features. By shifting the focus from explanation to intervention, structural causal explanations exploit causal reasoning to explicitly encode knowledge of the causal relationships governing the environment in which actions are performed [12].

In this regard, Chapter 6 reports exploratory findings on the application of causal constraints derived from SCMs during the generation of CEs (e.g., using the CEILS method [72]) for chronic disease prevention purposes in the context of T2D prevention (*Case Study 2*). Although preliminary, these results are promising and support the potential of causal learning to enhance XAI, towards the provision of practical recommendations that respect the underlying mechanism of the phenomenon of interest, integrating data-driven knowledge and prior knowledge, while satisfying quality metrics. Research efforts should focus on embedding CEILS within the MUCH algorithm to combine the advantages of the flexible, reliability-enhancing tools presented in Chapter 3 with the model-agnostic properties of CEILS and its capacity to enforce causal constraints. More generally, a key direction for future work is to relax the causal sufficiency assumption when generating structural causal explanations, by developing causal counterfactual explainers that can operate with a partial or incomplete SCM, reflecting the fact that full causal structures are rarely known in real-world clinical settings. While preliminary efforts in this direction have been made by Mahajan et al. [69], experiments applying this approach to the dataset used in Chapter 6 revealed significant challenges, including variational autoencoder collapse during training and the explainer consistently returning identical explanations, highlighting the need for more robust tuning. Moreover, clinicians should be engaged throughout all stages of the design process, including SCM specification, predictive model development, and counterfactual evaluation, to ensure that the resulting explanations are both clinically meaningful and feasible. This continuous expert involvement, combined with the integration of AI systems into the clinical workflow (Figure 6.1), will be essential to support personalized, actionable strategies for improving patient health.

Concluding remarks. In summary, this thesis work describes how counterfactual analysis can be used for chronic disease prevention purposes, leveraging retrospective data to estimate the effects of interventions on modifiable risk factors (e.g., lifestyle) at the individ-

ual level. The findings here presented illustrate how such analysis can, in principle, support clinicians in gaining an initial understanding of which treatments are most likely to be beneficial for individual patients, and in identifying which variables must be modified and to what extent. Besides, the methods here presented can be used to estimate heterogeneity in intervention effects across patients subgroups. This approach can also be used to analyze intervention effects in vulnerable patients or patients with rare diseases that may be difficult to recruit in clinical studies, thereby facilitating the selection of a tailored intervention also for scarcely represented patients groups [31]. Collectively, these findings underscore the value of counterfactual reasoning as a promising tool for translating data-driven models into actionable clinical knowledge. Future developments of the proposed approaches could inform the formulation of cost-effective, personalized prevention programs [139] integrating medical knowledge with causal inference from observational data. To enhance the applicability of these approaches, future research should prioritize comprehensive internal and external validation on independent cohorts, combining quantitative, functionality-grounded evaluations with qualitative assessments that involve human participation, either lay persons (human-grounded) or domain experts (application-grounded) [211].

Appendices

Appendix A

Evaluation of MUCH on benchmark data

This Appendix reports a comparative analysis of the MUCH counterfactual explainer, applied to MC-SVDD classification regions, using three frequently referenced multi-class open source tabular datasets, namely, the FIFA dataset, the IRIS dataset, and the Stellar Classification Dataset - SDSS17.

A.1 Evaluation metrics

The classification performance was evaluated in terms of accuracy, macro-averaged F1-score (i.e., the mean of F1-scores computed by class), Cohen's Kappa Coefficient (i.e., the level of agreement between ground truth and predicted values [132], ranging from -1 to 1) and the percentage of unclassified points (%OUT).

The retrieved CEs were evaluated in terms of availability, proximity, plausibility, and discriminative power (see Section 2.1.2).

Availability was defined as the ratio of the number of CEs of class b generated from class a to the total number of factual observations of class a:

$$Availability_{a\to b} = \frac{\mid \mathbf{C}^b_{\mathbf{F}_a} \mid}{\mathbf{F}_a}\%$$

Proximity was computed as the normalized distance between each available CE of class b generated from class a and its corresponding factual observation, compared to a maximum reference distance, hence:

$$Proximity_{a\to b}(mean) = \frac{1}{\mid \mathbf{C}^b_{\mathbf{F}_a} \mid} \sum_{\boldsymbol{x}_{fa}\in\mathbf{F}_a:\ \exists \boldsymbol{x}^{cf_b}_{fa}\in\mathbf{C}^b_{\mathbf{F}_a}} \frac{dist(\boldsymbol{x}_{f_a}, \boldsymbol{x}^{cf_b}_{f_a})}{d_{\max}}\%$$

where $dist(\boldsymbol{x}_{f_a}, \boldsymbol{x}_{f_a}^{cf_b})$ is the euclidean distance between a factual observation and its CE, obtained after normalizing data in $[0, 1]$, and $d_{\max}$ is the maximum theoretical distance in the standardized modifiable-feature space, i.e., $d_{\max} = \sqrt{|\mathbf{u}|}$[1].

Plausibility was computed as the Hellinger distance [212] between CEs of class b generated from factuals of class a and the training set distribution of class b, chosen as the reference population:

$$Plausibility_{a \to b} = dist_{HEL}(\mathbf{C}^b_{\mathbf{F}_a}, \mathbf{X}_b)$$

A small Hellinger distance (closer to 0) suggests that the generated CEs lie near the real data distribution, while a large Hellinger distance (closer to 1) indicates that the generated instances are far from the real data distribution, i.e., less representative of the target class. Hence, the lower the better.

Lastly, discriminative power for a certain set of factual observations was estimated by evaluating the macro averaged test accuracy of a k-Nearest Neighbor (KNN) classifier ($k = 5$, 5-fold cross-validation) in discriminating factual observations in $\mathbf{F}_a$ from CEs in $\mathbf{C}_{\mathbf{F}_a}$.

A.2 FIFA dataset

A.2.1 Dataset description

The FIFA dataset, named after the famous football video game series, includes 89 attributes for over 17,000 players across various football leagues. From the original collection of attributes, a subset of 50 attributes related to the player's physical and athletic characteristics was selected. Besides age, height, and weight, the selected attributes can be summarized in three main categories: mental, physical and technical Skills. These attributes depict different aspects of the player's individual abilities and are represented in terms of rating, on a scale from 1 to 100. Moreover, the main attributes can be combined in 6 fundamental attributes, namely *Pace* (55% sprint speed, 45% acceleration), *Shooting* (ability to score: 45% finishing, 20% shot power, 20% long shots, 5% penalties, 5% positioning, 5% volleys), *Passing* (capability to successfully pass the ball to other teammates: 35% short passing, 20% vision, 20% crossing, 15% long passing, 5% curve, 5% free kick accuracy), *Dribbling* (50% dribbling, 35% ball control, 10% agility, 5% balance), *Defending* (ability to intercept the ball and mark the opponent: 30% marking, 30% sliding tackle, 20% interception, 10% heading accuracy, 10% sliding tackle) and *Physical* (50% strength, 25% stamina, 20% aggression, 5% jumping). These key attributes can be directly derived from the others, and for this reason,

[1]The largest theoretical distance in a u-dimensional $[0, 1]$ hypercube is that between coordinates $(0, \ldots, 0)$ and $(1, \ldots, 1)$.

only the 44 secondary attributes were considered as input features for classification.

A.2.2 Multi-class classification performance

The classification task consisted in predicting the correct player's position among 4 possible classes: *Midfielder* (MF), *Defender* (DE), *Forward* (FO), or *Goalkeeper* (GK). A balanced dataset of 8,000 records was obtained by applying random undersampling, extracting 2,000 samples per player position to match the least represented class. The dataset was splitted in training set (70%) and test set (30%). Both the OvR and OvO versions of the MC-SVDD were trained and evaluated; hyperparameter tuning was performed by applying 3-fold cross validation on the training set. OvR and OvO MC-SVDD were compared with the multi-class generalization of the Support Vector Machine (MC-SVM).

The OvR MC-SVDD was trained following the procedure explained in Section 3.1. Three distinct kernels, namely linear, cubic polynomial, and Gaussian, were evaluated. Cross-validation results revealed that the Gaussian kernel is the most suitable choice for the dataset, as shown in Table A.1. MC-SVDD with Gaussian kernel outperformed the other two methods on both the training and test sets, with test performance nearly matching training performance, indicating strong generalization capabilities. This improved performance came at the cost of a slightly higher, but still limited, percentage of unclassified points (i.e., 1.25% on the test set). This means that the classification regions identified by MC-SVDD can enclose almost all points and the presence of outliers in the data set is limited.

	Linear		Cubic		Gaussian	
	Tr	Ts	Tr	Ts	Tr	Ts
ACC	70%	68%	48%	47%	78%	77%
Macro F1-SCORE	66%	64%	38%	38%	73%	73 %
Cohen's Kappa	0.66	0.65	0.33	0.33	0.71	0.70
%OUT	0.06%	0.07%	0.05%	0.04%	0.59%	1.25%

Table A.1: Comparison of OvR MC-SVDD trained with linear, cubic, and gaussian kernels

Implementing OvO MC-SVDD involved training six separate binary MC-SVDD classifiers. Final classification was achieved through majority voting, meaning that the final predicted class was determined as the class most frequently predicted by the binary classifiers. The application of majority voting ensured that each point was assigned to a class with OvO MC-SVDD, unlike in the OvR approach where some points remained unclassified. The classification performance of OvO MC-SVDD is reported in Table A.2. As it can be seen from the table, the OvO approach performed worse on the FIFA dataset compared to the

	OvR MC-SVDD		OvO MC-SVDD		MC-SVM	
	Tr	Ts	Tr	Ts	Tr	Ts
ACC	78%	77%	68%	57%	99%	68%
Macro F1-SCORE	73%	73%	66%	50%	99%	68%
Cohen's Kappa	0.71	0.70	0.57	0.41	0.99	0.57

Table A.2: Classification performance of OvR, OvO MC-SVDD and MC-SVM: FIFA dataset

OvR approach. Indeed, test accuracy and macro-averaged F1-score were both lower than 60% when using the OvO MC-SVDD, and overfitting was observed due to the presence of a train-test accuracy gap higher than 10%.

Another important benchmark method for comparison is the MC-SVM. SVDD and SVM are strictly related [105], with the slight difference that SVDD is more prone to deal with outlier samples, as it encloses data points based on hyperspheres instead of hyperrectangles. Table A.2 shows how the optimized multi-SVM with Gaussian kernel applied on the FIFA dataset heavily overfits on the training set, failing to generalize on the test set.

Classes DE, FO, and GK can be accurately classified using OvR MC-SVDD. On the contrary, class MF is more difficult to discriminate. Indeed, the single class F1-score on the test set is more than acceptable when considering DE, FO and GK (i.e., 84.78%, 79.24%, and 100%, respectively), whereas it is noticeably lower when considering MF (27.96%). This is due to the fact that points in the MF class are easily confused with those in DE and FO classes as the characteristics of MF players are, in practice, intermediate between those of DE and FO players. It can also be observed that GK players are perfectly distinguishable from footballers in other game positions, because of the peculiar skills that this kind of player must demonstrate.

A.2.3 Evaluation of counterfactual explanations

The MUCH algorithm was used to determine the minimal attribute modifications leading to a change in playing position for each record in the test set. A sufficiently large set of candidate CEs was obtained by sampling 10000 points for each of the $m-1$ MC-SVDD regions using Halton sampling (see Section 3.2.2). *Age* and *height* were considered as non-modifiable features, and thus they were constrained during counterfactual search. In practice, CEs were accepted within a certain tolerance δ (i.e., $\delta = \pm 2cm$ for *height*) to guarantee availability. Clearly, as the value of δ decreases, the likelihood of the algorithm failing to produce a counterfactual increases (i.e., decreased availability) due to a reduction in the search space, particularly when the number of non-modifiable variables increases.

Even though the OvO MC-SVDD presented poor classification performance on the FIFA dataset with respect to OvR MC-SVDD (Table A.2), CEs properties (i.e., availability, proximity, discriminative power, and plausibility as described in Section A.1) were evaluated for both methods for completeness. Such comparison provides a clear example of how a different classification performance could impact the generation of CEs according to quality metrics, under otherwise identical conditions. Notably, availability mainly depended on the characteristics of counterfactual explainer (e.g., number of non-modifiable features, sampling granularity), being agnostic to the previous classification step. Indeed, both OvO and OvR MC-SVDD successfully returned all CEs (100% availability), demonstrating a sufficiently dense sampling of the MC-SVDD regions. The OvO approach yielded a slightly lower, but comparable, discriminative power with respect to OvR (i.e., on average, 97.48% for OvO and 98.15% for OvR MC-SVDD) and a slightly worse proximity (i.e., a minimum proximity of 23.10% and 21.38%, respectively).

The negligible difference in discriminative power and proximity metrics when comparing MUCH applied to OvO and OvR MC-SVDD underscores how CEs should also be evaluated from a feasibility perspective in addition to a geometric basis. On the one hand, geometric properties measure how well the generation algorithm can induce minimal alterations to feature values that result in the factual prediction being converted to a specified output, while disregarding the true feature distribution underlying the observed phenomenon. Other properties like plausibility, on the other hand, focus on describing how the CEs resemble the real class to which they should belong. This aspect is of paramount importance as CEs should be representative of the real-world population they are targeting. In this regard, Table A.3 reports the plausibility in terms of Hellinger distance of CEs obtained with the two approaches, calculated in relation to the class-wise distributions from the training dataset. As illustrated in Table A.3, the distributions of CEs obtained using OvR classification were consistently closer to real data than those obtained by OvO, in some cases even by an order of magnitude. Hence, the poor classification provided by the OvO approach was also reflected in the quality and plausibility of counterfactual explanations.

Indeed, it can be observed that the average distribution of the OvO-based CEs (panel A.1a) did not adequately represent the class of destination, as assessed by comparing the obtained distribution to the training-set distribution (reference population, shown in Figure A.1c). For example, the average distribution of CEs of class GK was unrealistic compared to the actual data distribution, suggesting similar shooting skill as an MF and a FO and similar dribbling ability as that of an MF or DE. This remarks that the generation of counterfactual explanations first relies on a proper classification, without which there is no chance of obtaining a reliable and meaningful data representation.

	MF	DE	FO	GK
MF_{OvR}	-	0.003	0.002	0.002
MF_{OvO}	-	0.028	0.054	0.148
DE_{OvR}	0.068	-	0.058	0.151
DE_{OvO}	0.172	-	0.146	0.221
FO_{OvR}	0.026	0.005	-	0.050
FO_{OvO}	0.090	0.043	-	0.137
GK_{OvR}	0.123	0.088	0.082	-
MF_{OvO}	0.183	0.113	0.130	-

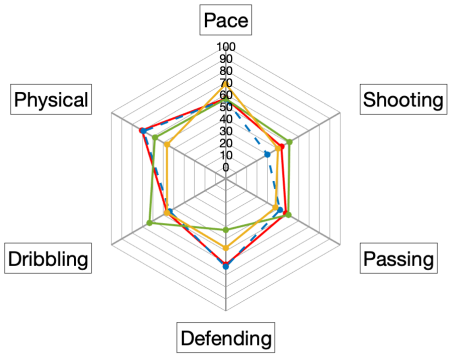
Table A.3: Plausibility wrt to the training set distribution using OvR and OvO MC-SVDD. Columns represent the factual class a, while rows represent the counterfactual class b

It is also worth noting that, in the case of OvO classification, the m classification regions ($S_1, \ldots, S_m$) to be used for the generation of CEs were not uniquely defined, since we had *m-1* different regions for each class, as each class was used in *m-1* classifiers. An empirical estimate of a distinct classification region for each class could be made by averaging the *m-1* profiles obtained. However, the regions obtained for different classes may overlap with each other, making the counterfactual search imprecise because the same data point can belong to more than one factual class. Instead, using the proposed OvR MC-SVDD we directly obtained m distinct and non-overlapping classification regions, obtained by training a smaller number of classifiers.

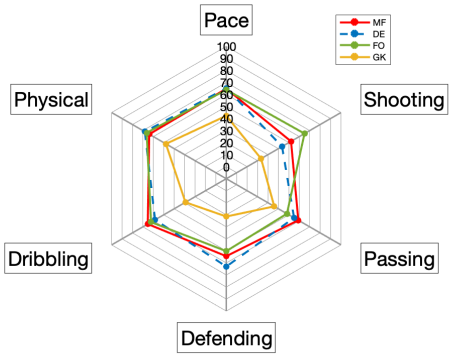
Table A.4 offers a more detailed analysis of the properties of CEs obtained with OvR MC-SVDD, for each factual class. Discriminative power was high, that is, above 95% for all factual classes. This indicates that CEs, although sought at a minimum distance, were easily distinguishable from points belonging to the factual class. Proximity values were also satisfactory, with average values between 21% (e.g., for $\boldsymbol{x}_{MF}^{DE}$) and 42% (e.g., for $\boldsymbol{x}_{GK}^{FO}$), depending on the factual class. Since the distribution of class GK differed prominently from the distributions of the other players, generating CEs for class GK typically requires larger feature modifications, resulting in higher (i.e., worst) average proximity. Lastly, the low plausibility values (i.e., $<< 1$) indicated that CEs were close to the real distribution of the class they aimed for.

Table A.5 lists three examples of factual observations belonging to class MF, FO, and GK and the corresponding counterfactual explanations belonging to class DE. These examples quantify the changes that specific players, trained for a specific role, would have to make in terms of fundamental characteristics to transition to the DE class.

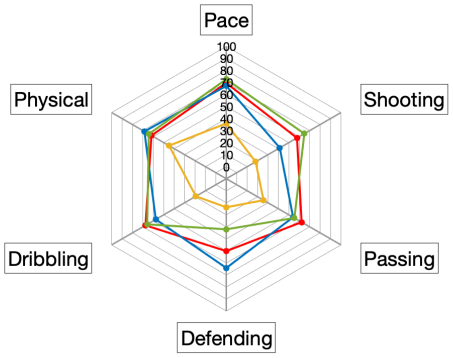
Each CE, being a local explanation, provides insights related to a specific input observation. However, considering the aggregate (global) pattern of changes required across all



(a) $\mathbf{C}_{\mathbf{F}_{DE}}$ (OvO MC-SVDD)



(b) $\mathbf{C}_{\mathbf{F}_{DE}}$ (OvR MC-SVDD)



(c) Training distribution

Figure A.1: Panel (a): Spiderplot representing the average fundamental skills of the factuals belonging to DE class (dashed line) and related counterfactuals (solid lines) obtained after OvO MC-SVDD classification. Panel (b): Spiderplot representing the average fundamental skills of the factuals belonging to DE class (dashed line) and related counterfactuals (solid lines) obtained after OvR MC-SVDD classification. Panel (C): average training data distributions grouped by output class (MF: red, DE: blue, FO: green, and GK: yellow). At a glance, it is evident that poor classification results in poor generation of CEs.

counterfactual explanations can be highly informative. As an example, Figure A.2 analyzes the average behavior of the set of factual observations belonging to class DE (i.e., $\mathbf{F}_{DE}$, dashed lines) with respect to their CEs ($\mathbf{C}^{MF}_{\mathbf{F}_{DE}}$, $\mathbf{C}^{FO}_{\mathbf{F}_{DE}}$, and $\mathbf{C}^{GK}_{\mathbf{F}_{DE}}$), showing a spiderplot for each attribute category (Mental, Physical, Technical and Fundamental skills). The retrieved counterfactual explanations, being intuitive and easy to interpret, can offer practical value for domain experts (e.g., athletic coaches) by informing the selection of key attributes and differentiate training plans according to target roles.

Factual Class	**MF**	**DE**	**FO**	**GK**
C1 Class	DE	MF	MF	MF
Availability%	100.00	100.00	100.00	100.00
Proximity%	21.73 (13.49, 29.96)	21.38 (12.74, 30.02)	21.39 (13.48, 29.31)	40.14 (35.80, 44.48)
C2 Class	FO	FO	DE	DE
Availability%	100.00	100.00	100.00	100.00
Proximity%	23.35 (15.80, 30.89)	24.05 (16.94, 31.17)	24.34 (16.65, 32.04)	38.21 (34.11, 42.31)
C3 Class	GK	GK	GK	FO
Availability%	100.00	100.00	100.00	100.00
Proximity%	40.13 (30.65, 49.61)	36.66 (27.71, 45.62)	37.60 (28.45, 46.75)	41.48 (36.95, 46.01)
Discriminative Power%	95.58	98.27	98.89	99.84

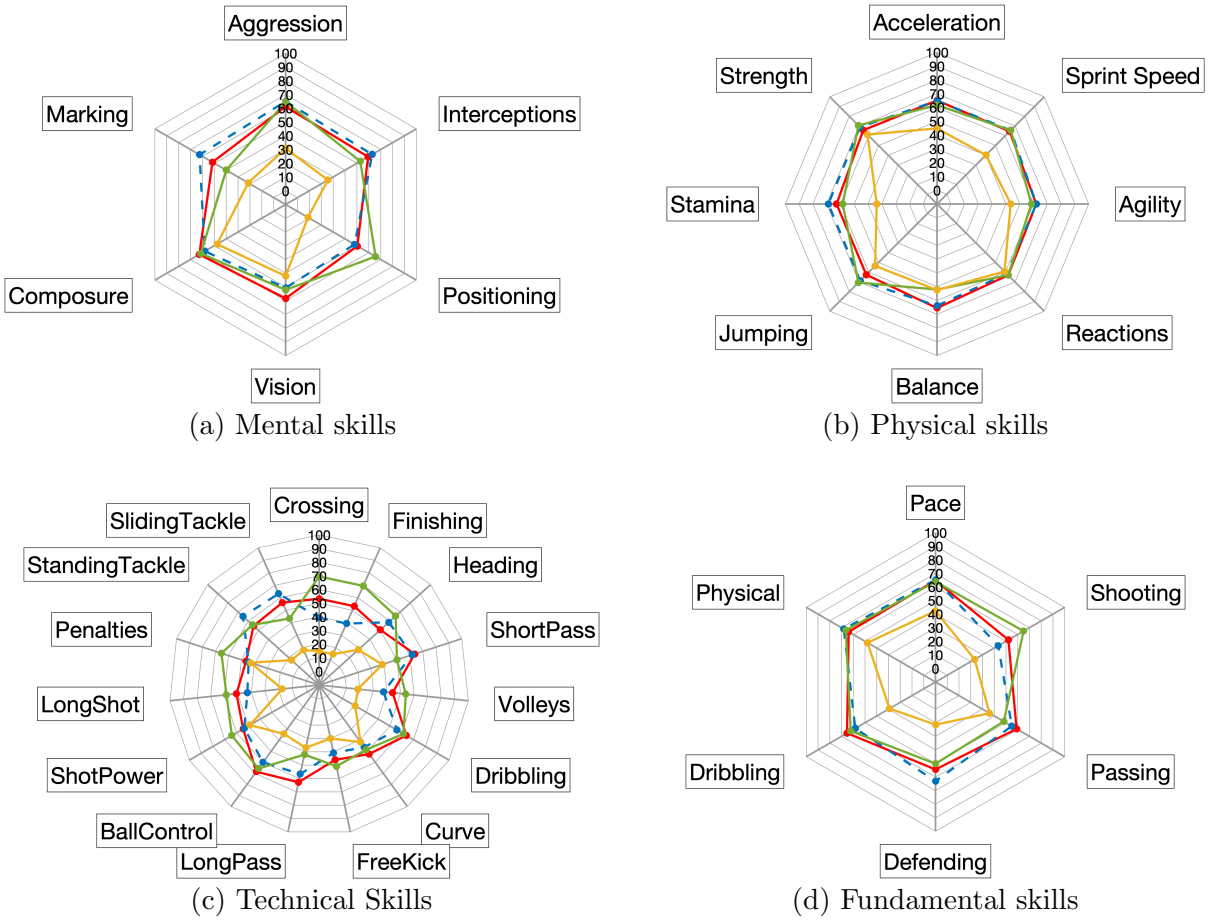
Table A.4: Availability (%), proximity (%, mean (95% CI)), discriminative power (%), and plausibility of counterfactuals generated from FIFA dataset, for different factuals classes.

	Example 1		**Example 2**		**Example 3**	
	$\boldsymbol{x}_{MF}$	$\boldsymbol{x}_{MF}^{DE}$	$\boldsymbol{x}_{FO}$	$\boldsymbol{x}_{FO}^{DE}$	$\boldsymbol{x}_{GK}$	$\boldsymbol{x}_{GK}^{DE}$
Pace	89.30	74.66	86.65	71.36	39.35	53.43
Shooting	61.3	51.32	72.60	52.37	16.20	35.79
Passing	52.15	50.05	62.45	52.75	19.80	39.51
Defending	50.40	62.44	44.90	66.44	10.60	64.99
Dribbling	65.75	53.46	84.25	63.26	12.20	47.68
Physical	67.20	60.13	63.85	61.18	43.20	62.73

Table A.5: Example of factuals ($\boldsymbol{x}_{MF}$, $\boldsymbol{x}_{FO}$, and $\boldsymbol{x}_{GK}$) and related counterfactual explanations ($\boldsymbol{x}_{MF}^{DE}$, $\boldsymbol{x}_{FO}^{DE}$, and $\boldsymbol{x}_{GK}^{DE}$).

A.3 Iris dataset

The Iris dataset consists of 150 observations related to four peculiar characteristics of three different iris species (i.e., *Setosa*, *Versicolor*, and *Virginica*). This dataset is a good benchmark because of its small size that simplifies visualization, experimentation and comparison of model performance. Data records are equally balanced in terms of classes and records of the Setosa species are linearly separable from the others, making it a simpler classification task with respect to the FIFA dataset. model performance. Table A.6 shows the training and test classification performance obtained by applying the OvR MC-SVDD model with Gaussian kernel to the Iris dataset, as compared to the MC-SVM. The dataset was split in training (70%) and test set (30%). Additionally, the MC-SVDD classification performance was compared with state-of-the-art multi-class classifiers including Decision Tree (criterion: "entropy"), Random Forest (criterion: "gini", bootstrap: true), and Gradient Boosting (cri-



(a) Mental skills

(b) Physical skills

(c) Technical Skills

(d) Fundamental skills

Figure A.2: Spiderplots representing the average distribution of the factual observations belonging to class DE ($\mathbf{F}_{DE}$, dashed line) and their CEs (solid line), for each attribute category (Mental, Physical, Technical and Fundamental Skills). The value scale ranges from 0 to 100, MF: red, DE: blue, FO: green, and GK: yellow.

terion: "gini") as shown in Table A.7. The parameters were optimized using Optuna [213] to achieve the best model performance. As summarized in Tables A.6 and A.7, all methods achieved high test accuracy ($> 98\%$) and F1 scores ($> 98\%$, except for MC-SVM), reflecting the simplicity and separability of the Iris dataset. More specifically, MC-SVDD yielded comparable, but slightly lower accuracy on the test set (i.e., 1-2% lower) than the four well-established methods. The dataset's separability is also evident from the fact that, when using MC-SVDD, all points were correctly assigned to a class with no outliers detected ($\%OUT = 0$).

Table A.8 shows the properties of the sets of CEs obtained with OvR MC-SVDD, for each factual class. The algorithm successfully returned all CEs (100% availability). Plausibility values were lower than 1, suggesting that the CEs were close to the real distribution of the class they aimed for. However, given the small sample size, the distribution of training

points may not accurately represent the true distribution of the class, making the computed plausibility measure less reliable. Since class Setosa in the Iris dataset is linearly separable from the other two classes, CEs belonging to classes Versicolor and Virginica were very easily distinguishable from factual points (discriminative power equal to 100%). Also in this case, achieving better separability and stronger discriminative power came at the expense of worse class proximity (i.e., higher distance).

	MC-SVDD		MC-SVM	
	Training	Test	Training	Test
ACC	95%	98%	99%	98%
Macro F1-SCORE	95%	98%	98%	75%
Cohen's Kappa	0.93	0.97	0.99	0.97

Table A.6: Classification performance of OvR MC-SVDD and MC-SVM: Iris dataset.

	Decision Tree	Random Forest	Gradient Boosting
ACC_{ts}	99.54%	99.32%	99.39%
Macro F1-$SCORE_{ts}$	99.00%	99.99%	99.29%

Table A.7: Classification performance of Decision Tree, Random Forest, and Gradient Boosting (test set): Iris dataset.

	IRIS		
Factual Class	***Setosa***	***Versicolor***	***Virginica***
C1 Class	*Versicolor*	*Setosa*	*Setosa*
Availability%	100.00	100.00	100.00
Proximity%	33.93 (27.80, 40.07)	28.77 (16.89, 40.66)	49.93 (38.72, 61.14)
Plausibility	0.32	0.29	0.17
C2 Class	*Virginica*	*Virginica*	*Versicolor*
Availability%	100.00	100.00	100.00
Proximity%	39.93 (33.92, 45.95)	11.83 (1.38, 22.29)	19.19 (9.13, 29.25)
Plausibility	0.32	0.19	0.38
Discriminative Power%	100.00	82.91	91.99

Table A.8: Availability (%), proximity (%, mean (95% CI)), discriminative power (%), and plausibility of CEs generated from the IRIS dataset, for different factuals classes.

A.4 Stellar Classification dataset

The Stellar Classification dataset includes 100,000 records of 3 type of objects (i.e., *Galaxy*, *Star* or *Quasar*) described by different spectral characteristics. The dataset includes real astronomical observations taken by the Sloan Digital Sky Survey. Every observation consists of 17 input features, however only a subset of 10 meaningful features was considered in this experiment. The dataset has a large number of records and represents an intermediate level of complexity, in terms of both the number of features and structure, compared to the two datasets discussed in Sections A.2 and A.3.

As for the Iris dataset, the MC-SVDD classification performance (OvR approach, Gaussian kernel) has been compared with state-of-the-art multi-class classifiers including MC-SVM (Table A.9), Decision Tree, Random Forest and Gradient Boosting (Table A.10). The dataset was split in training (70%) and test set (30%). Although performance on this dataset was slightly lower compared to those of the Iris one, all algorithms still achieved very high test accuracy ($> 92\%$). Macro F1-scores were also high for all methods ($> 93\%$), except for MC-SVM. MC-SVDD left a small proportion of samples unclassified, but this value remained limited (i.e., OUT% equal to 0.01% and 0.02% on the training and test set, respectively). Also in this case, the MC-SVDD yielded comparable, but slightly lower accuracy values on the test set than the four well-established classifiers. Table A.11 shows the main properties of the set of CEs obtained applying the MUCH method. The algorithm successfully returned all CEs (100% availability). In general, the generated CEs closely matched the training distribution of their target class (plausibility $<< 1$) and exhibited low distance from the factual class (proximity $<< 100\%$). However, CEs with target class Galaxy, represented an exception, with comparatively higher plausibility and distance values (please refer to C1 target class = Galaxy and C2 target class = Galaxy in Table A.11). Discriminative power appeared to be high, that is, above 95% for all classes.

	MC-SVDD		MC-SVM	
	Training	Test	Training	Test
ACC	94%	92%	96%	92%
Macro F1-SCORE	95%	94%	72%	69%
Cohen's Kappa	0.93	0.88	0.94	0.88

Table A.9: Classification performance of OvR MC-SVDD and MC-SVM: Stellar classification dataset.

	Decision Tree	Random Forest	Gradient Boosting
ACC_{ts}	94.83%	96.00%	94.00%
Macro F1-$SCORE_{ts}$	94.75%	95.92%	93.75 %

Table A.10: Classification performance of Decision Tree, Random Forest, and Gradient Boosting (test set): Stellar classification dataset.

	Stellar classification		
Factual Class	***Galaxy***	***Star***	***Quasar***
C1 Class	*Star*	*Galaxy*	*Galaxy*
Availability%	100.00	100.00	100.00
Proximity%)	39.14 (18.79, 59.49)	16.15 (3.72, 28.58)	14.91 (2.50, 27.33)
Plausibility	0.24	0.12	0.04
C2 Class	*Quasar*	*Quasar*	*Star*
Availability %	100.00	100.00	100.00
Proximity%	14.78 (3.25, 26.31)	17.40 (6.29, 28.51)	38.68 (19.61, 57.76)
Plausibility	0.03	0.08	0.21
Discriminative Power%	95.09	98.16	98.10

Table A.11: Availability (%), proximity (%, mean (95% CI)), discriminative power (%), and plausibility of CEs generated from the Stellar Classification dataset.

Appendix B

The Canadian Primary Care Sentinel Surveillance Network

The Canadian Primary Care Sentinel Surveillance Network (CPCSSN) [51] is a nation-wide database of de-identified primary care EHRs from across Canada. During my PhD I had access to the 2000-2015 version of CPCSSN, including a portion of more than 1.2 million patients, through an agreement with Toronto Metropolitan University. A waiver of review (REB 2013-261) was granted by the Review Ethics Board of Toronto Metropolitan University as this portion of CPCSSN database included de-identified and anonymized records. Besides, in this particular portion of the CPCSSN, there was no sensitive data related to patients' residence, socioeconomic status, or ethnicity.

The relational structure of the CPCSSN allows the extraction of heterogeneous information related to physiological and clinical data of a patient. The subset of CPCSSN tables relevant to the scope of this Thesis are depicted in Figure B.1. The *Patient* table is a registry of consenting patients whose primary provider is a physician in the CPCSSN network and contains demographic details like sex at birth and date of birth. Each patient is uniquely identified by the Patient_ID attribute. Physical examinations performed during primary care medical encounters are stored in the *Exam* table. Such table includes, for example, systolic and diastolic blood pressure, weight, height, and BMI measured during the encounter, with each record featuring the original manually inserted value, its automatic correction (if available), the unit of measurement, and the date of the examination. Similarly, the *Lab* table contains the results of laboratory tests, for example blood tests collecting HDL, LDL, total cholesterol, FPG, HbA1c, 2hPG, and triglycerides. Table *Medication* stores the names, and the identifier (i.e., ATC code) of the medications prescribed during a medical encounter, along with posology (e.g., frequency of assumption, dispensed units) and treatment duration.

Table *Riskfactor* contains useful information about lifestyle habits that may be related to a possible worsening of a pathological condition, like smoking or frequency of physical activity. This information is sparse and mainly coded in free text, requirying further processing for the sake of identifying the presence/absence of a given risk factor and its quantification (e.g., intensity of physical activity). Tables *HealthCondition*, *EncounterDiagnosis* and *Billing* keep trace of the reported health conditions with their date of onset, the outcome after each primary care encounter and the billing data for the patients, respectively. These tables are extremely useful to provide an overview of a patient's medical conditions and diagnoses. Additionally, a subset of high-prevalence diseases is coded in a structured way within the database and included in the auxiliary *DiseaseCase* table. Such table was populated by a detection algorithm that scanned the database using specific criteria to automatically assign one of eight coded diseases to a patient, as described in the 2014 CPCSSN Case Definition document [214]. Those criteria are based on billing information, reported problems, medications and laboratory results. The table contains diagnoses of diabetes, hypertension, COPD, depression, osteoarthritis, epilepsy, Parkinson's disease and dementia, accompanied by the corresponding Patient_ID and date of onset. Table B.1 reports the operational definition of CPCSSN diagnoses of diabetes and COPD, which were explicitly used in this Thesis.

As the database was populated by clinicians, often working in time-limited settings, the extraction of data from CPCSSN presented many challenges due to the presence of typing errors, missing fields, heterogeneity in medical terms and units of measurement. First, it was crucial to ensure the reported values aligned with a reasonable biological range. Entries which fell outside the acceptable ranges specified in Table B.2, defined as in [200], were removed. To ensure comprehensive data coverage, both the feature identifier and the corresponding value were sought in both coded (e.g., *Name_calc* and *Value_calc*) and non-coded free-text fields (e.g., *Name_orig* and *TestResult_orig*). While data in coded fields tended to be cleaner, data in non-coded fields required semi-automatic processing, e.g., using regular expressions. For example, BMI extraction encountered obstacles due to the presence of many missing values and numerical errors. To maximize data availability, BMI was often retrieved or corrected from weight and height, if present. For instance, when extracting the study dataset presented in Section 6.3 the following approach was followed:

1. If a BMI value in the acceptable range of Table B.2 was present within the time window of the glycemic state, that value (or the average in presence of multiple values) was extracted.

2. If no value was found, the BMI was computed using valid weight and height measures belonging to that same window in time using the formula $BMI = \frac{weight[kg]}{height[m]^2}$

3. Since weight and height typically do not fluctuate significantly over short time periods, if weight or height were not available within the time window, the earliest measure of that feature before *StartDate* was used to compute the BMI.

Moreover, being a Canadian database, additional complication arose from heterogeneity in language (English vs. French) and units of measurement (e.g., height recorded in feet/inches or centimeters, often with unspecified unit), which frequently lead to coding errors and inconsistencies. Also, the *RiskFactors* table could be a precious source of information for prevention purposes, due to the presence of information related to exercise, smoking, alcohol consumption, and diet. However, this information was reported as free text (when available), making structured data extraction difficult and requiring subjective assumptions.

Overall, the database represents a rich source of retrospective information and offers valuable coverage of the Canadian general population from 2000 to 2015. However, working with this data was not exempt from challenges. Much of the information was coded as free text, requiring extensive processing both to extract relevant features and to clean them, particularly for risk factors and medications. Also, many fields were optional, resulting in sparse or poorly coded data and sensitive variables that could have been highly informative for the development of prevention models, such as socioeconomic factors, were not provided due to ethical concerns. Follow-up information was often lost when patients were referred to specialistic care, and healthcare encounters typically occurred only when patients were symptomatic, leading to more extensive longitudinal records (e.g., presence of specific disease-related blood biomarkers) for individuals with certain conditions. For instance, the high proportion of missing values for HbA1c can be explained by its limited use in routine clinical screenings in the early 2000s, when the test was typically prescribed only if a patient was suspected of having diabetes or after the diagnosis.

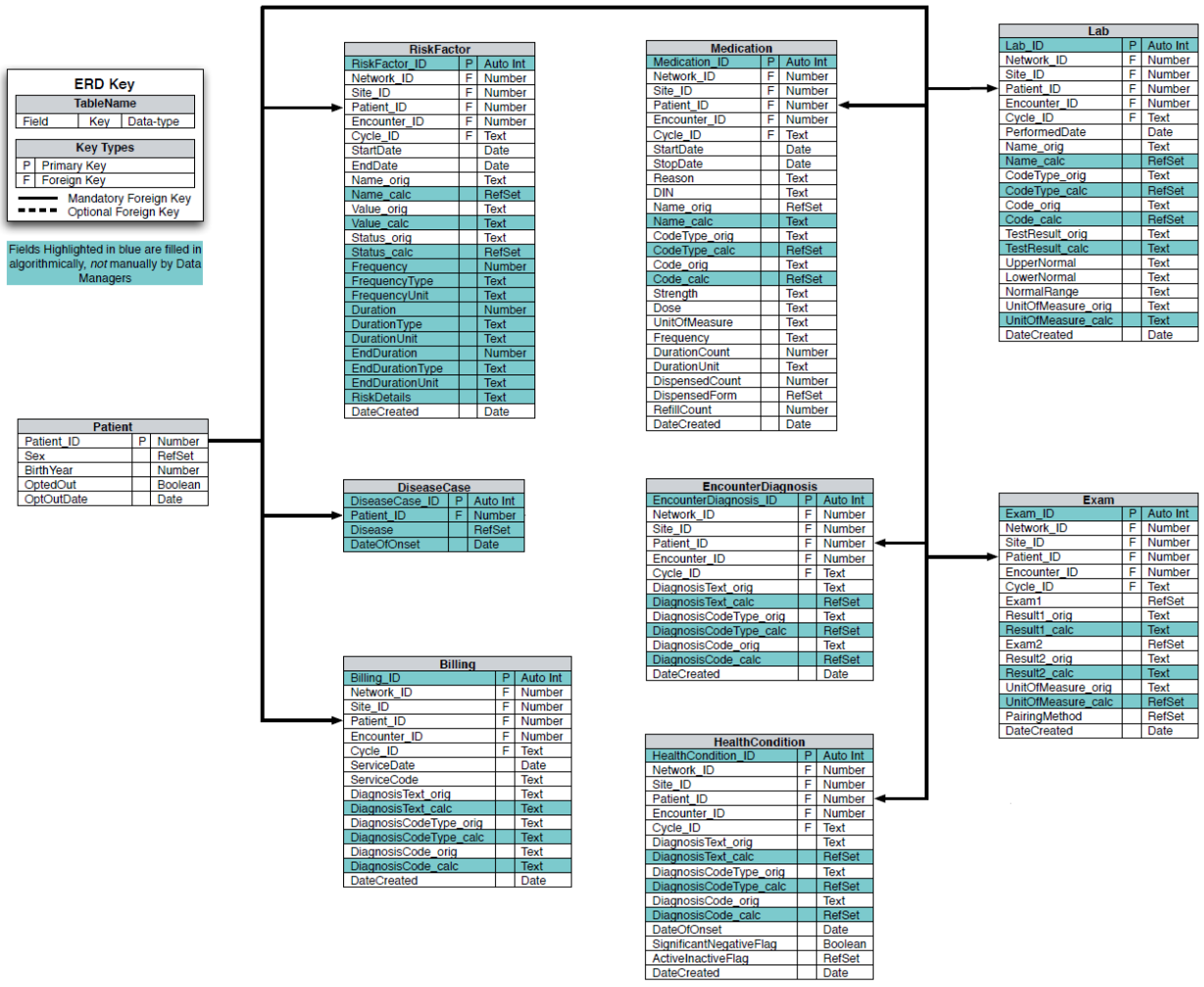


Figure B.1: Portion of the CPCSSN Entity Relationship Diagram considered for the purposes of this thesis work.

Disease	Billing	Problem List	Medication	Lab Results
Diabetes	Minimum two occurrences of the code '250, Diabetes mellitus' within two years	Any occurrence of the code '250, Diabetes mellitus'	ACARBOSE (A10BF01), GLIBENCLAMIDE (A10BB01), GLICLAZIDE (A10BB09), GLIMEPIRIDE (A10BB12), INSULIN (HUMAN) (A10AB01, A10AC01, A10AD01, A10AE01), INSULIN ASPART (A10AB05, A10AD05), INSULIN DETEMIR (A10AE05), INSULIN GLARGINE (A10AE04), INSULIN LISPRO (A10AB04, A10AD04), METFORMIN (A10BA02), METFORMIN AND ROSIGLITAZONE (A10BD03), SITAGLIPTIN (A10BH01), TOLBUTAMIDE (A10BB03), INSULIN (PORK) (A10AC03). The following diagnosis if exist in patient's problem list make the medication criteria alone insufficient: 256.4, Polycystic Ovarian Syndrome; 648.8, Gestational Diabetes; 249, Secondary (chemical induced) Diabetes; 790.29, Hyperglycemia NOS; 775.1, Neonatal diabetes mellitus	1. Any HbA1C >= 7 or two occurrences of FPG >7 within one year
COPD	Any occurrence of '491.2, Chronic bronchitis' or '492, Emphysema' or '496, Chronic airway obstruction not elsewhere classified'	Any occurrence of '491.2, Chronic bronchitis' or '492, Emphysema' or '496, Chronic airway obstruction not elsewhere classified'	TIOTROPIUM BROMIDE (R03BB04), IPRATROPIUM BROMIDE (R01AX03, R03BB02), SALBUTAMOL AND and SODIUM CROMOGLICATE (R03AK04) A diagnosis '493, Asthma' in patient's problem list make the medication criteria alone insufficient.	n.a.

Table B.1: Operational definition of diabetes and COPD [214].

Feature	Minimum value	Maximum value
FBS [mmol/L]	1.3	23
HDL [mmol/L]	0.6	3
LDL [mmol/L]	0.7	8
Total cholesterol [mmol/L]	2	13
Triglycerides [mmol/L]	0.1	20
sBP [mmHg]	50	266
dBP [mmHg]	20	192
Weight [Kg]	30	350
Height [cm]	80	210
Height [in]	31	79
Height [ft]	2.62	6.56
BMI [Kg/m^2]	10	60

Table B.2: Acceptable ranges used for features extraction.

Appendix C

Summary of additional contributions

This supplementary Chapter outlines additional activities carried out during the PhD that fall outside my primary research line but that led to journal or conference publications and student theses I supervised.

C.1 AI methods for hearing screening

During my PhD journey, I had the opportunity to continue the work started in my master's thesis [215, 216], further contributing to the design and assessment of a hearing-screening platform that integrates a battery of tests with AI models to support prevention of hearing loss in adults.

The WHISPER platform [217] includes: a survey with 18 multiple-choice questions about modifiable and non-modifiable risk factors for hearing loss (e.g., lifestyle, noise exposure, medical conditions); a self-administered speech-in-noise test [215] that is based on a one-up/three-down adaptive procedure using meaningless vowel-consonant-vowel stimuli (e.g., aba, ada) presented in stationary speech-shaped noise; and a battery of cognitive tests, including the digit span test (forward, backward, and ordering versions) in which the ability to recall sequences of digits of increasing length is tested as a proxy for working memory. In addition to the commonly measured Speech Recognition Threshold in noise (SRT), the platform extracts additional features, for example reaction times, percentage of correct responses, pattern of responses and self-adjusted test volume. By gathering over 40 distinct features, the WHISPER platform's test battery effectively characterizes several dimensions of individual performance, going beyond standard measures of hearing sensitivity (via pure-tone audiometry) and speech recognition (via speech-in-noise testing). As of November 2025, we examined 434 subjects aged 19-89 years (590 ears tested) across 12 native lan-

guages, tested in uncontrolled environmental noise settings in the lab and at local health screening initiatives (i.e., at universities of senior citizens, health prevention and awareness events for the general public). The study was approved by the Politecnico di Milano Research Ethical Committee (Opinion No. 13/2022, April 13, 2022). The analyses performed focused on three main research directions: (i) applying supervised learning techniques to evaluate how the extracted features contribute to predicting hearing loss, as defined by the gold-standard pure-tone audiometry test; (ii) employing unsupervised learning techniques to identify auditory profiles capable of revealing subgroups of individuals with similar hearing, cognitive, demographic, and risk-factor characteristics, thereby supporting data-driven clinical decision-making (e.g., personalized hearing-aid fitting and auditory rehabilitation); (iii) conducting a preliminary comparison of the WHISPER speech-in-noise test against other well-established tests, including the Digit Triplet Test and the Matrix Test.

C.2 XAI for the evaluation of synthetic health data

This section discusses the application of XAI models, alongside state-of-the-art metrics, to assess the quality of synthetic health data, a research activity initiated during my previous fellowship and completed in the early stage of the PhD. The methodology was assessed on a subset of 156 records of the WHISPER dataset and published in a Journal article [218].

Balanced synthetic datasets with 1000 records were generated starting from a subset of 156 records of the WHISPER dataset, using a Conditional Generative Adversarial Networks (GAN) [219, 220]. Different datasets were obtained by varying GAN parameters, namely the number of nodes per layer in the generator and discriminator networks, the batch size, and the number of epochs. To monitor the quality of the GAN generation process [221], we used a combination of the following measures: Maximum Mean Discrepancy (MMD, [222]), Classifier Two Sample (C2S) metric [223, 224], Hellinger Distance (HD, [225]) and Pairwise Correlation Difference (PCD, [226]) .

Besides standard measures, the quality of synthetic data was assessed using two complementary approaches: (i) comparing the classification performance of models trained on synthetic datasets in predicting slight/mild hearing impairment, and (ii) examining the decision rules generated by a native XAI model (the logic learning machine [227, 228]) to evaluate how closely the rules derived from synthetic data matched those produced using real data.

The classification performance was addressed by computing sensitivity, specificity, and F1-score in models deployed with the following combinations of training (Tr) and test (Te) sets:

- Condition A (baseline): TrR = training set from real dataset (80%), TeR = test set

from real dataset (20%)

- Condition B: TrS = training set from synthetic dataset (80%), TeS = test set from synthetic dataset (20%);
- Condition C: TrS = training set from synthetic dataset (80%), TeR = test set is the whole real dataset ;
- Condition D: TrR = training set from real dataset (80%), TeS = test set is the whole synthetic dataset.

A cross-classification (CC) measure [226] was introduced to summarize the similarity between real and synthetic datasets in terms of classification performance.

Decision rules were compared using a novel similarity measure[1] based on the cosine similarity between Bag of Words (BOW, [229]) representations of the set of rules. BOW is a widely used text representation approach (e.g., [230, 231]) where a text is decomposed into a matrix of words and their relative frequencies. Once the BOW matrix was created for both rulesets to be compared, cosine similarity was applied to all the combination of couples of rules, divided by class, to obtain a measure of similarity between rules. A global similarity metric between rulesets G_x was defined as the ratio of the number of real-synthetic rule pairs n_x with similarity greater than a pre-determined threshold value x (i.e, 0.6 in this study) to the total number of rules extracted from the real dataset.

This study demonstrated that XAI has the potential to provide additional insights in evaluating the quality of synthetic data, beyond the use of conventional utility metrics, in a hearing screening dataset. Specifically, a global similarity metric was introduced to assess the quality of synthetic data based on the similarity between the classification rule sets extracted from real and synthetic datasets. The proposed metric provided a comprehensive measure of similarity between decision rules by accounting for rule structure, cut-off values, and covering. Such measure helped in selecting the most suitable synthetic dataset(s) from a group of high-quality candidates that were otherwise considered equally similar based on standard utility metrics.

C.3 Consensus-clustering for Amyotrophic Lateral Sclerosis Phenotyping

This section briefly describes the use of unsupervised machine learning, specifically consensus clustering, to examine a retrospective Italian cohort of patients affected by Amyotrophic

[1]Code available at: https://github.com/lenattimarta/BOW_rule_similarity

Lateral Sclerosis (ALS), a neurodegenerative disease characterized by heterogeneous clinical, pathological, and genetic background, making disease phenotype classification challenging [232]. The objective was to determine whether clustering of patients using clinical, cognitive/behavioral, and neurophysiological measures could enhance the identification of disease heterogeneity at the time of diagnosis, thereby building the basis for designing personalized treatment approaches. This research activity [233, 234] was carried out within the RAISE[2] project.

The dataset included 184 patients with a validated ALS diagnosis according to the revised El Escorial Criteria[235], recruited at IRCCS Ospedale Policlinico San Martino (Comitato Etico Regionale della Liguria, LongALS, ID 11992), Genoa, Italy. Participants underwent a battery of neurophysiological, clinical, and extra-motor symptoms evaluations performed by experienced neurologists and neuropsychologists. Disease severity was evaluated using the ALS Functional Rating Scale-revised (ALSFRS-r) [236], muscular weakness was assessed using the Medical Research Council (MRC) scale [237], and the severity of UMN involvement was graded using the UMN score [238]. Demographics, family history of motoneuron diseases, onset site, type, side, limbs and muscular involvement, as well as disease duration, were recorded. Patients were grouped into four phenotypes: spinal onset ALS (S-ALS), bulbar onset ALS (B-ALS), pure/predominant upper motor neuron (pUMN), and pure/predominant lower motor neuron (pLMN), the latter aggregating flail arm, flail leg, and progressive muscular atrophy, while pyramidal and primary lateral sclerosis formed pUMN [232].

Two feature reduction strategies were evaluated: (i) Principal Component Analysis (PCA) and (ii) combination of PCA for numerical features and Multiple Discriminant Analysis (MDA) for categorical variables. After feature reduction, patient stratification was performed using consensus clustering [239]. In each of 100 runs, 80% of the data was randomly sampled and grouped into four clusters using K-means, hierarchical clustering, and their ensemble. Co-clustering frequencies across runs were aggregated into a consensus matrix, representing how consistently each pair of samples clustered together. Cluster labels were obtained by applying hierarchical clustering to this consensus matrix. The combination of PCA feature reduction and ensemble consensus clustering performed better than the other combinations in terms of external clustering metrics (Adjusted Rand Index, Fowlkes-Mallows index, homogeneity, and completeness [240]). Phenotypes distributions across the identified clusters were as follows: C1 (44 patients): 73% B-ALS, 9% S-ALS, 9% pUMN, and 9% pLMN; C2 (65 patients): 60% S-ALS, 37% pUMN, and 3% pLMN; C3 (9 patients): 45% S-ALS, 33% pLMN, and 22% pUMN; C4 (66 patients): 69.5% pLMN, 29% S-ALS and 1.5%

[2]Robotics and AI for Socio-economic Empowerment (RAISE), Mission 4, Component 2, Investment 1.5, Grant ECS00000035, National Recovery and Resilience Plan (NRRP)

pUMN. Key features of each retrieved clusters were determined based on the standardized mean difference (SMD) [241] with respect to the whole study population (absolute SMD $<$ 0.2: negligible effect; 0.2 $\leq$ absolute SMD $<$ 0.5: minor effect; absolute SMD $\geq$ 0.5: moderate effect). The four clusters showed distinctive patterns, indicated by SMDs between cluster samples and the entire population across the clinical variables considered. Onset variables appeared to have a moderate positive effect (SMD $\geq$ 0.5) within C1, compatibly with the prevalence of B-ALS patients in the cluster (key features: no side, bulbar onset site, no limbs and absence of muscular involvement), while the presence of ALS familiarity and definite familiar ALS El Escorial were key (interrelated) features for C3. All features in C2 and C4 showed a negligible or minor effect. Hence, consensus clustering revealed subgroups of patients with markedly different baseline cognitive/behavioral and familiarity characteristics indicating that a deeper analysis of patient subgroups beyond traditional phenotyping may help drive advancements in personalized treatment.

The average within-cluster consensus was above 0.7 for each cluster (0.87 for C1, 0.75 for C2, and 0.72 for C3 and C4), indicating consistent cluster stability across iterations, even given the relatively small sample size. Analysis of the within-phenotype consensus (the average pairwise consensus score among all samples assigned to a phenotype label) revealed a strong and well-defined profiling for B-ALS patients (0.92). In contrast, progressively weaker profiling was observed for pLMN (0.62), pUMN (0.59), and S-ALS (0.42), consistent with the onset characteristics of these phenotypes [232], reflecting substantial heterogeneity among S-ALS patients at the beginning of the disease, thus requiring longitudinal analysis to evaluate disease progression.

C.4 Dual-View Single-Shot Multibox Detector for a driver alert system

This section briefly describes research activities that were carried out within the Genova5G project, funded by the Italian Ministry of Economic Development (MiSE), with the objective to explore the potentials of 5G technology in a smart mobility context, within the metropolitan area of Genoa, Italy. The project's use case was an AI-based driver alert system for public transportation vehicles, based on video content analysis (VCA) and Message Queuing Telemetry Transport (MQTT) communication over 5G network. The base VCA system relied on a Single Shot Detector (SSD) [242] model trained on images from a single camera (annotated with YOLOv5x [243]) and three open-source datasets (Open Images Dataset [244], ETH Pedestrian Dataset [245] and EuroCity Dataset [246]) to identify,

localize, and eventually signal the presence of obstacles to public transportation vehicles approaching the surveilled area. The network was trained to recognize three classes of objects: 'vehicle', 'rider', and 'pedestrian'. The base SSD model was subsequently retrained on 10,000 images detected from a second camera, capturing a different field of view, using transfer learning. Such network was then evaluated on each camera separately and on their fusion, performed at the decision level. Evaluation was performed on 1000 frames for both cameras, opportunely aligned, in terms of i) object detection performance, that is, the ability of the system to correctly identify different classes of objects inside the region of interest (ROI), and ii) alert generation performance, that is the ability of the system to trigger an alert if and only if at least one object is present in the ROI.

By leveraging spatial redundancy (making decisions based on multiple camera views of the same area) and temporal continuity (analyzing two or more consecutive video frames) the alert system improved overall performance. This approach led to more correctly detected alarms (true positives), greater robustness in situations where a camera malfunctioned, and fewer missed alerts (false negatives). More details can be found in [247].

Appendix D

List of Publications

Peer reviewed Journal Articles

J1 L. Multerer, P. F. De Paola, M. **Lenatti**, A. Paglialonga, and L. Azzimonti, "A reduced model for the long-term effects of physical activity on type 2 diabetes," *Mathematical Biosciences* , vol. 394, p. 109645, 2026

J2 M. **Lenatti**, A. Carlevaro, A. Guergachi, K. Keshavjee, M. Mongelli, and A. Paglialonga, "Estimation and Conformity Evaluation of Multi-Class Counterfactual Explanations for Chronic Disease Prevention," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 9, pp. 6132–6142, 2025

J3 A. Carlevaro*, M. **Lenatti***, A. Paglialonga, and M. Mongelli, "Multiclass Counterfactual Explanations Using Support Vector Data Description," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 6, pp. 3046–3056, 2024

J4 M. **Lenatti**, A. Paglialonga, V. Orani, M. Ferretti, and M. Mongelli, "Characterization of Synthetic Health Data Using Rule-Based Artificial Intelligence Models," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 8, pp. 3760–3769, 2023

J5 M. **Lenatti**, S. Narteni, A. Paglialonga, V. Rampa, and M. Mongelli, "Dual-View Single-Shot Multibox Detector at Urban Intersections: Settings and Performance Evaluation," *Sensors*, vol. 23, no. 6, 2023

J6 F. Guida, M. **Lenatti**, K. Keshavjee, A. Khatami, A. Guergachi, and A. Paglialonga, "Characterization of Inclination Analysis for Predicting Onset of Heart Failure from Primary Care Electronic Medical Records," *Sensors*, vol. 23, no. 9, 2023

* co-first author

J7 M. **Lenatti**, A. Carlevaro, A. Guergachi, K. Keshavjee, M. Mongelli, and A. Paglialonga, "A novel method to derive personalized minimum viable recommendations for type 2 diabetes prevention based on counterfactual explanations," *PLOS ONE*, vol. 17, p. e0272825, Nov. 2022

J8 A. Carlevaro, M. **Lenatti**, A. Paglialonga, and M. Mongelli, "Counterfactual Building and Evaluation via eXplainable Support Vector Data Description," *IEEE Access*, vol. 10, pp. 60849–60861, 2022

J9 M. **Lenatti**, P. A. Moreno-Sánchez, E. M. Polo, M. Mollura, R. Barbieri, and A. Paglialonga, "Evaluation of Machine Learning Algorithms and Explainability Techniques to Detect Hearing Loss From a Speech-in-Noise Screening Test," *American Journal of Audiology*, vol. 31, p. 961–979, Sept 2022

J10 M. Zanet, E. M. Polo, M. **Lenatti**, T. van Waterschoot, M. Mongelli, R. Barbieri, and A. Paglialonga, "Evaluation of a Novel Speech-in-Noise Test for Hearing Screening: Classification Performance and Transducers' Characteristics," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, p. 4300–4307, Dec. 2021

Peer-reviewed conference articles

C1 L. Azzimonti*, M. **Lenatti***, M. Zaffalon*, D. Simeone, A. Guergachi, K. Keshavjee, M. Mongelli, and A. Paglialonga, "Data-Driven and Expert-Informed Causal Discovery for Type 2 Diabetes Risk in Primary Care." Provisionally accepted for podium presentation at *Medical Informatics Europe 2026 (EFMI MIE 2026)*, May 25-28, 2026, Genova, Italy

C2 P. Ferraro, S. Narteni, M. **Lenatti**, F. Oliveri, C. Gemelli, C. Cabona, A. Uccelli, A. Paglialonga, M. Mongelli, and A. Schenone, "A Consensus Clustering Approach to Amyotrophic Lateral Sclerosis Phenotyping." Provisionally accepted for podium presentation at *Medical Informatics Europe 2026 (EFMI MIE 2026)*, May 25-28, 2026, Genova, Italy

C3 M. **Lenatti**, M. Zaffalon, A. Antonucci, P. F. De Paola, L. Multerer, M. Mongelli, A. Paglialonga, and L. Azzimonti, "Counterfactual inference using ordinary differential equations to assess the effect of physical activity on type 2 diabetes onset," in *Artificial*

* co-first author

Intelligence in Medicine (R. Bellazzi, J. M. Juarez Herrero, L. Sacchi, and B. Zupan, eds.), (Cham), pp. 212–221, Springer Nature Switzerland, 2025

C4 D. Console, M. **Lenatti**, D. Simeone, K. Keshavjee, A. Guergachi, M. Mongelli, and A. Paglialonga, "Exploring Prediabetes Pathways Using Explainable AI on Data from Electronic Medical Records," in *Studies in Health Technology and Informatics*, IOS Press, 2024

C5 D. Simeone, M. **Lenatti**, C. Lagoa, K. Keshavjee, A. Guergachi, F. Dabbene, and A. Paglialonga, "Multi-Input Multi-Output Dynamic Modelling of Type 2 Diabetes Progression," in *Studies in health technology and informatics*, vol. 309, pp. 228–232, IOS Press, Oct. 2023

C6 A. Paglialonga, E. M. Polo, M. **Lenatti**, M. Mollura, and R. Barbieri, "A Screening Platform for Hearing Loss and Cognitive Decline: WHISPER (Widespread Hearing Impairment Screening and PrEvention of Risk)," in *Studies in Health Technology and Informatics*, IOS Press, Oct. 2023

C7 M. **Lenatti**, A. Carlevaro, K. Keshavjee, A. Guergachi, A. Paglialonga, and M. Mongelli, "Characterization of Type 2 Diabetes Using Counterfactuals and Explainable AI," in *Studies in Health Technology and Informatics*, IOS Press, May 2022

C8 E. M. Polo, M. Mollura, M. Zanet, M. **Lenatti**, A. Paglialonga, and R. Barbieri, "Analysis of the Effect of Emotion Elicitation on the Cardiovascular System," in *2021 Computing in Cardiology (CinC)*, vol. 48, pp. 1–4, 2021

C9 E. M. Polo, M. Mollura, M. **Lenatti**, M. Zanet, A. Paglialonga, and R. Barbieri, "Emotion recognition from multimodal physiological measurements based on an interpretable feature selection method," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 989–992, 2021

C10 E. M. Polo, M. Zanet, M. **Lenatti**, T. van Waterschoot, R. Barbieri, and A. Paglialonga, "Development and evaluation of a novel method for adult hearing screening: Towards a dedicated smartphone app," in *IoT Technologies for HealthCare* (R. Goleva, N. R. d. C. Garcia, and I. M. Pires, eds.), (Cham), pp. 3–19, Springer International Publishing, 2021

Conference abstracts

A1 M. **Lenatti**, M. Zaffalon, A. Antonucci, P. F. De Paola, L. Multerer, M. Mongelli,

A. Paglialonga, and L. Azzimonti, "Differential Equations-Driven Counterfactuals For Personalized Diabetes Risk Management." Presented at the *Precision Health Day*, March 28, 2025, Lugano, Switzerland

A2 P. Ferraro, M. **Lenatti**, S. Narteni, F. Oliveri, C. Gemelli, C. Cabona, A. Uccelli, A. Paglialonga, M. Mongelli, and A. Schenone, "Multivariate clustering of motor neuron diseases partially overlaps with current clinical classifications." Presented at PISA ON ALS 2025 - Artificial Intelligence driving research and cure in ALS, SMA, and other Motor Neuron Diseases, March 21-22, 2025, Pisa, Italy

A3 A. Paglialonga, M. **Lenatti**, D. Simeone, P. De Paola, A. Carlevaro, M. Mongelli, F. Dabbene, F. Castiglione, M. Palumbo, P. Stolfi, and P. Tieri, "Towards a Digital Twin For Personalized Diabetes Prevention: The PRAESIIDIUM Project." Presented at the *BUILD-IT2023 Workshop (BUILding a DIgital Twin: requirements, methods, and applications)*, October 19-20, 2023, Rome, Italy

A4 A. Carlevaro, G. De Bernardi, M. **Lenatti**, S. Narteni, M. Muselli, A. Paglialonga, F. Dabbene, and M. Mongelli, "Are Digital Twins Suitable To Drive Safe AI?." Presented at the *BUILD-IT2023 Workshop (BUILding a DIgital Twin: requirements, methods, and applications)*, October 19-20, 2023, Rome, Italy

A5 A. Paglialonga, M. **Lenatti**, V. Orani, A. Carlevaro, S. Narteni, M. Muselli, F. Dabbene, and M. Mongelli, "AI & Health: Methods and Applications." Presented at *Ital-IA22 (Ital-IA 2022 Convegno del Laboratorio nazionale CINI-AIIS)*, February 9-11, 2022,Turin, Italy

A6 A. Carlevaro, M. Mongelli, M. **Lenatti**, M. Mammarella, M. Muselli, S. Narteni, V. Orani, F. Dabbene, and A. Paglialonga, "eXplainable and Reliable AI Approaches to Trustworthy AI." Presented at *Ital-IA22 (Ital-IA 2022 Convegno del Laboratorio nazionale CINI-AIIS)*, February 9-11, 2022,Turin, Italy

A7 M. **Lenatti**, V. Orani, E. Polo, R. Barbieri, M. Mongelli, and A. Paglialonga, "A framework of Explainable Artificial Intelligence for adult hearing screening." Presented at *Hearing Across the Lifespan Conference (HEAL 2022)*, June 16-18,2022, Cernobbio, Italy

A8 A. Paglialonga, M. **Lenatti**, E. Polo, M. Paolini, L. Petrella, M. Mollura, and R. Barbieri, "WHISPER (Widespread Hearing Impairment Screening and PrEvention of Risk): a New Platform for Early Identification of Hearing Impairment and Cognitive Decline."

Presented at Hearing Across the Lifespan Conference (HEAL 2022), June 16-18,2022, Cernobbio, Italy

A9 M. **Lenatti**, E. Polo, M. Paolini, M. Mollura, M. Zanet, R. Barbieri, and A. Paglialonga, "Evaluation of multivariate classification algorithms for hearing loss detection through a speech-in-noise test." Presented at the *2nd Virtual Conference on Computational Audiology (VCCA2021)*, June 25, 2021

A10 E. Polo, M. **Lenatti**, M. Zanet, R. Barbieri, and A. Paglialonga, "Preliminary evaluation of the Speech Reception Threshold measured using a new language-independent screening test as a predictor of hearing loss." Presented at the *1st Virtual Conference on Computational Audiology (VCCA2020)*, June 19, 2020

Submitted contributions

S1 L. Multerer, M. Acquistapace, P. De Paola, M. **Lenatti**, A. Paglialonga, Z. Šmite, J. Sokolovska, and L. Azzimonti, "Computationally efficient patient-specific modeling of type 2 diabetes progression." Submitted to *Precision Health Day 2026 – AI-Enabled*, March 27, 2026, Lugano, Switzerland. Currently under review

Bibliography

[1] H. Schmidt, "Chronic Disease Prevention and Health Promotion," in *Public Health Ethics: Cases Spanning the Globe* (D. H. Barrett, L. W. Ortmann, A. Dawson, C. Saenz, A. Reis, and G. Bolan, eds.), vol. 3, pp. 137–176, Cham: Springer International Publishing, 2016. Series Title: Public Health Ethics Analysis.

[2] World Health Organization, *World Health Statistics 2023: Monitoring Health for the SDGs, Sustainable Development Goals*. Geneva: World Health Organization, 1st ed ed., 2023.

[3] World Health Organization, "Noncommunicable diseases." Available at: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases, 2024. Accessed: Aug 25, 2025.

[4] W. H. Organization, *World Health Statistics 2025: Monitoring Health for the SDGs, Sustainable Development Goals*. Geneva: World Health Organization, 1st ed ed., 2025.

[5] K. Hacker, "The Burden of Chronic Disease," *Mayo Clinic Proceedings: Innovations, Quality & Outcomes*, vol. 8, pp. 112–119, Feb. 2024.

[6] A. S. Ibraheem, C. B. Nwosu, R. K. Omowumi, and A. O. Ayodapo, "Managing population health through prevention and early detection of cancer and other non-communicable diseases: a call for action," *Discover Public Health*, vol. 22, p. 143, Apr. 2025.

[7] M. Pan, R. Li, J. Wei, H. Peng, Z. Hu, Y. Xiong, N. Li, Y. Guo, W. Gu, and H. Liu, "Application of artificial intelligence in the health management of chronic disease: bibliometric analysis," *Frontiers in Medicine*, vol. 11, p. 1506641, Jan. 2025.

[8] A. Balagopalan, I. Baldini, L. A. Celi, J. Gichoya, L. G. McCoy, T. Naumann, U. Shalit, M. Van Der Schaar, and K. L. Wagstaff, "Machine learning for healthcare that matters: Reorienting from technical novelty to equitable impact," *PLOS Digital Health*, vol. 3, p. e0000474, Apr. 2024.

[9] Y.-M. Chen, T.-H. Hsiao, C.-H. Lin, and Y. C. Fann, "Unlocking precision medicine: clinical applications of integrating health records, genetics, and immunology through artificial intelligence," *Journal of Biomedical Science*, vol. 32, p. 16, Feb. 2025.

[10] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera, "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence," *Information Fusion*, vol. 99, p. 101805, Nov. 2023.

[11] L. Farah, J. M. Murris, I. Borget, A. Guilloux, N. M. Martelli, and S. I. Katsahian, "Assessment of Performance, Interpretability, and Explainability in Artificial Intelligence–Based Health Technologies: What Healthcare Stakeholders Need to Know," *Mayo Clinic Proceedings: Digital Health*, vol. 1, pp. 120–138, June 2023.

[12] G. Carloni, A. Berti, and S. Colantonio, "The Role of Causality in Explainable Artificial Intelligence," *WIREs Data Mining and Knowledge Discovery*, vol. 15, June 2025. Publisher: Wiley.

[13] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[14] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)." Available at: https://data.europa.eu/eli/reg/2016/679/oj, 2016. Accessed: Nov 11, 2025.

[15] European Commission and Directorate-General for Communications Networks, Content and Technology and Grupa ekspertów wysokiego szczebla ds. sztucznej inteligencji, *Ethics guidelines for trustworthy AI*. Publications Office, 2019.

[16] European Commission: Directorate-General for Communications Networks, Content and Technology, *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment*. Publications Office, 2020.

[17] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139

and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)." Available at: https://eur-lex.europa.eu/eli/reg/2024/1689/oj, 2024. Accessed: Nov 11, 2025.

[18] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, p. 82–115, June 2020.

[19] T. Speith, "A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, (Seoul Republic of Korea), pp. 2239–2250, ACM, June 2022.

[20] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A Survey of Methods for Explaining Black Box Models," *ACM Computing Surveys*, vol. 51, pp. 1–42, Sept. 2019.

[21] M. Frasca, D. La Torre, G. Pravettoni, and I. Cutica, "Explainable and interpretable artificial intelligence in medicine: a systematic bibliometric review," *Discover Artificial Intelligence*, vol. 4, Feb. 2024. Publisher: Springer Science and Business Media LLC.

[22] H. Hakkoum, I. Abnane, and A. Idri, "Interpretability in the medical field: A systematic mapping and review study," *Applied Soft Computing*, vol. 117, p. 108391, Mar. 2022.

[23] M. Alsalem, A. Alamoodi, O. Albahri, A. Albahri, L. Martínez, R. Yera, A. M. Duhaim, and I. M. Sharaf, "Evaluation of trustworthy artificial intelligent healthcare applications using multi-criteria decision-making approach," *Expert Systems with Applications*, vol. 246, p. 123066, July 2024.

[24] C. Combi, B. Amico, R. Bellazzi, A. Holzinger, J. H. Moore, M. Zitnik, and J. H. Holmes, "A manifesto on explainability for artificial intelligence in medicine," *Artificial Intelligence in Medicine*, vol. 133, p. 102423, Nov. 2022.

[25] F. Mohsen, H. R. H. Al-Absi, N. A. Yousri, N. El Hajj, and Z. Shah, "A scoping review of artificial intelligence-based methods for diabetes risk prediction," *npj Digital Medicine*, vol. 6, Oct. 2023.

[26] M. Prosperi, Y. Guo, M. Sperrin, J. S. Koopman, J. S. Min, X. He, S. Rich, M. Wang, I. E. Buchan, and J. Bian, "Causal inference and counterfactual prediction in machine

learning for actionable healthcare," *Nature Machine Intelligence*, vol. 2, pp. 369–375, July 2020.

[27] J. G. Richens, C. M. Lee, and S. Johri, "Improving the accuracy of medical diagnosis with causal machine learning," *Nature Communications*, vol. 11, p. 3923, Aug. 2020.

[28] I. Bica, A. M. Alaa, C. Lambert, and M. van der Schaar, "From Real-World Patient Data to Individualized Treatment Effects Using Machine Learning: Current and Future Methods to Address Underlying Challenges," *Clinical Pharmacology & Therapeutics*, vol. 109, p. 87–100, June 2020.

[29] P. Msaouel, J. Lee, J. A. Karam, and P. F. Thall, "A Causal Framework for Making Individualized Treatment Decisions in Oncology," *Cancers*, vol. 14, no. 16, 2022.

[30] A. Curth, R. W. Peck, E. McKinney, J. Weatherall, and M. van der Schaar, "Using machine learning to individualize treatment effect estimation: Challenges and opportunities," *Clinical Pharmacology & Therapeutics*, vol. 115, p. 710–719, Jan. 2024.

[31] S. Feuerriegel, D. Frauen, V. Melnychuk, J. Schweisthal, K. Hess, A. Curth, S. Bauer, N. Kilbertus, I. S. Kohane, and M. van der Schaar, "Causal machine learning for predicting treatment outcomes.," *Nature Medicine*, vol. 30, p. 958–968, 2024.

[32] 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*, vol. 12, p. e1449, Mar. 2022.

[33] S. Esteban and A. Szmulewicz, "Making causal inferences from transactional data: A narrative review of opportunities and challenges when implementing the target trial framework," *Journal of International Medical Research*, vol. 52, Mar. 2024.

[34] J. Shi and B. Norgeot, "Learning Causal Effects From Observational Data in Healthcare: A Review and Summary," *Frontiers in Medicine*, vol. 9, p. 864882, July 2022.

[35] J. M. Smit, J. H. Krijthe, W. M. R. Kant, J. A. Labrecque, M. Komorowski, D. A. M. P. J. Gommers, J. Van Bommel, M. J. T. Reinders, and M. E. Van Genderen, "Causal inference using observational intensive care unit data: a scoping review and recommendations for future practice," *npj Digital Medicine*, vol. 6, p. 221, Nov. 2023.

[36] X.-E. Gao, J.-G. Hu, B. Chen, Y.-M. Wang, and S.-B. Zhou, "Causal discovery approach with reinforcement learning for risk factors of type II diabetes mellitus," *BMC Bioinformatics*, vol. 24, p. 296, July 2023.

[37] X. Shen, S. Ma, P. Vemuri, M. R. Castro, P. J. Caraballo, and G. J. Simon, "A novel method for causal structure discovery from EHR data and its application to type-2 diabetes mellitus," *Scientific Reports*, vol. 11, p. 21025, Oct. 2021.

[38] W. Wang, G. Hu, B. Yuan, S. Ye, C. Chen, Y. Cui, X. Zhang, and L. Qian, "Prior-Knowledge-Driven Local Causal Structure Learning and Its Application on Causal Discovery Between Type 2 Diabetes and Bone Mineral Density," *IEEE Access*, vol. 8, pp. 108798–108810, 2020.

[39] G. F. Marchezini, A. M. Lacerda, G. L. Pappa, J. Meira, Wagner, D. Miranda, M. A. Romano-Silva, D. S. Costa, and L. M. Diniz, "Counterfactual inference with latent variable and its application in mental health care," *Data Mining and Knowledge Discovery*, vol. 36, p. 811–840, Jan. 2022.

[40] A. Zanga, A. Bernasconi, P. J. F. Lucas, H. Pijnenborg, C. Reijnen, M. Scutari, and F. Stella, "Risk Assessment of Lymph Node Metastases in Endometrial Cancer Patients: A Causal Approach," *arXiv preprint*, 2023. Available at: http://arxiv.org/abs/2305.10041.

[41] A. Balordi, A. Bernasconi, A. Andreotti, S. Guzzinati, R. Cabañas De Paz, and A. Zanga, "On Counterfactual Explanations of Cardiovascular Risk in Adolescent and Young Adult Breast Cancer Survivors," *Journal of Medical Systems*, vol. 49, p. 140, Oct. 2025.

[42] H. Kern, G. Corani, D. Huber, N. Vermes, M. Zaffalon, M. Varini, C. Wenzel, and A. Fringer, "Impact on place of death in cancer patients: a causal exploration in southern Switzerland," *BMC Palliative Care*, vol. 19, p. 160, Dec. 2020.

[43] X. Shen, S. Ma, P. Vemuri, G. Simon, and the Alzheimer's Disease neuroimaging initiative, "Challenges and Opportunities with Causal Discovery Algorithms: Application to Alzheimer's Pathophysiology," *Scientific Reports*, vol. 10, Feb. 2020.

[44] B. K. Beaulieu-Jones, S. G. Finlayson, W. Yuan, R. B. Altman, I. S. Kohane, V. Prasad, and K. Yu, "Examining the Use of Real-World Evidence in the Regulatory Process," *Clinical Pharmacology & Therapeutics*, vol. 107, pp. 843–852, Apr. 2020.

[45] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *WIREs Data Mining and Knowledge Discovery*, vol. 9, Apr. 2019.

[46] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR," *Harvard Journal of Law & Technology*, pp. 841–887, 2018.

[47] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, Feb. 2019.

[48] Y.-L. Chou, C. Moreira, P. Bruza, C. Ouyang, and J. Jorge, "Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications," *Information Fusion*, vol. 81, pp. 59–83, May 2022. Publisher: Elsevier BV.

[49] A.-H. Karimi, B. Schölkopf, and I. Valera, "Algorithmic Recourse: from Counterfactual Explanations to Interventions," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, (Virtual Event Canada), pp. 353–362, ACM, Mar. 2021.

[50] N. Savage, "Why artificial intelligence needs to understand consequences," *Nature Outlook: Robotics and artificial intelligence*, Feb. 2023.

[51] "Canadian Primary Care Sentinel Surveillance Network (CPCSSN) [Online]." Available at: http://cpcssn.ca/. Accessed: Oct 23, 2025.

[52] P. F. De Paola, A. Borri, F. Dabbene, K. Keshavjee, P. Palumbo, and A. Paglialonga, "Modeling the cumulative benefits of regular physical activity on type 2 diabetes progression," *Computers in Biology and Medicine*, vol. 198, p. 111194, 2025.

[53] D. M. J. Tax and R. P. W. Duin, "Support Vector Data Description," *Machine Learning*, vol. 54, pp. 45–66, 2004.

[54] V. Belle and I. Papantonis, "Principles and Practice of Explainable Machine Learning," *Frontiers in Big Data*, vol. 4, p. 688969, July 2021.

[55] G. Vilone and L. Longo, "Classification of Explainable Artificial Intelligence Methods through Their Output Formats," *Machine Learning and Knowledge Extraction*, vol. 3, pp. 615–661, Aug. 2021.

[56] W. Samek and K.-R. Müller, *Towards Explainable Artificial Intelligence*, pp. 5–22. Cham: Springer International Publishing, 2019.

[57] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg,

S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[58] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), p. 1135–1144, Association for Computing Machinery, 2016.

[59] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation," *Journal of Computational and Graphical Statistics*, vol. 24, no. 1, pp. 44–65, 2015.

[60] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017.

[61] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, p. 56–67, Jan. 2020.

[62] J. A. McDermid, Y. Jia, Z. Porter, and I. Habli, "Artificial intelligence explainability: the technical and ethical dimensions," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 379, p. 20200363, Aug. 2021.

[63] R. Guidotti, "Counterfactual explanations and how to find them: literature review and benchmarking," *Data Mining and Knowledge Discovery*, vol. 38, p. 2770–2824, 2024.

[64] S. Verma, V. Boonsanong, M. Hoang, K. Hines, J. Dickerson, and C. Shah, "Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review," *ACM Computing Surveys*, vol. 56, pp. 1–42, Dec. 2024. Publisher: Association for Computing Machinery (ACM).

[65] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," Jan. 2020.

[66] I. Stepin, J. M. Alonso, A. Catala, and M. Pereira-Farina, "A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence," *IEEE Access*, vol. 9, pp. 11974–12001, 2021. Publisher: Institute of Electrical and Electronics Engineers (IEEE).

[67] K. Kanamori, T. Takagi, K. Kobayashi, and H. Arimura, "DACE: Distribution-Aware Counterfactual Explanation by Mixed-Integer Linear Optimization," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, (Yokohama, Japan), pp. 2855–2862, International Joint Conferences on Artificial Intelligence Organization, July 2020.

[68] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das, "Explanations based on the missing: towards contrastive explanations with pertinent negatives," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, (Red Hook, NY, USA), p. 590–601, Curran Associates Inc., 2018.

[69] D. Mahajan, C. Tan, and A. Sharma, "Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers," *arXiv preprint*, June 2020. Available at: http://arxiv.org/abs/1912.03277.

[70] "Github repository: Dice." Available at: https://github.com/interpretml/DiCE, 2020.

[71] F. Bodria, R. Guidotti, F. Giannotti, and D. Pedreschi, "Transparent Latent Space Counterfactual Explanations for Tabular Data," in *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, (Shenzhen, China), pp. 1–10, IEEE, Oct. 2022.

[72] R. Crupi, A. Castelnovo, D. Regoli, and B. San Miguel Gonzalez, "Counterfactual explanations as interventions in latent space," *Data Mining and Knowledge Discovery*, vol. 38, pp. 2733–2769, Sept. 2024. Publisher: Springer Science and Business Media LLC.

[73] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.

[74] J. Peters, D. Janzing, and B. Schlkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

[75] M. G. Judea Pearl and N. P. Jewell, *Causal Inference in Statistics: A Primer*. John Wiley & Sons, 2019.

[76] A. Zanga, E. Ozkirimli, and F. Stella, "A survey on causal discovery: Theory and practice," *International Journal of Approximate Reasoning*, vol. 151, pp. 101–129, 2022.

[77] M. C. Vonk, N. Malekovic, T. Bäck, and A. V. Kononova, "Disentangling causality: assumptions in causal discovery and inference," *Artificial Intelligence Review*, vol. 56, p. 10613–10649, Feb. 2023.

[78] N. K. Kitson and A. C. Constantinou, "Causal discovery using dynamically requested knowledge," *Knowledge-Based Systems*, vol. 314, p. 113185, 2025.

[79] P. Spirtes and C. Glymour, "An Algorithm for Fast Recovery of Sparse Causal Graphs," *Social Science Computer Review*, vol. 9, no. 1, pp. 62–72, 1991.

[80] P. Spirtes, C. Meek, and T. Richardson, "Causal inference in the presence of latent variables and selection bias," in *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, (San Francisco, CA, USA), p. 499–506, Morgan Kaufmann Publishers Inc., 1995.

[81] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The max-min hill-climbing Bayesian network structure learning algorithm," *Machine Learning*, vol. 65, p. 31–78, Mar. 2006.

[82] H. Akaike, *A New Look at the Statistical Model Identification*, pp. 215–222. New York, NY: Springer New York, 1998.

[83] G. Schwarz, "Estimating the Dimension of a Model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[84] R. R. Bouckaert, "Optimizing Causal Orderings for Generating DAGs from Data," in *Uncertainty in Artificial Intelligence* (D. Dubois, M. P. Wellman, B. D'Ambrosio, and P. Smets, eds.), pp. 9–16, Morgan Kaufmann, 1992.

[85] D. M. Chickering, "Optimal structure identification with greedy search," *J. Mach. Learn. Res.*, vol. 3, p. 507–554, Mar. 2003.

[86] N. Friedman, "Learning Belief Networks in the Presence of Missing Values and Hidden Variables," in *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, (San Francisco, CA, USA), p. 125–133, Morgan Kaufmann Publishers Inc., 1997.

[87] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[88] A. C. Constantinou, Z. Guo, and N. K. Kitson, "The impact of prior knowledge on causal structure learning," *Knowledge and Information Systems*, vol. 65, pp. 3385–3434, Aug. 2023.

[89] J. Peters and P. Bühlmann, "Structural intervention distance for evaluating causal graphs," *Neural Computation*, vol. 27, p. 771–799, Mar. 2015.

[90] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies.," *Journal of Educational Psychology*, vol. 66, pp. 688–701, 1974.

[91] D. Galles and J. Pearl, "An Axiomatic Characterization of Causal Counterfactuals," *Foundations of Science*, vol. 3, p. 151–182, Jan. 1998.

[92] D. Ibeling and T. Icard, "Comparing Causal Frameworks: Potential Outcomes, Structural Models, Graphs, and Abstractions," in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 80130–80141, Curran Associates, Inc., 2023.

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

[94] J. Pearl, "An Introduction to Causal Inference," *The International Journal of Biostatistics*, vol. 6, no. 2, 2010.

[95] H. Geffner, R. Dechter, and J. Y. Halpern, eds., *Probabilistic and Causal Inference: The Works of Judea Pearl*, vol. 36. New York, NY, USA: Association for Computing Machinery, 1 ed., 2022.

[96] J. Pearl, "Causal and Counterfactual Inference," in *The Handbook of Rationality* (M. Knauff and W. Spohn, eds.), pp. 427–438, The MIT Press, Dec. 2021.

[97] A. Balke and J. Pearl, "Bounds on Treatment Effects from Studies with Imperfect Compliance," *Journal of the American Statistical Association*, vol. 92, no. 439, pp. 1171–1176, 1997.

[98] J. Zhang, J. Tian, and E. Bareinboim, "Partial counterfactual identification from observational and experimental data," in *Proceedings of the 39th International Conference on Machine Learning* (K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, eds.), vol. 162, pp. 26548–26558, PMLR, Jul 2022.

[99] M. Zaffalon, A. Antonucci, R. Cabañas, D. Huber, and D. Azzimonti, "Efficient computation of counterfactual bounds," *International Journal of Approximate Reasoning*, p. 109111, 2024.

[100] M. Zaffalon and A. Antonucci, "A Note on Bayesian Networks with Latent Root Variables," *arXiv preprint*, 2024.

[101] A. Carlevaro*, M. **Lenatti***, A. Paglialonga, and M. Mongelli, "Multiclass Counterfactual Explanations Using Support Vector Data Description," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 6, pp. 3046–3056, 2024.

[102] M. **Lenatti**, A. Carlevaro, A. Guergachi, K. Keshavjee, M. Mongelli, and A. Paglialonga, "Estimation and Conformity Evaluation of Multi-Class Counterfactual Explanations for Chronic Disease Prevention," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 9, pp. 6132–6142, 2025.

[103] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. Springer-Verlag, 2005.

[104] P. Mills, "Solving for multi-class: a survey and synthesis," *arXiv preprint*, 2018. Available at: http://arxiv.org/abs/1809.05929.

[105] D. M. J. Tax and R. P. W. Duin, "Support vector domain description," *Pattern Recognition Letters*, vol. 20, pp. 1191–1199, 1999.

[106] S. S. Khan and M. G. Madden, "One-class classification: taxonomy of study and review of techniques," *The Knowledge Engineering Review*, vol. 29, no. 3, pp. 345–374, 2014.

[107] M. Turkoz, S. Kim, Y. Son, M. K. Jeong, and E. A. Elsayed, "Generalized support vector data description for anomaly detection," *Pattern Recognition*, vol. 100, p. 107119, 2020.

[108] G. Huang, H. Chen, Z. Zhou, F. Yin, and K. Guo, "Two-class support vector data description," *Pattern Recognition*, vol. 44, no. 2, pp. 320–329, 2011.

[109] L. Duan, M. Xie, T. Bai, and J. Wang, "A new support vector data description method for machinery fault diagnosis with unbalanced datasets," *Expert Systems with Applications*, vol. 64, pp. 239–246, 2016.

[110] W. Guo, Z. Wang, S. Hong, D. Li, H. Yang, and W. Du, "Multi-kernel Support Vector Data Description with boundary information," *Engineering Applications of Artificial Intelligence*, vol. 102, p. 104254, 2021.

[111] H. Hou and H. Ji, "Improved multiclass support vector data description for planetary gearbox fault diagnosis," *Control Engineering Practice*, vol. 114, p. 104867, 2021.

[112] J. Fang, W. Wang, X. Wang, Z. Long, D. Liang, and Q. Zhou, "A SVDD method based on maximum distance between two centers of spheres," *Chinese Journal of Electronics*, vol. 21, no. 1, p. 107 – 111, 2012.

[113] J. Mercer, "Functions of Positive and Negative Type, and their Connection with the Theory of Integral Equations," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 209, pp. 415–446, 1909.

[114] A. Carlevaro and M. Mongelli, "A New SVDD Approach to Reliable and Explainable AI," *IEEE Intelligent Systems*, vol. 37, p. 55–68, Mar. 2022.

[115] C. Cervellera, M. Gaggero, D. Macciò, and R. Marcialis, "Quasi-random sampling for approximate dynamic programming," in *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2013.

[116] N. Cesa-Bianchi, Y. Mansour, and O. Shamir, "On the Complexity of Learning with Kernels," in *Proceedings of The 28th Conference on Learning Theory*, vol. 40 of *Proceedings of Machine Learning Research*, (Paris, France), pp. 297–325, 2015.

[117] A. Carlevaro, M. **Lenatti**, A. Paglialonga, and M. Mongelli, "Counterfactual Building and Evaluation via eXplainable Support Vector Data Description," *IEEE Access*, vol. 10, pp. 60849–60861, 2022.

[118] S. Sen, T. Samanta, and A. Reese, "Quasi-versus pseudo-random generators: Discrepancy, complexity and integration-error based comparison," *Int J Innov Comput Info Control*, vol. 2, 2006.

[119] Y. Burago and D. Shoenthal, "Metric geometry," in *New Analytic and Geometric Methods in Inverse Problems* (K. Bingham, Y. V. Kurylev, and E. Somersalo, eds.), (Berlin, Heidelberg), pp. 3–50, Springer Berlin Heidelberg, 2004.

[120] M. Blum, R. W. Floyd, V. Pratt, R. L. Rivest, and R. E. Tarjan, "Time Bounds for Selection," *J. Comput. Syst. Sci.*, vol. 7, no. 4, p. 448–461, 1973.

[121] X. Huang, G. Jin, and W. Ruan, *Machine Learning Safety*. Springer Nature Singapore, 2023.

[122] A. N. Angelopoulos and S. Bates, "Conformal Prediction: A Gentle Introduction," *Foundations and Trends in Machine Learning*, vol. 16, no. 4, p. 494–591, 2023.

[123] Global initiative for chronic obstructive lung disease, "Global strategy for prevention, diagnosis and management of COPD: 2025 Report [Online]." Available at: https://goldcopd.org/2025-gold-report/, 2025. Accessed: Nov 21, 2025.

[124] D. Adeloye, P. Song, Y. Zhu, H. Campbell, A. Sheikh, and I. Rudan, "Global, regional, and national prevalence of, and risk factors for, chronic obstructive pulmonary disease (COPD) in 2019: a systematic review and modelling analysis," *The Lancet Respiratory Medicine*, vol. 10, p. 447–458, May 2022.

[125] A. D. Morgan, R. Zakeri, and J. K. Quint, "Defining the relationship between COPD and CVD: what are the implications for clinical practice?," *Therapeutic Advances in Respiratory Disease*, vol. 12, Jan. 2018.

[126] W. Chen, J. Thomas, M. Sadatsafavi, and J. M. FitzGerald, "Risk of cardiovascular comorbidity in patients with chronic obstructive pulmonary disease: a systematic review and meta-analysis," *The Lancet Respiratory Medicine*, vol. 3, p. 631–639, Aug. 2015.

[127] H. Müllerová, J. Marshall, E. de Nigris, P. Varghese, N. Pooley, N. Embleton, C. Nordon, and Z. Marjenberg, "Association of COPD exacerbations and acute cardiovascular events: a systematic review and meta-analysis," *Therapeutic Advances in Respiratory Disease*, vol. 16, Jan. 2022.

[128] K. Brassington, S. Selemidis, S. Bozinovski, and R. Vlahos, "New frontiers in the treatment of comorbid cardiovascular disease in chronic obstructive pulmonary disease," *Clinical Science*, vol. 133, p. 885–904, Apr. 2019.

[129] K. Brassington, S. Selemidis, S. Bozinovski, and R. Vlahos, "Chronic obstructive pulmonary disease and atherosclerosis: common mechanisms and novel therapeutics," *Clinical Science*, vol. 136, p. 405–423, Mar. 2022.

[130] Canadian Cardiovascular Society, "Framingham Risk Score Worksheet [Online]." Available at: https://ccs.ca/app/uploads/2020/12/FRS_eng_2017_fnl_greyscale.pdf. Accessed: Jan 16, 2024.

[131] Y.-L. Chen, Y. F. Zheng, and Y. Liu, "Margin and domain integrated classification for images," *International Journal of Information Acquisition*, vol. 08, p. 1–16, Mar. 2011.

[132] J. Cohen, "A Coefficient of Agreement for Nominal Scales," 1960.

[133] M. **Lenatti**, A. Carlevaro, A. Guergachi, K. Keshavjee, M. Mongelli, and A. Paglialonga, "A novel method to derive personalized minimum viable recommendations for type 2 diabetes prevention based on counterfactual explanations," *PLOS ONE*, vol. 17, p. e0272825, Nov. 2022.

[134] L. Azzimonti*, M. **Lenatti***, M. Zaffalon*, D. Simeone, A. Guergachi, K. Keshavjee, M. Mongelli, and A. Paglialonga, "Data-Driven and Expert-Informed Causal Discovery for Type 2 Diabetes Risk in Primary Care." Provisionally accepted for podium presentation at *Medical Informatics Europe 2026 (EFMI MIE 2026)*, May 25-28, 2026, Genova, Italy.

[135] International Diabetes Federation, "IDF Diabetes Atlas, 11th edn. Brussels, Belgium [Online]." Available at: https://diabetesatlas.org, 2025. Accessed: Nov 21, 2025.

[136] American Diabetes Association, "2. Classification and diagnosis of diabetes: Standards of medical care in diabetes—2021," *Diabetes Care*, vol. 44, no. Supplement_1, pp. S15–S33, 2020.

[137] K. Keshavjee, J. Candeliere, F. Cepeda, M. Mittal, S. Ali, and A. Guergachi, "A framework for implementing disease prevention and behavior change evidence at scale," *Studies in Health Technology and Informatics*, pp. 3–8, Feb. 2024.

[138] The Diabetes Prevention Program (DPP) Research Group, "The Diabetes Prevention Program (DPP): Description of lifestyle intervention," *Diabetes Care*, vol. 25, pp. 2165–2171, 12 2002.

[139] J. B. Sussman, D. M. Kent, J. P. Nelson, and R. A. Hayward, "Improving diabetes prevention with benefit based tailored treatment: risk based reanalysis of diabetes prevention program," *BMJ*, vol. 350, 2015.

[140] S. C. Y. Wang, G. Nickel, K. P. Venkatesh, M. M. Raza, and J. C. Kvedar, "AI-based diabetes care: risk prediction models and implementation concerns," *npj Digital Medicine*, vol. 7, pp. 36, s41746–024–01034–7, Feb. 2024.

[141] S. Kodama, K. Fujihara, C. Horikawa, M. Kitazawa, M. Iwanaga, K. Kato, K. Watanabe, Y. Nakagawa, T. Matsuzaka, H. Shimano, and H. Sone, "Predictive ability of current machine learning algorithms for type 2 diabetes mellitus: A meta-analysis," *Journal of Diabetes Investigation*, vol. 13, p. 900–908, Jan. 2022.

[142] A. D. Sarma and M. Devi, "Artificial intelligence in diabetes management: transformative potential, challenges, and opportunities in healthcare," *Hormones*, Mar. 2025.

[143] Y. Wang, W. S. Zhang, Y. T. Hao, C. Q. Jiang, Y. L. Jin, K. K. Cheng, T. H. Lam, and L. Xu, "A Bayesian network model of new-onset diabetes in older Chinese: The Guangzhou biobank cohort study," *Frontiers in Endocrinology*, vol. 13, p. 916851, Aug. 2022.

[144] S. Kalia, O. Saarela, B. O'Neill, C. Meaney, R. Moineddin, F. Sullivan, and M. Greiver, "Emulating a Target Trial Using Primary-Care Electronic Health Records: Sodium-Glucose Cotransporter 2 Inhibitor Medications and Hemoglobin A1c," *American Journal of Epidemiology*, vol. 192, pp. 782–789, 01 2023.

[145] D. Simeone, M. **Lenatti**, C. Lagoa, K. Keshavjee, A. Guergachi, F. Dabbene, and A. Paglialonga, "Multi-Input Multi-Output Dynamic Modelling of Type 2 Diabetes Progression," in *Studies in health technology and informatics*, vol. 309, pp. 228–232, IOS Press, Oct. 2023.

[146] M. Scutari, "Learning Bayesian Networks with the bnlearn R Package," *Journal of Statistical Software*, vol. 35, no. 3, pp. 1–22, 2010.

[147] R. Cabañas, A. Antonucci, D. Huber, and M. Zaffalon, "CREDICI: A Java Library for Causal Inference by Credal Networks," in *Proceedings of the 10th International Conference on Probabilistic Graphical Models* (M. Jaeger and T. D. Nielsen, eds.), vol. 138 of *Proceedings of Machine Learning Research*, pp. 597–600, PMLR, 23–25 Sep 2020.

[148] G. Schwarz, "Estimating the Dimension of a Model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[149] M. A. Brookhart, T. Stürmer, R. J. Glynn, J. Rassen, and S. Schneeweiss, "Confounding Control in Healthcare Database Research," *Medical Care*, vol. 48, p. S114–S120, jun2010.

[150] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian Network Classifiers," *Machine Learning*, vol. 29, no. 2/3, p. 131–163, 1997.

[151] A. Hajizadeh, S. Howes, A. Theodoulou, E. Klemperer, J. Hartmann-Boyce, J. Livingstone-Banks, and N. Lindson, "Antidepressants for smoking cessation," *Cochrane Database of Systematic Reviews*, vol. 2023, May 2023.

[152] D. T. Felson, "Obesity and Knee Osteoarthritis," *Annals of Internal Medicine*, vol. 109, p. 18, July 1988.

[153] D. Strikić, A. Vujević, D. Perica, D. Leskovar, K. Paponja, I. Pećin, and I. Merćep, "Importance of Dyslipidaemia Treatment in Individuals with Type 2 Diabetes Mellitus—A Narrative Review," *Diabetology*, vol. 4, p. 538–552, Dec. 2023.

[154] S. Paredes, L. Fonseca, L. Ribeiro, H. Ramos, J. C. Oliveira, and I. Palma, "Novel and traditional lipid profiles in Metabolic Syndrome reveal a high atherogenicity," *Scientific Reports*, vol. 9, Aug. 2019.

[155] M. L. Matey-Hernandez, F. M. K. Williams, T. Potter, A. M. Valdes, T. D. Spector, and C. Menni, "Genetic and microbiome influence on lipid metabolism and dyslipidemia," *Physiological Genomics*, vol. 50, p. 117–126, Feb. 2018.

[156] R. Ruze, T. Liu, X. Zou, J. Song, Y. Chen, R. Xu, X. Yin, and Q. Xu, "Obesity and type 2 diabetes mellitus: connections in epidemiology, pathogenesis, and treatments," *Frontiers in Endocrinology*, vol. 14, Apr. 2023.

[157] B. Hemmingsen, G. Gimenez-Perez, D. Mauricio, M. Roqué i Figuls, M.-I. Metzendorf, and B. Richter, "Diet, physical activity or both for prevention or delay of type 2 diabetes mellitus and its associated complications in people at increased risk of developing type 2 diabetes mellitus," *Cochrane Database of Systematic Reviews*, vol. 2017, Dec. 2017.

[158] T. A. Khan, D. Field, V. Chen, S. Ahmad, S. B. Mejia, H. Kahleová, and et al., "Combination of multiple low-risk lifestyle behaviors and incident type 2 diabetes: a systematic review and dose-response meta-analysis of prospective cohort studies," *Diabetes Care*, vol. 46, no. 3, pp. 643–656, 2023.

[159] J. Lindström, A. Louheranta, M. Mannelin, M. Rastas, V. Salminen, J. Eriksson, M. Uusitupa, J. Tuomilehto, and for the Finnish Diabetes Prevention Study Group, "The Finnish Diabetes Prevention Study (DPS): Lifestyle intervention and 3-year results on diet and physical activity ," *Diabetes Care*, vol. 26, pp. 3230–3236, 12 2003.

[160] World Health Organization, *WHO guidelines on physical activity and sedentary behaviour*. World Health Organization, 2020.

[161] A. Böhm, C. Weigert, H. Staiger, and H.-U. Häring, "Exercise and diabetes: relevance and causes for response variability," *Endocrine*, vol. 51, no. 3, p. 390–401, 2016.

[162] L. M. Ross, C. A. Slentz, and W. E. Kraus, "Evaluating Individual Level Responses to Exercise for Health Outcomes in Overweight or Obese Adults," *Frontiers in Physiology*, vol. 10, 2019.

[163] M. **Lenatti**, M. Zaffalon, A. Antonucci, P. F. De Paola, L. Multerer, M. Mongelli, A. Paglialonga, and L. Azzimonti, "Counterfactual inference using ordinary differential equations to assess the effect of physical activity on type 2 diabetes onset," in *Artificial Intelligence in Medicine* (R. Bellazzi, J. M. Juarez Herrero, L. Sacchi, and B. Zupan, eds.), (Cham), pp. 212–221, Springer Nature Switzerland, 2025.

[164] R. Bergman, L. Phillips, and C. Cobelli, "Physiologic Evaluation of Factors Controlling Glucose Tolerance in Man: Measurement of Insulin Sensitivity and Beta-Cell Glucose Sensitivity from the Response to Intravenous Glucose.," *Journal of Clinical Investigation*, vol. 68, pp. 1456–1467, Dec. 1981.

[165] A. Roy and R. Parker, "Dynamic Modeling of Exercise Effects on Plasma Glucose and Insulin Levels," *Journal of Diabetes Science and Technology*, vol. 1, pp. 338–347, May 2007.

[166] B. Topp, K. Promislow, G. Devries, R. Miura, and D. Finegood, "A Model of Beta-Cell Mass, Insulin, and Glucose Kinetics: Pathways to Diabetes," *Journal of Theoretical Biology*, vol. 206, pp. 605–619, Oct. 2000.

[167] J. Ha, L. Satin, and A. Sherman, "A Mathematical Model of the Pathogenesis, Prevention, and Reversal of Type 2 Diabetes," *Endocrinology*, vol. 157, pp. 624–635, Feb. 2016.

[168] P. De Paola, A. Paglialonga, P. Palumbo, K. Keshavjee, F. Dabbene, and A. Borri, "The Long-Term Effects of Physical Activity on Blood Glucose Regulation: A Model to Unravel Diabetes Progression," *IEEE Control Systems Letters*, vol. 7, pp. 2916–2921, 2023.

[169] R. Bergman, "Minimal Model: Perspective from 2005," *Hormone Research in Paediatrics*, vol. 64, no. Suppl. 3, pp. 8–15, 2005.

[170] M. Palumbo, M. Morettini, P. Tieri, F. Diele, M. Sacchetti, and F. Castiglione, "Personalizing physical exercise in a computational model of fuel homeostasis," *PLOS Computational Biology*, vol. 14, p. e1006073, Apr. 2018.

[171] M. Palumbo, A. De Graaf, M. Morettini, P. Tieri, S. Krishnan, and F. Castiglione, "A computational model of the effects of macronutrients absorption and physical exercise on hormonal regulation and metabolic homeostasis," *Computers in Biology and Medicine*, vol. 163, p. 107158, Sept. 2023.

[172] A. De Gaetano, T. Hardy, B. Beck, E. Abu-Raddad, P. Palumbo, J. Bue-Valleskey, and N. Pørksen, "Mathematical Models of Diabetes Progression," *American Journal of Physiology-Endocrinology and Metabolism*, vol. 295, pp. E1462–E1479, Dec. 2008.

[173] A. De Gaetano and T. Hardy, "A Novel Fast-Slow Model of Diabetes Progression: Insights into Mechanisms of Response to the Interventions in the Diabetes Prevention Program," *PLOS ONE*, vol. 14, p. e0222833, Oct. 2019.

[174] M. Morettini, M. Palumbo, M. Sacchetti, F. Castiglione, and C. Mazzà, "A System Model of the Effects of Exercise on Plasma Interleukin-6 Dynamics in Healthy Individuals: Role of Skeletal Muscle and Adipose Tissue," *PLOS ONE*, vol. 12, p. e0181224, July 2017.

[175] L. Multerer, P. F. De Paola, M. **Lenatti**, A. Paglialonga, and L. Azzimonti, "A reduced model for the long-term effects of physical activity on type 2 diabetes," *Mathematical Biosciences* , vol. 394, p. 109645, 2026.

[176] N. Bakhvalov and G. Panasenko, *Homogenisation: Averaging Processes in Periodic Media*, vol. 36 of *Mathematics and its Applications*. Dordrecht: Springer Netherlands, 1989.

[177] D. Cioranescu and P. Donato, *An Introduction to Homogenization*. No. 17 in Oxford lecture series in mathematics and its applications, Oxford ; New York: Oxford University Press, 1999.

[178] J. Sanders, F. Verhulst, and J. Murdock, *Averaging Methods in Nonlinear Dynamical Systems*, vol. 59 of *Applied Mathematical Sciences*. New York, NY: Springer New York, 2007.

[179] G. Pavliotis and A. Stuart, *Multiscale Methods: Averaging and Homogenization*. No. 53 in Texts in Applied Mathematics, New York, NY: Springer, 2008.

[180] F. Verhulst, *Methods and Applications of Singular Perturbations*, vol. 50 of *Texts in Applied Mathematics*. New York, NY: Springer New York, 2005.

[181] P. F. De Paola, A. Borri, A. Paglialonga, P. Palumbo, and F. Dabbene, "A model-based approach for glucose control via physical activity," in *Studies in Health Technology and Informatics*, 2025. In press.

[182] H. Langtangen and G. Pedersen, *Scaling of Differential Equations*. Cham: Springer International Publishing, 2016.

[183] S. Bongers, T. Blom, and J. M. Mooij, "Causal modeling of dynamical systems," *arXiv preprint*, 2022. Available at: http://arxiv.org/abs/1803.08784.

[184] J. Pearl, "Probabilities of causation: three counterfactual interpretations and their identification," in *Probabilistic and Causal Inference*, pp. 317–372, New York, NY, USA: ACM, 1 ed., 2022.

[185] I. M. Sobol', "Sensitivity estimates for nonlinear mathematical models," *Matematicheskoe modelirovanie*, vol. 2, no. 1, pp. 112–118, 1990.

[186] T. Iwanaga, W. Usher, and J. Herman, "Toward SALib 2.0: Advancing the accessibility and interpretability of global sensitivity analyses," *Socio-Environmental Systems Modelling*, vol. 4, p. 18155, 2022.

[187] P. P. Khin, J. H. Lee, and H.-S. Jun, "Pancreatic Beta-cell Dysfunction in Type 2 Diabetes," *European Journal of Inflammation*, vol. 21, 2023.

[188] J. Boonpor, S. Parra-Soto, F. Petermann-Rocha, N. Lynskey, V. Cabanas-Sánchez, N. Sattar, and et al., "Dose–response relationship between device-measured physical activity and incident type 2 diabetes: findings from the uk biobank prospective cohort study," *BMC Medicine*, vol. 21, no. 1, 2023.

[189] M. Naser, "An engineer's guide to explainable artificial intelligence and interpretable machine learning: Navigating causality, forced goodness, and the false perception of inference," *Automation in Construction*, vol. 129, p. 103821, 2021.

[190] E. H. Simpson, "The interpretation of interaction in contingency tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 13, no. 2, pp. 238–241, 1951.

[191] R. Hamon, H. Junklewitz, I. Sanchez, G. Malgieri, and P. De Hert, "Bridging the Gap Between AI and Explainability in the GDPR: Towards Trustworthiness-by-Design in Automated Decision-Making," *IEEE Computational Intelligence Magazine*, vol. 17, no. 1, pp. 72–85, 2022.

[192] D. Watson, "Rational shapley values," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, (New York, NY, USA), p. 1083–1094, Association for Computing Machinery, 2022.

[193] A. Zapaishchykova, D. Dreizin, Z. Li, J. Y. Wu, S. Faghihroohi, and M. Unberath, "An interpretable approach to automated severity scoring in pelvic trauma," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* (M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, eds.), (Cham), pp. 424–433, Springer International Publishing, 2021.

[194] C. Zednik and H. Boelsen, "Scientific exploration and explainable artificial intelligence," *Minds and Machines*, vol. 32, p. 219–239, Mar. 2022.

[195] S. Ghaffarian, F. R. Taghikhah, and H. R. Maier, "Explainable artificial intelligence in disaster risk management: Achievements and prospective futures," *International Journal of Disaster Risk Reduction*, vol. 98, p. 104123, 2023.

[196] D. Console, M. **Lenatti**, D. Simeone, K. Keshavjee, A. Guergachi, M. Mongelli, and A. Paglialonga, "Exploring Prediabetes Pathways Using Explainable AI on Data from Electronic Medical Records," in *Studies in Health Technology and Informatics*, IOS Press, 2024.

[197] T. L. D. . Endocrinology, "Prediabetes: much more than just a risk factor," *The Lancet Diabetes & Endocrinology*, vol. 13, p. 165, Mar. 2025.

[198] A. M. Nawi, P. S. N. M. Kamaruddin, N. R. M. Nordin, S. S. S. Soffian, and M. Baharom, "Machine learning models in prediabetes screening: A systematic review.," *Journal of Clinical & Diagnostic Research*, vol. 16, no. 5, 2022.

[199] S. R. Barber, M. J. Davies, K. Khunti, and L. J. Gray, "Risk assessment tools for detecting those with pre-diabetes: a systematic review," *Diabetes research and clinical practice*, vol. 105, no. 1, pp. 1–13, 2014.

[200] F. Guida, M. **Lenatti**, K. Keshavjee, A. Khatami, A. Guergachi, and A. Paglialonga, "Characterization of Inclination Analysis for Predicting Onset of Heart Failure from Primary Care Electronic Medical Records," *Sensors*, vol. 23, no. 9, 2023.

[201] M. **Lenatti**, A. Carlevaro, K. Keshavjee, A. Guergachi, A. Paglialonga, and M. Mongelli, "Characterization of Type 2 Diabetes Using Counterfactuals and Explainable AI," in *Studies in Health Technology and Informatics*, IOS Press, May 2022.

[202] "The English Longitudinal Study of Ageing. [Online]." at:https://www.elsa-project.ac.uk/.

[203] S. Mueller and J. Pearl, "Personalized decision making – A conceptual introduction," *Journal of Causal Inference*, vol. 11, no. 1, p. 20220050, 2023.

[204] L. Multerer, M. Acquistapace, P. De Paola, M. **Lenatti**, A. Paglialonga, Z. Šmite, J. Sokolovska, and L. Azzimonti, "Computationally efficient patient-specific modeling of type 2 diabetes progression." Submitted to *Precision Health Day 2026 – AI-Enabled*, March 27, 2026, Lugano, Switzerland. Currently under review.

[205] J. M. Robins, M. A. Hernan, and B. Brumback, "Marginal Structural Models and Causal Inference in Epidemiology," *Epidemiology*, vol. 11, p. 550–560, sep2000.

[206] I. Bica, A. M. Alaa, and M. Van Der Schaar, "Time series deconfounder: estimating treatment effects over time in the presence of hidden confounders," in *Proceedings of the 37th International Conference on Machine Learning*, ICML'20, JMLR.org, 2020.

[207] D. Cao, J. Enouen, and Y. Liu, "Estimating Treatment Effects in Continuous Time with Hidden Confounders." Presented at *ICML 2022 (Workshop Continuous time methods for machine learning)*. ArXiv preprint available at: https://arxiv.org/abs/2302.09446 .

[208] J. M. Robins, A. Rotnitzky, and D. O. Scharfstein, "Sensitivity Analysis for Selection bias and unmeasured Confounding in missing Data and Causal inference models," in *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, (New York, NY), pp. 1–94, Springer New York, 2000.

[209] A. M. Franks, A. D'Amour, and A. Feller, "Flexible Sensitivity Analysis for Observational Studies Without Observable Implications," *Journal of the American Statistical Association*, vol. 115, no. 532, pp. 1730–1746, 2020.

[210] B. Ustun, A. Spangher, and Y. Liu, "Actionable Recourse in Linear Classification," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, (New York, NY, USA), p. 10–19, Association for Computing Machinery, 2019.

[211] A. M. Salih, I. B. Galazzo, P. Gkontra, E. Rauseo, A. M. Lee, K. Lekadir, P. Radeva, S. E. Petersen, and G. Menegaz, "A review of evaluation approaches for explainable AI with applications in cardiology," *Artificial Intelligence Review*, vol. 57, p. 240, Aug. 2024.

[212] E. Hellinger, "Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen.," *Journal für die reine und angewandte Mathematik*, vol. 136, pp. 210–271, 1909.

[213] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, (New York, NY, USA), p. 2623–2631, Association for Computing Machinery, 2019.

[214] T. Williamson, "CPCSSN Disease Definitions: Canadian Primary Care Sentinel Surveillance Network (CPCSSN)." Available at: http://cpcssn.ca/research-resources/case-definitions, 2014. Accessed: Oct 23, 2025.

[215] M. Zanet, E. M. Polo, M. **Lenatti**, T. van Waterschoot, M. Mongelli, R. Barbieri, and A. Paglialonga, "Evaluation of a Novel Speech-in-Noise Test for Hearing Screening: Classification Performance and Transducers' Characteristics," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, p. 4300–4307, Dec. 2021.

[216] M. **Lenatti**, P. A. Moreno-Sánchez, E. M. Polo, M. Mollura, R. Barbieri, and A. Paglialonga, "Evaluation of Machine Learning Algorithms and Explainability Techniques to Detect Hearing Loss From a Speech-in-Noise Screening Test," *American Journal of Audiology*, vol. 31, p. 961–979, Sept 2022.

[217] A. Paglialonga, E. M. Polo, M. **Lenatti**, M. Mollura, and R. Barbieri, "A Screening Platform for Hearing Loss and Cognitive Decline: WHISPER (Widespread Hearing Impairment Screening and PrEvention of Risk)," in *Studies in Health Technology and Informatics*, IOS Press, Oct. 2023.

[218] M. **Lenatti**, A. Paglialonga, V. Orani, M. Ferretti, and M. Mongelli, "Characterization of Synthetic Health Data Using Rule-Based Artificial Intelligence Models," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 8, pp. 3760–3769, 2023.

[219] R. Atienza, *Advanced Deep Learning with Keras: Apply deep learning techniques, autoencoders, GANs, variational autoencoders, deep reinforcement learning, policy gradients, and more*. Packt Publishing, 2018.

[220] Y. Zhou, B. Wang, X. He, S. Cui, and L. Shao, "DR-GAN: Conditional Generative Adversarial Network for Fine-Grained Lesion Synthesis on Diabetic Retinopathy Images," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, p. 56–66, Jan. 2022.

[221] A. Borji, "Pros and cons of GAN evaluation measures," *Computer Vision and Image Understanding*, vol. 179, p. 41–65, Feb. 2019.

[222] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, "A Kernel Method for the Two-Sample-Problem," in *Advances in Neural Information Processing Systems* (B. Schölkopf, J. Platt, and T. Hoffman, eds.), vol. 19, MIT Press, 2006.

[223] I. Kim, A. Ramdas, A. Singh, and L. Wasserman, "Classification accuracy as a proxy for two-sample testing," *The Annals of Statistics*, vol. 49, Feb. 2021.

[224] H. Cai, B. Goggin, and Q. Jiang, "Two-sample test based on classification probability," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 13, p. 5–13, Oct. 2019.

[225] L. Le Cam and G. Lo Yang, *Asymptotics in Statistics*. Springer New York, 2000.

[226] A. Goncalves, P. Ray, B. Soper, J. Stevens, L. Coyle, and A. P. Sales, "Generation and evaluation of synthetic patient data," *BMC Medical Research Methodology*, vol. 20, May 2020.

[227] M. Muselli, "Switching Neural Networks: A New Connectionist Model for Classification," 2006.

[228] M. Muselli and E. Ferrari, "Coupling Logical Analysis of Data and Shadow Clustering for Partially Defined Positive Boolean Function Reconstruction," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, p. 37–50, Jan. 2011.

[229] Y. Goldberg, *Neural Network Methods for Natural Language Processing*. Springer International Publishing, 2017.

[230] K. Juluru, H.-H. Shih, K. N. Keshava Murthy, and P. Elnajjar, "Bag-of-Words Technique in Natural Language Processing: A Primer for Radiologists," *RadioGraphics*, vol. 41, p. 1420–1426, Sept 2021.

[231] D. Polap and M. Wlodarczyk-Sielicka, "Classification of Non-Conventional Ships Using a Neural Bag-Of-Words Mechanism," *Sensors*, vol. 20, p. 1608, Mar. 2020.

[232] A. Chio, A. Calvo, C. Moglia, L. Mazzini, and G. Mora, "Phenotypic heterogeneity of amyotrophic lateral sclerosis: a population based study," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 82, p. 740–746, Mar. 2011.

[233] P. Ferraro, M. **Lenatti**, S. Narteni, F. Oliveri, C. Gemelli, C. Cabona, A. Uccelli, A. Paglialonga, M. Mongelli, and A. Schenone, "Multivariate clustering of motor neuron diseases partially overlaps with current clinical classifications." Presented at PISA ON ALS 2025 - Artificial Intelligence driving research and cure in ALS, SMA, and other Motor Neuron Diseases, March 21-22, 2025, Pisa, Italy.

[234] P. Ferraro, S. Narteni, M. **Lenatti**, F. Oliveri, C. Gemelli, C. Cabona, A. Uccelli, A. Paglialonga, M. Mongelli, and A. Schenone, "A Consensus Clustering Approach to Amyotrophic Lateral Sclerosis Phenotyping." Provisionally accepted for podium presentation at *Medical Informatics Europe 2026 (EFMI MIE 2026)*, May 25-28, 2026, Genova, Italy.

[235] B. R. Brooks, R. G. Miller, M. Swash, and T. L. Munsat, "El Escorial revisited: Revised criteria for the diagnosis of amyotrophic lateral sclerosis," *Amyotrophic Lateral Sclerosis and Other Motor Neuron Disorders*, vol. 1, p. 293–299, Jan. 2000.

[236] J. M. Cedarbaum, N. Stambler, E. Malta, C. Fuller, D. Hilt, B. Thurmond, and A. Nakanishi, "The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function," *Journal of the Neurological Sciences*, vol. 169, p. 13–21, Oct. 1999.

[237] G. M. Bove, "Epi-perineurial anatomy, innervation, and axonal nociceptive mechanisms," *Journal of Bodywork and Movement Therapies*, vol. 12, p. 185–190, July 2008.

[238] M. Turner, A. Cagnin, F. Turkheimer, C. Miller, C. Shaw, D. Brooks, P. Leigh, and R. Banati, "Evidence of widespread cerebral microglial activation in amyotrophic lateral sclerosis: an [11c](r)-pk11195 positron emission tomography study," *Neurobiology of Disease*, vol. 15, p. 601–609, Apr. 2004.

[239] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data," *Machine Learning*, vol. 52, p. 91–118, July 2003.

[240] B. A. Hassan, N. B. Tayfor, A. A. Hassan, A. M. Ahmed, T. A. Rashid, and N. N. Abdalla, "From a-to-z review of clustering validation indices," *Neurocomputing*, vol. 601, p. 128198, Oct. 2024.

[241] Y. Qin, L. Xuan, Z. Wu, Y. Deng, B. Liu, and S. Wang, "Use of consensus clustering to identify distinct subtypes of chronic kidney disease and associated mortality risk," *Scientific Reports*, vol. 14, Dec. 2024.

[242] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision – ECCV 2016* (B. Leibe, J. Matas, N. Sebe, and M. Welling, eds.), (Cham), pp. 21–37, Springer International Publishing, 2016.

[243] G. Jocher, "Yolov5 by ultralytics (version 7.0)[computer software]." Accessible at: https://zenodo.org/record/7347926/#.ZBGNcnZByUk, 2020. Accessed: Nov 10, 2022.

[244] I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, S. Kamali, and et al., "Openimages: A public dataset for large-scale multi-label and multi-class image classification." Accessible at: https://storage.googleapis.com/openimages/web/index.html. Accessed: Jan 11, 2023.

[245] A. Ess, B. Leibe, K. Schindler, and L. Van Gool, "A mobile vision system for robust multi-person tracking," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[246] M. Braun, S. Krebs, F. Flohr, and D. M. Gavrila, "Eurocity persons: A novel benchmark for person detection in traffic scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1844–1861, 2019.

[247] M. **Lenatti**, S. Narteni, A. Paglialonga, V. Rampa, and M. Mongelli, "Dual-View Single-Shot Multibox Detector at Urban Intersections: Settings and Performance Evaluation," *Sensors*, vol. 23, no. 6, 2023.

[248] E. M. Polo, M. Mollura, M. Zanet, M. **Lenatti**, A. Paglialonga, and R. Barbieri, "Analysis of the Effect of Emotion Elicitation on the Cardiovascular System," in *2021 Computing in Cardiology (CinC)*, vol. 48, pp. 1–4, 2021.

[249] E. M. Polo, M. Mollura, M. **Lenatti**, M. Zanet, A. Paglialonga, and R. Barbieri, "Emotion recognition from multimodal physiological measurements based on an interpretable feature selection method," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 989–992, 2021.

[250] E. M. Polo, M. Zanet, M. **Lenatti**, T. van Waterschoot, R. Barbieri, and A. Paglialonga, "Development and evaluation of a novel method for adult hearing screening: Towards a dedicated smartphone app," in *IoT Technologies for HealthCare* (R. Goleva, N. R. d. C. Garcia, and I. M. Pires, eds.), (Cham), pp. 3–19, Springer International Publishing, 2021.

[251] M. **Lenatti**, M. Zaffalon, A. Antonucci, P. F. De Paola, L. Multerer, M. Mongelli, A. Paglialonga, and L. Azzimonti, "Differential Equations-Driven Counterfactuals For Personalized Diabetes Risk Management." Presented at the *Precision Health Day*, March 28, 2025, Lugano, Switzerland.

[252] A. Paglialonga, M. **Lenatti**, D. Simeone, P. De Paola, A. Carlevaro, M. Mongelli, F. Dabbene, F. Castiglione, M. Palumbo, P. Stolfi, and P. Tieri, "Towards a Digital Twin For Personalized Diabetes Prevention: The PRAESIIDIUM Project." Presented at the *BUILD-IT2023 Workshop (BUILding a DIgital Twin: requirements, methods, and applications)*, October 19-20, 2023, Rome, Italy.

[253] A. Carlevaro, G. De Bernardi, M. **Lenatti**, S. Narteni, M. Muselli, A. Paglialonga, F. Dabbene, and M. Mongelli, "Are Digital Twins Suitable To Drive Safe AI?." Presented at the *BUILD-IT2023 Workshop (BUILding a DIgital Twin: requirements, methods, and applications)*, October 19-20, 2023, Rome, Italy.

[254] A. Paglialonga, M. **Lenatti**, V. Orani, A. Carlevaro, S. Narteni, M. Muselli, F. Dabbene, and M. Mongelli, "AI & Health: Methods and Applications." Presented at *Ital-IA22 (Ital-IA 2022 Convegno del Laboratorio nazionale CINI-AIIS)*, February 9-11, 2022,Turin, Italy.

[255] A. Carlevaro, M. Mongelli, M. **Lenatti**, M. Mammarella, M. Muselli, S. Narteni, V. Orani, F. Dabbene, and A. Paglialonga, "eXplainable and Reliable AI Approaches to Trustworthy AI." Presented at *Ital-IA22 (Ital-IA 2022 Convegno del Laboratorio nazionale CINI-AIIS)*, February 9-11, 2022,Turin, Italy.

[256] M. **Lenatti**, V. Orani, E. Polo, R. Barbieri, M. Mongelli, and A. Paglialonga, "A framework of Explainable Artificial Intelligence for adult hearing screening." Presented at *Hearing Across the Lifespan Conference (HEAL 2022)*, June 16-18,2022, Cernobbio, Italy.

[257] A. Paglialonga, M. **Lenatti**, E. Polo, M. Paolini, L. Petrella, M. Mollura, and R. Barbieri, "WHISPER (Widespread Hearing Impairment Screening and PrEvention of Risk): a New Platform for Early Identification of Hearing Impairment and Cognitive Decline." Presented at Hearing Across the Lifespan Conference (HEAL 2022), June 16-18,2022, Cernobbio, Italy.

[258] M. **Lenatti**, E. Polo, M. Paolini, M. Mollura, M. Zanet, R. Barbieri, and A. Paglialonga, "Evaluation of multivariate classification algorithms for hearing loss detection

through a speech-in-noise test." Presented at the *2nd Virtual Conference on Computational Audiology (VCCA2021)*, June 25, 2021.

[259] E. Polo, M. **Lenatti**, M. Zanet, R. Barbieri, and A. Paglialonga, "Preliminary evaluation of the Speech Reception Threshold measured using a new language-independent screening test as a predictor of hearing loss." Presented at the *1st Virtual Conference on Computational Audiology (VCCA2020)*, June 19, 2020.