Universiteit Leiden

Evaluating Counterfactual Robustness Across Independently Trained Predictive Models: A Multi-Dimensional Framework for Healthcare

Duru Emekçi (s3844919)

Supervisors:
Hazel Doughty

Bachelor Thesis– Data Science and Artificial Intelligence

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

July 13, 2026

Abstract

Counterfactual explanations have emerged as one of the most widely adopted explainable artificial intelligence (XAI) techniques because they provide actionable recommendations for changing the prediction of a machine learning model. Although existing research has extensively investigated properties such as validity, sparsity, plausibility, and actionability, comparatively little attention has been given to whether counterfactual explanations remain reliable when predictive models are independently retrained for the same prediction task.

This thesis investigates the robustness of counterfactual explanations across independently trained machine learning models using a real-world healthcare prediction problem. Five XGBoost classifiers were independently trained on the Diabetes 130-US Hospitals dataset using identical data, preprocessing procedures, and hyperparameters while varying only the random seed controlling stochastic training operations. Counterfactual explanations were generated using the DiCE framework and evaluated through a multi-dimensional robustness framework consisting of four complementary metrics: cross-model validity, feature-set consistency, direction consistency, and magnitude consistency. Statistical uncertainty was quantified using patient-level bootstrap confidence intervals based on 5,000 bootstrap resamples.

The experimental results demonstrate that robustness varies substantially across different explanation dimensions. Cross-model validity (0.6475) and feature-set consistency (0.5134) indicate considerable variability in whether explanations transfer across retrained models and in the selection of actionable features. In contrast, direction consistency (0.8063) and magnitude consistency (0.8450) remain substantially higher, suggesting that once models identify common intervention features, they generally agree on both the direction and relative magnitude of the recommended changes.

These findings demonstrate that the robustness of counterfactual explanations is inherently multi-dimensional and cannot be adequately characterized using a single evaluation metric. The proposed evaluation framework provides a systematic methodology for assessing explanation reliability under model retraining and contributes to the growing body of research on trustworthy explainable artificial intelligence for high-stakes healthcare applications.

Contents

1 Introduction

1.1 Motivation

Artificial intelligence is becoming an increasingly important component of modern healthcare, supporting tasks such as disease diagnosis, patient risk stratification, treatment planning, and hospital resource management. Recent advances in machine learning have enabled predictive models to identify complex patterns in large-scale electronic health records and achieve strong performance across a wide range of clinical prediction tasks. Consequently, such models are increasingly being considered for integration into clinical decision-support systems [19, 10].

Predictive performance alone, however, is rarely sufficient in healthcare. Clinical decisions may directly affect patient safety and treatment outcomes, requiring healthcare professionals to understand not only what a model predicts, but also the basis for that prediction. The reasoning of complex machine learning models is often difficult to inspect, which can limit their practical acceptance and responsible use. Interpretability and transparency have therefore become central requirements for the deployment of artificial intelligence in clinical settings [10, 14].

Counterfactual explanations are particularly relevant in this context because they provide intuitive and action-oriented information about individual predictions. Rather than merely identifying which variables influenced an outcome, they describe how an input would need to change for the model to produce a different prediction. This form of explanation resembles reasoning through alternative scenarios and may therefore offer useful support for understanding model behaviour in healthcare applications [21, 15].

The usefulness of such explanations nevertheless depends on their robustness. Healthcare prediction models may be retrained as new data become available or as modelling pipelines are updated. Even when independently trained predictive models achieve comparable performance on the same task, they may generate different counterfactual recommendations for the same patient. Substantial variation across retrained models would suggest that an explanation reflects the behaviour of a particular model instance rather than a stable characteristic of the prediction task.

This concern motivates the present thesis. Rather than proposing a new counterfactual generation algorithm, this study examines the extent to which counterfactual explanations remain consistent across independently trained predictive models developed for the same healthcare prediction task. By evaluating explanation robustness under model retraining, the thesis aims to clarify an important aspect of counterfactual explanation reliability in healthcare.

1.2 Research Gap and Thesis Objectives

Over the past decade, substantial progress has been made in the development of counterfactual explanation methods. Existing research has proposed numerous algorithms that improve properties such as validity, sparsity, diversity, plausibility, and actionability while remaining applicable to complex black-box machine learning models [16, 13]. As a result, the literature now provides a wide range of techniques capable of generating high-quality counterfactual explanations for individual predictions.

Despite these advances, an important limitation remains. The vast majority of existing studies evaluate counterfactual explanations with respect to a single trained predictive model. While

previous studies have explored robustness from different perspectives, robustness across independently retrained predictive models remains comparatively underexplored. In particular, relatively little attention has been given to evaluating complementary aspects of robustness—including explanation validity, feature selection, direction of change, and intervention magnitude—within a unified evaluation framework. In this setting, explanation quality is typically assessed using metrics such as validity, sparsity, proximity, diversity, or plausibility, while implicitly assuming that the underlying predictive model remains fixed throughout the evaluation process [8, 15]. However, this assumption is often unrealistic in real-world machine learning systems.

In practical healthcare applications, predictive models are frequently retrained as additional patient data become available, preprocessing pipelines evolve, or model hyperparameters are updated. Although these retrained models may achieve nearly identical predictive performance on the same clinical task, they do not necessarily learn identical decision boundaries. Consequently, two equally accurate models may produce substantially different counterfactual explanations for the same patient. From a clinical perspective, such variability raises important concerns regarding the reliability and trustworthiness of counterfactual explanations, particularly when they are intended to support decision-making in high-stakes environments.

While robustness has received increasing attention within explainable artificial intelligence, relatively little work has systematically investigated whether counterfactual explanations remain stable across independently trained models solving the same prediction task. In particular, limited attention has been given to whether different models consistently identify the same actionable features, recommend changes in the same direction, or suggest modifications of comparable magnitude for identical patients. These properties are essential if counterfactual explanations are to be regarded as reliable decision-support tools rather than model-specific artifacts.

This thesis addresses this gap by evaluating the robustness of counterfactual explanations under model retraining using a real-world healthcare prediction task. Counterfactual explanations are generated using the DiCE framework for multiple independently trained predictive models, after which their consistency is assessed across four complementary robustness dimensions: cross-model validity, feature-set consistency, direction consistency, and magnitude consistency. Together, these dimensions provide a comprehensive evaluation of whether counterfactual explanations remain reliable when the underlying predictive model changes despite performing the same clinical prediction task. Accordingly, the primary research question investigated in this thesis is:

> *To what extent do counterfactual explanations remain reliable across independently trained machine learning models for the same healthcare prediction task?*

1.3 Thesis Contributions

This thesis investigates the reliability of counterfactual explanations beyond the conventional setting of a single trained predictive model. Rather than proposing a new counterfactual generation algorithm, the primary objective is to evaluate whether counterfactual explanations remain reliable when independently trained machine learning models are developed for the same healthcare prediction task.

To achieve this objective, this thesis makes four main contributions. First, it provides an empirical evaluation of the reliability of counterfactual explanations across independently trained predictive models using a real-world healthcare prediction task. Second, it introduces a multidimensional

robustness evaluation framework based on four complementary robustness dimensions, namely cross-model validity, feature-set consistency, direction consistency, and magnitude consistency. Third, it develops a multi-patient experimental protocol in which counterfactual explanations are generated for a cohort of eligible patients across multiple independently trained models, enabling systematic cross-model comparisons. Finally, the reported robustness estimates are complemented with patient-level bootstrap confidence intervals, providing a statistical characterization of the uncertainty associated with the experimental findings.

1.4 Thesis Organization

This bachelor thesis was conducted within the Data Science and Artificial Intelligence programme at the Leiden Institute of Advanced Computer Science (LIACS), under the supervision of Hazel Doughty.

The remainder of this thesis is organized as follows. Chapter 2 presents the theoretical background on explainable artificial intelligence, counterfactual explanations, robustness in explainable AI, and existing evaluation metrics for counterfactual explanations. Chapter 3 describes the proposed methodology, including the predictive modelling pipeline, the counterfactual generation process, the robustness evaluation framework, and the statistical analysis. Chapter 4 presents the experimental results. Chapter 5 reports the patient-level bootstrap confidence intervals used to quantify the statistical uncertainty of the robustness estimates. Chapter 6 discusses the findings in relation to the existing literature, interprets the proposed robustness dimensions, highlights the practical implications and limitations of the study, and outlines directions for future research. Finally, Chapter 7 concludes the thesis by summarizing its main findings and contributions.

2 Background

2.1 Explainable Artificial Intelligence

Recent advances in machine learning have led to remarkable improvements in predictive performance across a wide range of domains, including computer vision, natural language processing, finance, and healthcare. In particular, complex machine learning models such as ensemble methods and deep neural networks have demonstrated superior predictive accuracy compared to traditional statistical approaches by learning highly nonlinear relationships directly from data [17, 15]. Despite these advances, the increasing complexity of modern predictive models has made their internal decision-making processes difficult to understand, giving rise to the well-known "black-box" problem.

In many real-world applications, predictive accuracy alone is insufficient. Users often require an explanation of *why* a model reached a particular decision before that prediction can be trusted or acted upon. This requirement is particularly critical in high-stakes domains such as healthcare, where machine learning systems may support decisions related to diagnosis, treatment planning, patient risk assessment, or hospital resource allocation. In these settings, opaque predictions without meaningful explanations can reduce clinician trust, hinder clinical adoption, and make it difficult to identify erroneous or biased model behaviour [10, 14].

The growing demand for transparent and trustworthy artificial intelligence has motivated the development of Explainable Artificial Intelligence (XAI), an active research area that seeks

to make machine learning models more understandable to human users. Rather than treating predictive models as inaccessible black boxes, XAI aims to provide explanations that allow users to understand the factors influencing a prediction, evaluate whether the model's reasoning is plausible, and determine whether the prediction should be trusted in practice [9, 15]. Depending on the application, these explanations may support model debugging, fairness assessment, regulatory compliance, or informed decision-making.

Existing XAI approaches can broadly be categorized into intrinsic interpretability and post-hoc explanation methods. Intrinsically interpretable models, such as linear regression, decision trees, or rule-based systems, are designed to remain transparent by construction, allowing their decision processes to be understood directly. In contrast, post-hoc explanation techniques seek to explain already-trained black-box models without modifying their internal structure. Common examples include feature attribution methods such as SHAP and LIME, example-based explanations, saliency methods, and counterfactual explanations [15, 2]. These approaches enable practitioners to retain the predictive performance of complex machine learning models while providing additional insight into individual predictions.

Healthcare has emerged as one of the most important application areas for explainable machine learning. Clinical decision support systems are increasingly used to predict disease progression, patient deterioration, hospital readmission, and treatment outcomes using electronic health records and other heterogeneous medical data. Because these predictions may directly influence patient care, clinicians must be able to understand not only the predicted outcome but also the reasoning behind it. Consequently, explainability has become an essential requirement for trustworthy medical AI rather than merely a desirable model characteristic [10, 19].

Although considerable progress has been made in developing explainability techniques, important challenges remain. Many existing explanation methods focus primarily on producing plausible explanations for a single trained model while paying comparatively little attention to whether those explanations remain reliable when the underlying predictive model changes. Small variations in the training process, such as different random seeds on the same dataset, may alter the resulting explanations even when predictive performance remains nearly identical. This raises an important question regarding the reliability of explanations themselves, motivating further investigation into the robustness of explanation methods. The following section therefore introduces counterfactual explanations, one of the most widely adopted post-hoc explanation techniques for tabular healthcare machine learning models.

2.2 Counterfactual Explanations

Among the various post-hoc explainability techniques proposed in recent years, counterfactual explanations have emerged as one of the most intuitive and actionable approaches for explaining individual predictions. Rather than identifying which input features contributed most to a prediction, counterfactual explanations answer a simple hypothetical question: What is the smallest change to the input that would have altered the model's prediction [21, 16]. This formulation closely resembles the way humans naturally reason about alternative outcomes, making counterfactual explanations particularly accessible to non-technical users.

A counterfactual explanation consists of an alternative version of an input instance that is minimally modified while changing the predicted class. For example, if a machine learning model predicts that a patient is at high risk of hospital readmission, a counterfactual explanation may

suggest that reducing the number of previous inpatient visits or shortening the expected length of hospital stay would have changed the prediction to a lower-risk outcome. By explicitly describing how an outcome could be different, counterfactual explanations provide users with concrete and interpretable information about the model's local decision boundary rather than simply indicating which variables were influential.

Compared with feature attribution methods such as SHAP or LIME, counterfactual explanations offer several practical advantages. Feature attribution methods quantify the contribution of individual variables to a prediction but generally do not indicate how those variables should change to obtain a different outcome. Counterfactual explanations, in contrast, are inherently action-oriented because they identify specific feature modifications associated with an alternative prediction. This property has made counterfactual explanations particularly attractive for decision support applications in domains where actionable recommendations are more valuable than descriptive feature importance alone [20, 15].

Healthcare represents one of the most promising application areas for counterfactual explanations. Machine learning models are increasingly used to estimate patient risks such as mortality, disease progression, treatment response, and hospital readmission. While accurate predictions are valuable, clinicians often require additional insight into whether a patient's predicted risk could realistically change under different clinical circumstances. Counterfactual explanations naturally address this requirement by describing hypothetical but clinically meaningful scenarios under which the prediction would differ. Consequently, they have attracted growing attention as an explanation technique capable of supporting clinical reasoning while remaining relatively easy for healthcare professionals to interpret [13].

Generating useful counterfactual explanations, however, involves several practical challenges. Ideally, counterfactuals should satisfy multiple desirable properties simultaneously. They should remain close to the original instance, modify only a small number of features (sparsity), correspond to realistic data points (plausibility), respect immutable patient characteristics such as age or biological sex (actionability), and produce valid predictions according to the underlying machine learning model [16, 13]. Balancing these often competing objectives remains an active area of research, and numerous optimization algorithms have been proposed to generate clinically meaningful counterfactual explanations.

Among the available counterfactual generation frameworks, DiCE (Diverse Counterfactual Explanations) has become one of the most widely adopted methods for tabular machine learning applications because it can efficiently generate multiple diverse counterfactuals for an individual prediction while remaining model-agnostic [16]. Owing to its flexibility and widespread adoption in the literature, DiCE serves as the counterfactual generation framework employed throughout this thesis.

2.3 Robustness in Explainable Artificial Intelligence

As explainable artificial intelligence (XAI) has become increasingly important for the deployment of machine learning systems in high-stakes domains, attention has gradually shifted from generating explanations to evaluating their quality and reliability. While early research primarily focused on developing explanation methods capable of interpreting complex predictive models, more recent work has emphasized that explanations themselves should satisfy desirable properties beyond interpretability alone. Among these properties, robustness has emerged as one of the most important

characteristics of trustworthy explanations [2, 15].

In the context of explainable AI, robustness refers to the stability of explanations under small changes that do not substantially alter the underlying prediction task. Intuitively, if two predictive models exhibit similar predictive performance or produce the same prediction for a particular instance, their corresponding explanations should not differ dramatically. Explanations that vary considerably under minor perturbations may reduce user confidence and raise concerns regarding whether the explanation reflects meaningful model behaviour or merely artifacts of the learning process [1, 14].

The notion of robustness in XAI differs from robustness in predictive machine learning. Traditional machine learning robustness typically concerns the stability of model predictions under adversarial attacks, noisy inputs, or distributional shifts. In contrast, explanation robustness focuses on the consistency of the explanations themselves. Consequently, even when predictive performance remains essentially unchanged, explanations may still vary substantially if the underlying predictive model changes. This distinction is particularly important in applications where explanations, rather than predictions alone, are intended to support human decision-making.

Several studies have demonstrated that popular explanation methods may exhibit considerable instability. Small perturbations to the training data, random initialization, or model parameters can lead to noticeably different explanations despite producing nearly identical predictive accuracy [1, 7]. Such findings suggest that evaluating explanation quality requires considering not only traditional metrics such as interpretability or fidelity, but also the robustness of the generated explanations.

The increasing recognition of explanation robustness has motivated growing interest in developing systematic evaluation methodologies for explainable AI. Rather than assuming that explanations generated by a single trained model are inherently reliable, recent research has highlighted the importance of assessing whether explanations remain consistent under realistic variations in the underlying learning process. This perspective provides the conceptual foundation for the robustness evaluation performed in this thesis.

2.4 Evaluation Metrics for Counterfactual Explanations

The quality of a counterfactual explanation is commonly assessed using a number of quantitative evaluation metrics that capture different desirable properties. Since no single metric can fully characterize explanation quality, existing studies typically evaluate counterfactual explanations using multiple complementary criteria [16, 13].

One of the most fundamental evaluation criteria is *validity*, which measures whether the generated counterfactual successfully changes the prediction of the underlying machine learning model to the desired target class. A counterfactual that fails to alter the model prediction cannot be considered a successful explanation, regardless of its other characteristics. Consequently, validity is generally regarded as the minimum requirement that every counterfactual explanation should satisfy.

Among valid counterfactuals, researchers generally seek explanations that remain as close as possible to the original instance. This property is commonly referred to as *proximity* and is typically measured using a distance metric between the original observation and the generated counterfactual. Closely related to proximity is *sparsity*, which encourages modifying as few input features as possible. Sparse counterfactuals are generally considered easier for users to understand and more

realistic to implement in practice because they require fewer simultaneous changes [21, 16].

When multiple counterfactual explanations are generated for the same instance, *diversity* becomes another desirable property. Rather than producing nearly identical explanations, diverse counterfactuals provide multiple alternative pathways that could achieve the desired prediction, allowing users to consider different actionable scenarios. In practical decision-support applications, diversity may increase the usefulness of counterfactual explanations by offering several feasible options rather than a single recommended modification [16].

Beyond optimization-based metrics, recent research has increasingly emphasized the importance of *plausibility* and *actionability*. Plausibility refers to whether generated counterfactuals resemble realistic observations drawn from the underlying data distribution, while actionability requires suggested feature changes to be feasible in practice and to respect immutable characteristics such as age or biological sex. These considerations are particularly important in healthcare applications, where unrealistic or clinically impossible recommendations could reduce the practical value of counterfactual explanations [13].

Although these evaluation metrics have become well established in the counterfactual explanation literature, they share an important limitation. Existing metrics primarily evaluate the quality of counterfactual explanations generated for a single trained predictive model. They do not assess whether equally accurate models trained for the same prediction task produce consistent explanations for identical instances. Consequently, while existing metrics provide valuable information about the quality of individual counterfactual explanations, they offer limited insight into the reliability of explanations under model retraining. Addressing this limitation forms the central methodological focus of the present thesis.

3 Methodology

3.1 Dataset

This study uses the Diabetes 130-US Hospitals dataset, a large real-world electronic health record (EHR) dataset introduced by Strack et al. [18]. The dataset contains records collected from 130 hospitals in the United States between 1999 and 2008 and has become one of the most widely used benchmark datasets for hospital readmission prediction. It includes demographic information, admission details, laboratory measurements, diagnoses, and diabetes-related medication records for hospitalized patients.

The objective of this thesis is to predict whether a patient will be readmitted within 30 days following hospital discharge. Following the preprocessing procedure adopted in this study, the original readmission outcome is converted into a binary classification problem, where patients readmitted within 30 days constitute the positive class and all remaining patients constitute the negative class. Predicting short-term hospital readmission represents an important clinical prediction task because early readmissions are associated with increased healthcare costs, reduced quality of care, and poorer patient outcomes.

The Diabetes 130-US Hospitals dataset was selected for three primary reasons. First, hospital readmission prediction is a clinically relevant problem that has received considerable attention within healthcare machine learning. Second, the dataset contains a rich combination of demographic, clinical, and treatment-related variables, providing a heterogeneous feature space in which

constrained counterfactual explanations can be evaluated. Finally, its widespread use in previous healthcare machine learning and explainable artificial intelligence studies facilitates comparison with existing literature.

After preprocessing, the final dataset consists of both numerical and categorical variables describing patient demographics, healthcare utilization, laboratory procedures, prescribed medications, and grouped diagnostic information. These variables form the feature space used throughout the predictive modelling and counterfactual explanation experiments presented in this thesis.

To ensure an unbiased evaluation, the processed dataset was divided into training and test partitions using an 80–20 stratified split, preserving the original class distribution. The training partition was used to construct the predictive models, whereas the held-out test partition was reserved exclusively for counterfactual generation and robustness evaluation. Categorical variables were represented using one-hot encoding prior to model training. Unlike ordinal encoding, one-hot encoding avoids introducing artificial ordering relationships between nominal categories and is therefore more appropriate for tree-based models trained on heterogeneous healthcare data.

3.2 Predictive Modelling Pipeline

Before generating counterfactual explanations, a predictive model capable of estimating the probability of short-term hospital readmission must first be constructed. Since the objective of this thesis is to evaluate the robustness of counterfactual explanations under model retraining rather than to compare different predictive algorithms, a single modelling approach is adopted throughout all experiments.

Gradient-boosted decision trees implemented through XGBoost were selected as the predictive modelling framework. XGBoost provides a scalable tree-boosting implementation capable of modelling nonlinear relationships in tabular data [5]. In the model comparison conducted in this study, XGBoost achieved the highest balanced accuracy and ROC–AUC and was therefore selected for the subsequent counterfactual experiments.

The processed dataset was first divided into training and test partitions using the stratified split described in the previous section. All predictive models were trained exclusively on the training partition, while the test partition remained unseen throughout model development and was reserved solely for counterfactual generation and robustness evaluation.

To evaluate whether counterfactual explanations remain reliable under model retraining, multiple predictive models were trained independently using identical training data and modelling procedures. The only source of variation between models was the random seed controlling the stochastic elements of the training process. Consequently, all models solved exactly the same prediction task using the same dataset, feature representation, and hyperparameter configuration while differing only through random initialization. This experimental design enables any observed variation in counterfactual explanations to be attributed to model retraining rather than differences in data or modelling choices.

Although the dataset exhibits class imbalance, synthetic oversampling techniques such as SMOTE [4] were intentionally not applied. Since the objective of this thesis is to evaluate the robustness of counterfactual explanations rather than maximize predictive performance, generating synthetic patient records could complicate the interpretation of explanation robustness. Instead, the original data distribution was preserved throughout the experiments. After training, predictive performance was evaluated on the held-out test set to verify that all independently trained

models achieved comparable classification performance before robustness analysis was performed. Counterfactual explanations were subsequently generated only for patients that were consistently classified as belonging to the positive class by every predictive model, ensuring that all models explained identical prediction scenarios throughout the robustness evaluation.

3.3 Counterfactual Generation

Counterfactual explanations were generated using the Diverse Counterfactual Explanations (DiCE) framework proposed by Mothilal et al. [16]. DiCE was selected because it is a model-agnostic counterfactual generation framework that has been widely adopted in the explainable artificial intelligence literature for tabular machine learning problems. In addition to generating valid counterfactual explanations, DiCE is capable of producing multiple diverse alternatives for a single prediction, making it particularly suitable for evaluating explanation robustness.

For each predictive model, counterfactual explanations were generated independently using the same generation procedure and parameter configuration. Maintaining an identical generation process across all predictive models ensures that any observed differences in the resulting explanations originate from differences in the trained models rather than variations introduced during counterfactual generation.

Since the objective of this study is to evaluate explanation robustness under model retraining, counterfactual explanations were generated only for patients that satisfied two conditions. First, the patient had to belong to the positive class in the held-out test set. Second, all independently trained predictive models had to correctly classify that patient as positive. Restricting the analysis to patients that satisfy both conditions guarantees that all models explain the same clinical prediction scenario, thereby enabling a meaningful comparison of the generated counterfactual explanations. This restriction ensures that the generated counterfactual explanations explain successful model decisions rather than prediction errors. Since counterfactual explanations are intended to explain model behaviour rather than ground-truth labels, evaluating explanations for misclassified instances would confound explanation robustness with predictive errors.

Among the eligible patients, a cohort of 50 individuals was randomly selected for robustness evaluation. Rather than treating individual counterfactual explanations as independent observations, the present study performs robustness evaluation at the patient level. For each selected patient, counterfactual explanations are generated across multiple independently trained predictive models, resulting in several hundred counterfactual explanations and thousands of pairwise robustness comparisons. Consequently, the selected cohort provides a computationally feasible yet sufficiently diverse evaluation set, while the large number of cross-model counterfactual comparisons and patient-level bootstrap analysis enable statistically meaningful robustness estimates. Although the evaluation cohort consisted of 50 patients, the experimental design produced 750 counterfactual explanations and several thousand cross-model robustness comparisons, providing a sufficiently rich basis for the subsequent robustness analysis. Furthermore, statistical uncertainty is quantified through patient-level bootstrap confidence intervals, allowing the robustness estimates to be interpreted together with their associated variability. Notably, patient eligibility depended solely on satisfying the predefined inclusion criteria and successful counterfactual generation under all models, rather than on the robustness scores ultimately obtained. Consequently, the evaluation cohort was selected independently of the reported experimental outcomes.

Counterfactual generation was restricted to four permitted-to-vary care-process variables: time

in hospital, number of laboratory procedures, number of procedures, and number of medications. All remaining predictors, including demographic characteristics, diagnoses, admission information, and medication-status variables, were held fixed throughout the optimization process. For each permitted feature, the allowable range was derived from the training data using the 1st and 99th percentile values (rounded outward to the nearest integer), preventing unrealistic feature modifications during optimization.

For every eligible patient and each independently trained predictive model, three counterfactual explanations were generated. Producing multiple candidate explanations rather than a single alternative allows DiCE to capture different feasible pathways towards the desired prediction and reduces the likelihood that the robustness evaluation depends on a single optimization outcome. Throughout the subsequent robustness analysis, comparisons were performed using the generated counterfactual explanations corresponding to identical patients across independently trained predictive models.

3.4 Robustness Evaluation Framework

Conventional counterfactual evaluation typically examines whether an explanation is valid, sparse, proximal, plausible, and actionable with respect to a single fixed predictive model. Although these properties remain important, they do not establish whether the explanation would remain reliable if the predictive model were independently retrained. Recent work on robust counterfactual explanations has therefore emphasized the need to evaluate counterfactual validity and stability under plausible changes to the underlying model [3, 11, 12].

This thesis evaluates robustness by generating counterfactual explanations generated by five independently trained XGBoost models trained using identical data, preprocessing, hyperparameters, and learning procedures but different random seeds. For every selected patient, each model generates three counterfactual explanations under the same actionable-feature and permitted-range constraints. The resulting explanations are then evaluated across four complementary dimensions: cross-model validity, feature-set consistency, direction consistency, and magnitude consistency.

These dimensions capture distinct aspects of explanation reliability. Cross-model validity measures whether a counterfactual generated from one model produces the desired prediction across the full set of independently trained models, including the model from which it was generated. Feature-set consistency examines whether independently trained models rely on similar actionable variables. Direction consistency evaluates whether explanations that use the same feature recommend changing it in the same direction. Magnitude consistency evaluates whether same-direction recommendations have comparable normalized sizes.

The four dimensions are reported separately rather than combined into a single composite score. This avoids imposing unsupported weights on conceptually different properties and allows cases in which validity and structural consistency behave differently to remain visible. For example, a counterfactual may remain valid across models even when the models disagree about which features should be modified.

3.4.1 Cross-Model Validity

Let $M = \{m_1, \ldots, m_K\}$ denote the set of K independently trained predictive models, and let $x'_{i,s,c}$ denote counterfactual c generated for patient i using source model m_s. In the present experiment,

$K = 5$ and three counterfactuals are generated by each source model.

Each generated counterfactual is evaluated using every model in M, including its source model. For the desired non-readmission class $y^* = 0$, the cross-model robust validity of $x'_{i,s,c}$ is defined as

$$V\left(x'_{i,s,c}\right) = \frac{1}{K} \sum_{k=1}^{K} I\left[m_k\left(x'_{i,s,c}\right) = y^*\right], \tag{1}$$

where $I[\cdot]$ is the indicator function. A value of 1 indicates that the counterfactual remains valid under all independently trained models, whereas a value of $1/K$ indicates validity only under one model. This metric operationalizes the central robustness question of whether a counterfactual remains capable of changing the predicted outcome after model retraining [11, 12].

Patient-level cross-model validity is calculated by averaging the robust validity scores of all counterfactual explanations generated for the patient. Cohort-level results are subsequently summarized using the distribution of patient-level estimates.

3.4.2 Feature-Set Consistency

Cross-model validity does not determine whether independently trained models provide structurally similar explanations. Two counterfactuals may both produce the desired outcome while recommending changes to entirely different variables. Feature-set consistency is therefore used to quantify agreement in the actionable features modified by explanations.

For a patient i and counterfactual explanation x', let

$$F_i(x') = \left\{ f \in \mathcal{A} \;\middle|\; x'_f \neq x_{i,f} \right\}, \tag{2}$$

where $\mathcal{A}$ is the set of actionable features and x_i is the original patient instance. For two counterfactuals x'_a and x'_b generated for the same patient by different source models, feature-set similarity is measured using the Jaccard coefficient:

$$J(x'_a, x'_b) = \frac{|F_i(x'_a) \cap F_i(x'_b)|}{|F_i(x'_a) \cup F_i(x'_b)|}. \tag{3}$$

If neither explanation modifies any permitted feature, the similarity is defined as 1; this edge case did not affect the reported analysis. A value of 1 indicates that both explanations modify identical actionable feature sets, whereas a value of 0 indicates no overlap.

Counterfactual identifiers such as *CF 1*, *CF 2*, and *CF 3* do not represent aligned explanations across independently trained models. Consequently, the first counterfactual generated by one model is not assumed to correspond to the first counterfactual generated by another. For each patient and each pair of source models, all 3×3 cross-model counterfactual combinations are therefore compared. With five models, this produces

$$52 \times 3 \times 3 = 90 \tag{4}$$

feature-set comparisons per patient. Patient-level feature-set consistency is calculated as the mean of these pairwise Jaccard similarities.

3.4.3 Direction Consistency

When independently generated explanations modify the same actionable feature, they may nevertheless disagree about whether that feature should increase or decrease. Direction consistency measures the extent to which observed recommendations for a given patient–feature combination agree in sign.

For patient i, feature f, and counterfactual x', the direction of change is defined as

$$d_{i,f}(x') = \{+1, if x'_f - x_{i,f} > 0, -1, if x'_f - x_{i,f} < 0. \tag{5}$$

Counterfactuals in which feature f remains unchanged are excluded from the direction analysis for that feature. Let $n^+_{i,f}$ and $n^-_{i,f}$ denote the numbers of observed increases and decreases, respectively. Direction consistency is defined as the proportion of observed changes following the dominant direction:

$$D_{i,f} = \frac{\max\left(n^+_{i,f}, n^-_{i,f}\right)}{n^+_{i,f} + n^-_{i,f}}. \tag{6}$$

The metric takes values greater than or equal to 0.5, with values near 0.5 indicating substantial disagreement and a value of 1 indicating complete directional agreement. If the numbers of increases and decreases are equal, the dominant direction is recorded as a tie.

Because the purpose of this metric is to assess agreement across independently trained models, a patient–feature combination is included only when the feature is modified by counterfactuals originating from at least two distinct source models. Features used by only one model cannot provide evidence of cross-model directional consistency and are therefore excluded. Patient-level direction consistency is obtained by averaging $D_{i,f}$ across all eligible actionable features for that patient.

3.4.4 Magnitude Consistency

Direction agreement alone does not establish that explanations recommend quantitatively similar interventions. Two models may agree that a feature should increase while suggesting substantially different amounts of change. Magnitude consistency therefore evaluates agreement in the normalized size of same-direction feature modifications.

Because actionable variables have different numerical scales, each absolute change is normalized using the permitted range of its corresponding feature. For feature f, the normalized change magnitude is defined as

$$\Delta_{i,f}(x') = \frac{\left|x'_f - x_{i,f}\right|}{u_f - l_f}, \tag{7}$$

where l_f and u_f are the lower and upper permitted bounds of feature f, derived from the training distribution.

For two counterfactuals x'_a and x'_b generated for the same patient by different models, modifying the same feature in the same direction, magnitude consistency is defined as

$$G_{i,f}(x'_a, x'_b) = \max\left\{0,\, 1 - \left|\Delta_{i,f}(x'_a) - \Delta_{i,f}(x'_b)\right|\right\}. \tag{8}$$

The score is bounded to the interval $[0, 1]$, with negative values truncated to zero. A value of 1 indicates identical normalized change magnitudes, while lower values indicate greater quantitative disagreement. Counterfactual pairs that modify the feature in opposite directions are excluded because such disagreements are already captured by the direction-consistency metric.

As in the feature-set analysis, counterfactual positions are not assumed to be aligned across models. For each patient, feature, direction, and pair of source models, all available cross-model counterfactual combinations are evaluated. Patient–feature magnitude consistency is calculated by averaging these pairwise scores, after which patient-level magnitude consistency is obtained by averaging across the eligible features of each patient.

3.5 Statistical Analysis

Descriptive robustness summaries were first calculated at the natural evaluation-unit level of each metric. Cross-model validity was summarized across generated counterfactual explanations, feature-set consistency across cross-model counterfactual pairs, direction consistency across eligible patient–feature combinations, and magnitude consistency across eligible same-direction cross-model comparisons.

For statistical uncertainty estimation, each robustness metric was additionally aggregated to one mean score per patient. These patient-level scores ensured that every patient contributed equally to the bootstrap analysis, regardless of the number of eligible feature-level or pairwise comparisons associated with that patient. Cohort-level patient-level means were then computed by averaging the corresponding patient-level scores.[6]

To quantify the statistical uncertainty associated with these estimates, non-parametric bootstrap resampling was performed at the patient level. Bootstrap sampling was chosen because the distribution of the robustness metrics is unknown and no assumptions regarding normality are required. In each bootstrap iteration, patients were sampled with replacement from the evaluation cohort, after which the cohort-level mean robustness score was recomputed. This procedure was repeated 5,000 times for each robustness metric.

The resulting bootstrap distributions were used to estimate 95% percentile confidence intervals for all reported robustness metrics. Sampling at the patient level preserves the dependence structure among multiple counterfactual explanations generated for the same patient and therefore provides a more appropriate measure of uncertainty than treating individual counterfactual explanations as independent observations.

Throughout the experimental evaluation, robustness results are reported as patient-level mean estimates together with their corresponding bootstrap confidence intervals.

4 Results

4.1 Predictive Model Performance

Four supervised classification models were evaluated for the prediction of 30-day hospital readmission: Logistic Regression, Random Forest, Tuned Random Forest, and XGBoost. Model performance was assessed on the held-out test set using accuracy, balanced accuracy, precision, recall, F1-score, and ROC–AUC.

Table 1 summarizes the predictive performance of all evaluated models.

Table 1: Predictive performance of the evaluated classification models.

Model	Accuracy	Balanced Accuracy	Precision	Recall	F1-score	ROC–AUC
Logistic Regression	0.6686	0.6258	0.1834	0.5707	0.2776	0.6801
Random Forest	0.8878	0.5232	0.4766	0.0537	0.0966	0.6754
Tuned Random Forest	0.7683	0.6172	0.2199	0.4227	0.2893	0.6873
XGBoost	0.6691	0.6355	0.1880	0.5923	0.2854	0.6890

The XGBoost classifier achieved the highest ROC–AUC (0.6890) and balanced accuracy (0.6355) among the evaluated models. Consequently, XGBoost was selected as the predictive model for all subsequent counterfactual generation experiments. The robustness evaluation presented in the remainder of this thesis was then performed by retraining this model architecture multiple times using different random seeds while keeping all other experimental settings fixed.

4.2 Counterfactual Generation

Counterfactual explanations were generated for a subset of correctly classified positive patients from the held-out test set. The objective was to construct an evaluation cohort for the subsequent robustness analysis while ensuring that counterfactual explanations could be successfully generated for every selected patient.

Among the test instances, 1,205 patients were correctly classified as requiring readmission (true positives). The common true-positive population was randomly ordered using a fixed random seed. Patients were then screened sequentially for complete counterfactual generation. A patient was included only if each of the five independently trained predictive models successfully generated all three requested counterfactual explanations under the shared generation constraints. The first 50 patients satisfying this criterion formed the final evaluation cohort. In total, 121 candidate patients were screened to construct the cohort. After applying the counterfactual generation eligibility criteria described in Chapter 3, 50 patients remained for the robustness evaluation.

For each patient, counterfactual explanations were generated independently using five retrained XGBoost models initialized with different random seeds. Three counterfactual explanations were produced for each model, yielding a total of fifteen counterfactual explanations per patient.

Overall, the experimental pipeline generated 750 counterfactual explanations across the 50-patient evaluation cohort. These explanations formed the basis for all subsequent cross-model robustness analyses.

Table 2: Summary of the counterfactual generation pipeline.

Quantity	Value
Correctly classified true-positive patients	1205
Candidate patients screened	121
Eligible patients selected	50
Retrained XGBoost models	5
Counterfactuals per model	3
Counterfactuals per patient	15
Total counterfactual explanations	750

These counterfactual explanations were subsequently compared across independently trained models to evaluate the robustness of their predictions, selected features, directions of change, and required intervention magnitudes.

4.3 Robustness Evaluation

The following sections present the experimental results for cross-model validity, feature-set consistency, direction consistency, and magnitude consistency. The robustness of counterfactual explanations was evaluated using the four complementary metrics introduced in Chapter 3. The following subsections report descriptive summaries at the natural evaluation-unit level of each metric: generated counterfactuals for cross-model validity, cross-model counterfactual pairs for feature-set consistency, eligible patient–feature combinations for direction consistency, and eligible same-direction cross-model comparisons for magnitude consistency. Patient-level equally weighted estimates and their bootstrap confidence intervals are reported separately in Chapter 5.

4.3.1 Cross-Model Validity

Cross-model validity evaluates whether a counterfactual explanation generated for one predictive model remains valid when evaluated by the remaining independently trained models performing the same prediction task. Higher values indicate that the desired prediction is preserved across a larger proportion of independently trained models.

Table 3 summarizes the cross-model validity scores obtained for all generated counterfactual explanations.

Table 3: Summary statistics of cross-model validity across generated counterfactual explanations.

Statistic	Value
Mean	0.6475
Median	0.6000
Minimum	0.2000
Maximum	1.0000

Across the 750 generated counterfactual explanations, the mean cross-model validity was 0.6475, with a median of 0.6000. Individual validity scores ranged from 0.2000 to 1.0000, indicating considerable variation in how well explanations generalized across independently trained predictive

models. Overall, 26.53% of generated counterfactual explanations remained valid across all five independently trained models, whereas the remaining explanations exhibited varying degrees of disagreement between models.

4.3.2 Feature-Set Consistency

Feature-set consistency evaluates the extent to which independently trained predictive models rely on the same features when generating counterfactual explanations for the same patient.

Table 4 summarizes the feature-set consistency scores obtained across all pairwise comparisons of generated counterfactual explanations.

Table 4: Summary statistics of feature-set consistency across pairwise counterfactual comparisons.

Statistic	Value
Mean	0.5134
Median	0.3333
Minimum	0.0000
Maximum	1.0000

Across the 4,500 pairwise comparisons of generated counterfactual explanations, the mean feature-set consistency was 0.5134, while the median was 0.3333. Feature-set consistency scores ranged from 0.0000 to 1.0000, indicating substantial variability in the features selected by independently trained predictive models. Although complete agreement was observed for some explanation pairs, other pairs shared no common modified features, demonstrating that equally valid predictive models may identify different feature subsets when producing counterfactual explanations.

4.3.3 Direction Consistency

Direction consistency evaluates whether independently trained predictive models recommend changing the same feature in the same direction when generating counterfactual explanations.

Table 5 summarizes the direction consistency scores obtained across all eligible patient–feature combinations.

Table 5: Summary statistics of direction consistency across eligible patient–feature combinations.

Statistic	Value
Mean	0.8094
Median	0.8000
Minimum	0.5000
Maximum	1.0000

Across the 173 eligible patient–feature combinations, the mean direction consistency was 0.8094, with a median of 0.8000. Direction consistency scores ranged from 0.5000 to 1.0000. These results indicate that, for features modified by multiple independently trained predictive models, the recommended direction of change frequently agreed across models, although complete agreement was not observed in every case.

4.3.4 Magnitude Consistency

Magnitude consistency evaluates the extent to which independently trained predictive models recommend similar amounts of change for the same feature when the recommended direction of change is identical.

Table 6 summarizes the magnitude consistency scores obtained across all eligible pairwise counterfactual comparisons.

Table 6: Summary statistics of magnitude consistency across eligible pairwise counterfactual comparisons.

Statistic	Value
Mean	0.8755
Median	0.9231
Minimum	0.1538
Maximum	1.0000

Across the 3,659 eligible pairwise counterfactual comparisons, the mean magnitude consistency was 0.8755, with a median of 0.9231. Magnitude consistency scores ranged from 0.1538 to 1.0000. These results indicate that, when independently trained predictive models recommended changing the same feature in the same direction, the normalized magnitude of the recommended changes was frequently similar across models.

5 Bootstrap Confidence Intervals

To quantify the statistical uncertainty associated with the robustness estimates, non-parametric bootstrap resampling was performed at the patient level, as described in Chapter 3. For each robustness dimension, 5,000 bootstrap samples were generated by sampling patients with replacement and recomputing the cohort-level patient mean. The resulting percentile-based 95% confidence intervals are reported in Table 7.

Table 7: Patient-level robustness estimates, bootstrap standard errors, and 95% percentile confidence intervals.

Robustness dimension	Patient-level mean	Bootstrap SE	95% CI lower	95% CI upper
Cross-model validity	0.6475	0.0205	0.6059	0.6869
Feature-set consistency	0.5134	0.0155	0.4846	0.5443
Direction consistency	0.8063	0.0106	0.7852	0.8265
Magnitude consistency	0.8450	0.0079	0.8288	0.8599

The bootstrap confidence intervals were relatively narrow for all four robustness metrics, suggesting that the patient-level robustness estimates remained stable across repeated bootstrap resampling. Overall, the relatively narrow bootstrap confidence intervals suggest that the observed differences between the four robustness dimensions are unlikely to be attributable solely to sampling variability, supporting the stability of the reported robustness rankings across the evaluation cohort.

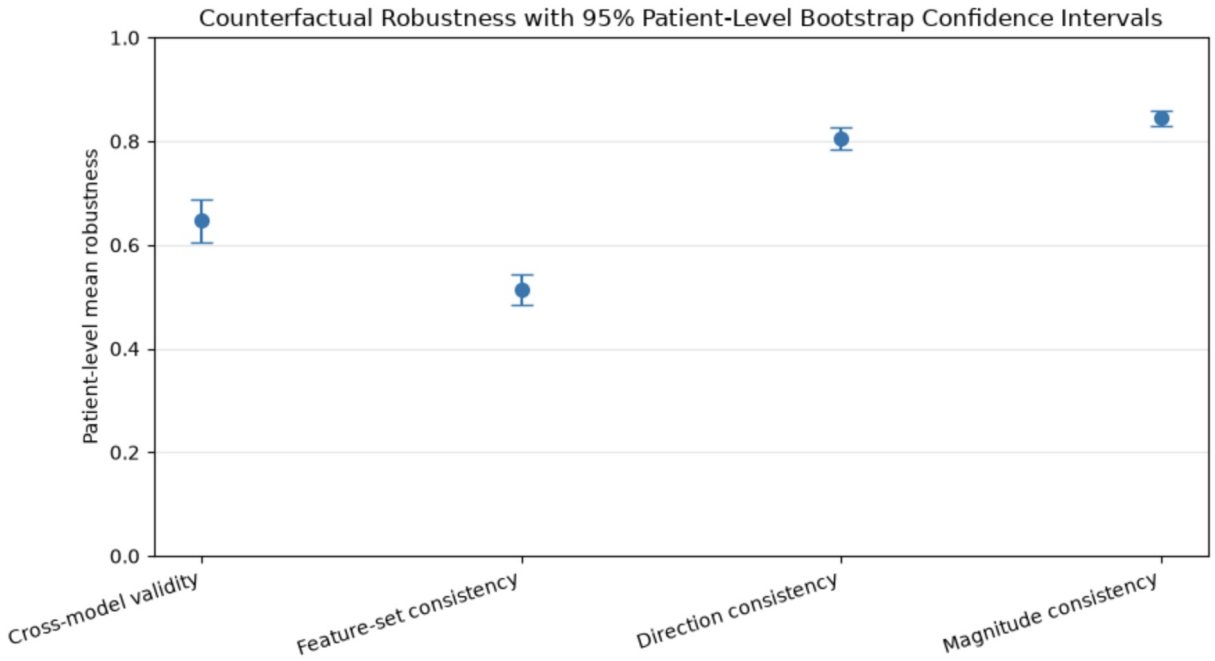

Figure 1: Patient-level bootstrap estimates and 95% confidence intervals for the four robustness metrics.

6 Discussion

6.1 Summary of Main Findings

This thesis investigated the reliability of counterfactual explanations across independently trained machine learning models performing the same healthcare prediction task. Rather than evaluating explanations generated by a single predictive model, the study examined whether independently trained models produced similar counterfactual recommendations for the same patients.

Four complementary robustness dimensions were evaluated: cross-model validity, feature-set consistency, direction consistency, and magnitude consistency. The experimental results revealed a clear pattern. Cross-model validity remained moderate, indicating that counterfactual explanations did not consistently transfer across independently trained models. Feature-set consistency produced the lowest robustness estimates, suggesting considerable variability in the choice of modified features. In contrast, direction consistency and magnitude consistency achieved substantially higher robustness scores, indicating that when different models modified the same features, they generally agreed on both the direction and the relative magnitude of the recommended changes.

Overall, these findings suggest that the robustness of counterfactual explanations depends strongly on the aspect of explanation being evaluated. Although independently trained models frequently disagreed on which features should be modified, they tended to provide similar recommendations once a common feature was selected. This highlights the importance of evaluating counterfactual explanations using multiple complementary robustness dimensions rather than relying on a single robustness metric.

6.2 Interpretation of the Four Robustness Dimensions

The four robustness dimensions evaluated in this study capture complementary aspects of counterfactual explanation reliability. Rather than representing competing evaluation criteria, they provide a multi-dimensional assessment of whether explanations remain consistent across independently trained predictive models.

Cross-model validity produced a moderate robustness score, indicating that counterfactual explanations generated for one predictive model did not consistently achieve the desired prediction when evaluated on other independently trained models. This finding suggests that counterfactual explanations remain partially dependent on the specific predictive model from which they are generated, even when different models are trained for the same prediction task. Such behaviour is consistent with previous observations that counterfactual explanations are influenced by model-specific decision boundaries rather than solely by the underlying data distribution [1, 3].

Among the four robustness dimensions, feature-set consistency produced the lowest robustness estimates. Independently trained predictive models frequently selected different subsets of features when constructing valid counterfactual explanations for the same patient. This observation is unsurprising given that multiple valid recourse strategies often exist for a single prediction problem. Because counterfactual generation is inherently an optimization problem with potentially multiple feasible solutions, different predictive models may identify different combinations of actionable features while still producing successful counterfactual explanations [21, 16].

In contrast, direction consistency remained substantially higher than feature-set consistency. When independently trained predictive models selected the same feature for modification, they generally agreed on whether that feature should increase or decrease. This suggests that although models may disagree about *which* features to modify, they often capture similar local trends in the relationship between individual features and the prediction outcome.

Magnitude consistency achieved the highest robustness estimates across all evaluated dimensions. Once independently trained models agreed on both the selected feature and the direction of change, they also tended to recommend similar normalized amounts of change. This result indicates that disagreement between models primarily arises during feature selection rather than during the estimation of the required modification itself. Consequently, the largest source of variability in counterfactual explanations appears to lie in identifying the feature-modification strategy, whereas the recommended change magnitude remains comparatively stable once a common strategy has been selected.

Taken together, these findings demonstrate that counterfactual robustness cannot be adequately characterized by a single evaluation metric. Different robustness dimensions capture different sources of variability, and considering only one aspect may lead to an incomplete assessment of explanation reliability. A comprehensive robustness evaluation should therefore consider multiple complementary dimensions, particularly in high-stakes applications such as healthcare. Therefore, reporting only a single robustness score may conceal important differences between explanation properties. The proposed framework instead enables robustness to be analysed as a collection of complementary behaviours rather than a single aggregated quantity.

6.3 Comparison with Previous Literature

The findings of this study are broadly consistent with previous research highlighting the model dependence of counterfactual explanations. Alvarez-Melis and Jaakkola demonstrated that interpretability methods can be sensitive to small changes in predictive models, suggesting that explanations may not remain stable under model perturbations [1]. Similarly, Artelt et al. reported that counterfactual explanations can exhibit substantial variability when generated under different conditions, emphasizing the importance of evaluating robustness beyond the success of individual explanations [3].

While previous studies primarily investigated robustness with respect to perturbations of a single predictive model or individual explanation methods, the present work considers a different setting by comparing explanations generated from multiple independently trained models performing the same prediction task. This perspective reflects a realistic deployment scenario in which several predictive models with comparable predictive performance may coexist or be retrained over time.

The results obtained in this study suggest that robustness is not a single property but rather depends on the aspect of explanation being considered. Although feature selection exhibited considerable variability across independently trained models, substantially higher agreement was observed in the direction and magnitude of recommended feature changes once common features were selected. This observation complements previous robustness studies by demonstrating that different robustness dimensions may reveal different forms of explanation stability.

More broadly, these findings support recent survey papers arguing that robustness should be regarded as a multi-dimensional property of counterfactual explanations rather than being assessed through a single evaluation criterion [12]. The experimental framework presented in this thesis follows this perspective by evaluating four complementary robustness dimensions, thereby providing a more comprehensive characterization of explanation reliability across independently trained predictive models.

6.4 Practical Implications

The findings of this study have several practical implications for the deployment and evaluation of counterfactual explanations in healthcare applications. Counterfactual explanations are increasingly proposed as a means of supporting clinical decision-making because they provide actionable recommendations that indicate how a prediction could change. However, the present results suggest that explanations generated by a single predictive model should not automatically be interpreted as universally reliable.

The observed variability in cross-model validity and feature-set consistency indicates that different predictive models trained for the same clinical prediction task may recommend different intervention strategies for the same patient. Consequently, relying on a single counterfactual explanation may provide an incomplete picture of the available recourse options. Evaluating explanations across multiple independently trained models can therefore offer additional insight into which recommendations are consistently supported and which are model-dependent.

At the same time, the relatively high direction and magnitude consistency observed in this study suggests that agreement between independently trained models becomes substantially stronger once common intervention features are identified. This finding indicates that robustness assessments should distinguish between disagreement in feature selection and disagreement in the recommended

modification itself, as these represent different sources of uncertainty.

More broadly, the experimental framework presented in this thesis provides a practical methodology for evaluating counterfactual explanation reliability beyond predictive performance alone. Since independently trained models may achieve comparable predictive accuracy while producing different explanations, robustness evaluation can serve as an additional criterion when selecting predictive models for high-stakes decision-support systems.

6.5 Limitations

Several limitations of the present study should be acknowledged when interpreting the reported findings.

First, the experimental evaluation was conducted using a single real-world healthcare dataset, namely the Diabetes 130-US Hospitals dataset. Although this dataset represents a widely used benchmark for medical machine learning research, the observed robustness characteristics may differ for other clinical prediction tasks, patient populations, or data distributions. Consequently, the generalizability of the reported robustness estimates should be confirmed using additional healthcare datasets.

Second, the study focused exclusively on counterfactual explanations generated using the DiCE framework. While DiCE is one of the most widely adopted methods for generating diverse counterfactual explanations, alternative generation algorithms may exhibit different robustness characteristics. The conclusions presented in this thesis should therefore be interpreted as being specific to the explanation method evaluated.

Third, all experiments were performed using independently trained XGBoost predictive models. Although the models differed through stochastic training variation induced by different random seeds, the study did not investigate whether similar robustness patterns hold across fundamentally different model architectures, such as neural networks or logistic regression models.

Finally, this thesis evaluated robustness exclusively from the perspective of independently trained predictive models performing the same prediction task. Other robustness settings, such as robustness to distribution shifts, adversarial perturbations, measurement noise, and temporal dataset drift, were beyond the scope of the present work. In particular, meaningful perturbation-based analyses were not considered because the Diabetes 130-US Hospitals dataset contains a large number of categorical and clinically constrained variables, making the design of realistic perturbations substantially more challenging than for purely numerical datasets. Consequently, the reported robustness dimensions should be viewed as one component of a broader robustness evaluation framework for counterfactual explanations. In addition, the present study evaluates robustness using a single random train-test split. Although multiple independently trained models were considered, future work could investigate whether similar robustness patterns persist across different data partitions.

6.6 Future Work

The findings of this thesis open several promising directions for future research on the robustness of counterfactual explanations, extending beyond the specific experimental setting considered in this study.

An immediate extension of the present work is the evaluation of the proposed robustness framework using additional counterfactual generation algorithms beyond DiCE, such as the method proposed by Wachter et al. [21]. Similarly, investigating robustness across different predictive model families, including neural networks, logistic regression, and alternative ensemble methods, would provide a broader understanding of how explanation robustness depends on both the explanation algorithm and the underlying predictive model.

Another important research direction is the evaluation of robustness under additional sources of uncertainty. While this thesis focused on independently retrained predictive models, future studies could investigate robustness under clinically meaningful distribution shifts, temporal dataset drift, and realistic perturbation scenarios. Such analyses would complement the model-retraining perspective adopted in this work and contribute to a more comprehensive understanding of explanation reliability in real-world healthcare environments. Recent surveys have emphasized that robustness remains one of the central open challenges in counterfactual explanation research [12].

Beyond extending the evaluation framework, future work may explore robustness-aware counterfactual generation. Existing counterfactual generation algorithms primarily optimize objectives such as validity, proximity, sparsity, and diversity [21, 16]. An interesting research direction is to incorporate robustness directly into the optimization objective, encouraging the generation of explanations that remain stable across independently trained predictive models rather than evaluating robustness only after explanations have been generated.

A particularly promising direction is the development of consensus counterfactual explanations. Instead of relying on the explanation produced by a single predictive model, future systems could aggregate recommendations from multiple independently trained models to identify intervention strategies that are consistently supported across models. Such consensus-based explanations may better capture stable characteristics of the underlying prediction task while reducing dependence on the behaviour of any individual trained model.

Finally, as explainable artificial intelligence increasingly transitions from academic research toward deployment in real-world clinical decision-support systems [10, 19], robustness may become an essential component of explanation quality rather than an optional evaluation criterion. Just as predictive models are routinely assessed for predictive performance before deployment, future explainable AI systems may also require systematic robustness validation to ensure that generated explanations remain reliable under realistic sources of model uncertainty.

7 Conclusion

This thesis investigated the robustness of counterfactual explanations across independently trained predictive models for healthcare prediction. While previous research has primarily focused on generating valid, sparse, and actionable counterfactual explanations for a single predictive model, considerably less attention has been devoted to understanding whether these explanations remain reliable when predictive models are independently retrained.

To address this problem, a multi-dimensional robustness evaluation framework was proposed and applied to counterfactual explanations generated using the DiCE framework on the Diabetes 130-US Hospitals dataset. Five independently trained XGBoost models were used to generate counterfactual explanations for the same patients, allowing robustness to be evaluated from four complementary perspectives: cross-model validity, feature-set consistency, direction consistency,

and magnitude consistency. Statistical uncertainty was subsequently quantified using patient-level bootstrap confidence intervals.

The experimental results demonstrated that different robustness dimensions exhibit substantially different behaviour. While cross-model validity and feature-set consistency indicated that independently trained predictive models frequently produced different counterfactual explanations, considerably stronger agreement was observed in the direction and magnitude of recommended feature changes once common intervention features were identified. These findings suggest that robustness is inherently multi-dimensional and cannot be adequately characterized using a single evaluation metric. Rather than proposing yet another counterfactual generation algorithm, this thesis argues that robustness itself should be treated as an essential evaluation criterion for existing explanation methods. This perspective complements current evaluation practices by emphasizing the reliability of explanations in addition to their validity and actionability. Taken together, these findings suggest that evaluating explanation robustness should complement, rather than replace, traditional counterfactual quality metrics such as validity, sparsity, and actionability.

More broadly, this thesis highlights the importance of evaluating explanation reliability in addition to predictive performance. As explainable artificial intelligence continues to play an increasingly important role in high-stakes healthcare decision-support systems [10, 17, 19], understanding whether explanations remain stable under realistic model uncertainty becomes an essential aspect of trustworthy machine learning. Although additional validation across different datasets, explanation algorithms, and predictive model families remains necessary, the proposed evaluation framework establishes a systematic foundation for studying robustness across independently trained predictive models and contributes to the growing body of research on trustworthy counterfactual explanations. By treating robustness as a multi-dimensional property rather than a single scalar quantity, this thesis offers a practical basis for evaluating explanation reliability under realistic model retraining scenarios. Rather than concluding that counterfactual explanations are simply robust or non-robust, the findings demonstrate that robustness depends on the aspect of explanation being evaluated. This highlights the importance of assessing multiple complementary robustness dimensions when evaluating the reliability of counterfactual explanations under model retraining.

References

[1] David Alvarez-Melis and Tommi Jaakkola. On the robustness of interpretability methods. In *ICML Workshop on Human Interpretability in Machine Learning*, 2018.

[2] Alejandro Barredo Arrieta et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[3] André Artelt, Valerie Vaquet, Riza Velioglu, Fabian Hinder, Johannes Brinkrolf, Malte Schilling, and Barbara Hammer. Evaluating robustness of counterfactual explanations. In *2021 IEEE Symposium Series on Computational Intelligence*, pages 1–9. IEEE, 2021.

[4] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[5] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[6] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

[7] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3681–3688, 2019.

[8] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):1–42, 2018.

[9] David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI Magazine*, 40(2):44–58, 2019.

[10] Andreas Holzinger et al. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*, 2017.

[11] Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. Formalising the robustness of counterfactual explanations for neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14901–14909, 2023.

[12] Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. Robust counterfactual explanations in machine learning: A survey. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 8086–8094. International Joint Conferences on Artificial Intelligence Organization, 2024.

[13] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: From counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 353–362. ACM, 2021.

[14] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[15] Christoph Molnar. *Interpretable Machine Learning*. Lulu.com, 2 edition, 2024.

[16] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617. ACM, 2020.

[17] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[18] Beata Strack, Jonathan P. DeShazo, Carissa Hernandez, et al. Impact of hba1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed Research International*, 2014:781670, 2014.

[19] Eric Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, 2019.

[20] Sahil Verma, John P. Dickerson, and Keegan Hines. Counterfactual explanations and algorithmic recourse: A review. *arXiv preprint arXiv:2010.10596*, 2020.

[21] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. In *Harvard Journal of Law & Technology*, volume 31, pages 841–887, 2017.