Universiteit Leiden

Master Computer Science

Applying Counterfactual Explanations and Multivariate Forecasting to Medical Prediction Tasks

Name:	Tomke Meyer
Student ID:	s2231086
Date:	25/08/2025
Specialisation:	Bioinformatics
1st supervisor:	Jan N. van Rijn
2nd supervisor:	Panagiotis Papapetrou (Stockholm University)

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Forecasting patient outcomes is a critical task in healthcare, with applications ranging from treatment planning to risk assessment. Although recent advances in time series forecasting have explored counterfactual explanations, existing approaches usually modify past observations, which limits their practical utility in clinical settings where future-oriented interventions are more relevant. This thesis introduces a new method for counterfactual time series forecasting, which generates realistic future trajectories by adjusting exogenous variables within the forecast horizon to achieve a desired clinical outcome. By directly altering future treatment plans, this approach could support healthcare on a much more personal level. Counterfactual explanations offer clinicians an intuitive way of exploring alternative future scenarios and evaluating the potential impact of different interventions, and therefore enhancing transparency and supporting more informed and goal-aligned decision-making.

We incorporate multiple forecasting models, such as SARIMAX, OLS, GRU, and N-BEATS, to learn the relationships between exogenous inputs and target variables, enabling the generation of interpretable and constrained counterfactuals. Extensive evaluation across two healthcare applications, glucose level forecasting using the OhioT1DM and the SimGlucose datasets, and mortality prediction in HFpEF patients using the MIMIC-IV, demonstrates the effectiveness and realism of the proposed method. In particular, the GRU model achieves superior predictive performance, with a 78% adherence to the clinical constraints with minimal intervention costs of 1.5% average input changes on the OhioT1DM dataset, and a low prediction error for the MIMIC-IV dataset. It consistently produces counterfactuals that adhere closely to known clinical values.

Our findings show that effective counterfactual interventions typically involve changes across multiple variables and that realistic, health-aligned adjustments can be achieved without significantly deviating from natural data distributions. Despite the promising results, challenges remain, particularly the need for clinical validation, such as healthcare expert review and real-world testing, to ensure the recommendations are clinically sound and actionable. Additionally, the realism of the generated counterfactuals inherently depends on the quality of the underlying forecasts. Inaccuracies in prediction could lead to misleading or clinically implausible results. We outline future directions for improving clinical relevance, such as collaborating with domain experts, integrating more extensive patient context, and optimising for adherence.

This work establishes a foundation for the development of transparent, applicable, and personalised forecasting tools in healthcare, that can simulate "what-if" scenarios to support better clinical decision-making.

Contents

1 Introduction

With the increasing availability of large and complex healthcare datasets, the usage of machine learning and deep learning techniques for clinical decision support has increased [ASFWHY+25, MPS19]. In particular, time series forecasting plays a crucial role in monitoring patient states and predicting clinical outcomes based on continuously recorded data, such as vital signs and laboratory results [KCK+20, JBS+23]. Accurate forecasts can enable early interventions, prevent complications, and improve personalised treatment strategies, which are essential in medical settings where timely and informed decisions can directly affect patient outcomes.
Despite advances in predictive accuracy, many deep learning models remain black-box models, limiting their interpretability and acceptance by clinicians. Clinicians require transparent reasoning and actionable insights to be able to trust and act on model outputs. Counterfactual explanations have emerged as a promising approach to enhance model interpretability by identifying minimal and feasible changes in input features that could lead to a more desirable prediction [VBH+24, Gui24]. While counterfactual explanations have been widely studied in static tabular data contexts, their extension to time series forecasting, especially in real-world, multivariate medical applications, remains limited. Existing methods often focus on altering historical data points, which is impractical in clinical practice because the past is fixed [WSMP24].
This thesis addresses this gap by combining multivariate forecasting with counterfactual explanations to support clinical decision-making. We focus on medical prediction tasks where forecasts of patient outcomes are critical, and actionable recommendations are required. Specifically, we consider scenarios in which clinicians need to anticipate changes in patient states, such as blood glucose levels in diabetes management or cardiac function in heart failure, so that interventions can be planned proactively. By integrating forecasts with counterfactual reasoning, our approach not only predicts future outcomes but also identifies which controllable factors could be adjusted to improve these outcomes, ensuring that the predictions are both accurate and actionable.

1.1 Motivation

Healthcare presents an interesting domain for counterfactual forecasting for several reasons. Patient data is complex, multivariate, and collected over time, creating rich time series suitable for predictive modelling. Accurate forecasts can inform early interventions, prevent complications, and improve personalised treatment strategies. Additionally, interpretability is crucial, as clinicians must understand the reasoning behind predictions to be willing to use the models. Two concrete applications are examined in this thesis:

- **Diabetes management:** Continuous glucose monitoring (CGM) enables real-time tracking of glucose levels. By forecasting future glucose trends and suggesting adjustments in insulin, diet, or activity, counterfactual predictions can proactively reduce risks of hyper- and hypoglycaemia [EMA+24a, EMA+24b].
- **Heart failure with preserved ejection fraction (HFpEF):** Early detection of deteriorating cardiac function through time series forecasting of vital signs and other patient data can guide lifestyle or treatment interventions, helping prevent hospital readmissions and improve outcomes [BP10, PVA+16].

These examples highlight why exogenous interventions, rather than modifying historical data, are critical for practical counterfactual forecasting in healthcare.

1.2 Main Contributions

Time series forecasting can play an important role in healthcare by predicting how well a treatment works or predicting the risk of complications, relapse, or mortality. While recent works, such as COMET [WSMP24], have explored the use of counterfactuals in time series forecasting, many existing methods work by retroactively modifying historical observations, a strategy that is not applicable in many medical contexts. This thesis contributes to the ongoing research in counterfactual time series analysis by proposing an alternative forecasting method that identifies optimal changes in exogenous variables during the forecast horizon to achieve desirable outcomes. More specifically, we propose a method that learns the relationship between the forecasted targets and exogenous variables, which leads to a more effective and interpretable decision-making in healthcare.
The main contributions of this thesis can be summarised as follows:

- We formally define a new variant of applying counterfactual explanations and multivariate forecasting to medical prediction tasks.
- We introduce a new counterfactual time series forecasting method to achieve a desired constrained forecast by modifying exogenous variables within the forecast horizon.
- We incorporate existing forecasting models, such as SARIMAX, OLS, GRU and N-BEATS, for learning the relationship between exogenous variables and a target variable to ensure actionable and interpretable predictions.
- We evaluate the models on two applications in healthcare, specifically for glucose level prediction and HFpEF management.
- We demonstrate the practical utility of incorporating counterfactuals for medical prediction tasks.

All code, models, and the publicly available datasets used in this thesis are available on our GitHub Repository,[1] providing full reproducibility and facilitating future research.

1.3 Problem Formulation

In this section, we formally define the problem of counterfactual time series generation in the context of time series forecasting. We introduce the notation used throughout the thesis and describe the objectives and constraints that guide the proposed methodology. Figure 1 illustrates the core idea of hybrid counterfactual forecasting. The upper line shows the evolution of the target variable $\mathbf{y}$ (for example blood glucose level). Based on the historical window (back horizon n), the forecasting model f produces an original forecast $\hat{\mathbf{y}}$ (blue curve) for the next t steps. In many medical applications, these forecasts may fall outside of clinically acceptable ranges. In this example, the predicted values exceed the upper bound $\boldsymbol{\beta}$, violating the desired constraints (red lines). A typical use case

[1] https://github.com/TomkeMeyer/ThesisTomkeMeyer.git

would be a glucose prediction that rises to 200 mg/dL, which is too high for a safe range. Here, the goal is to keep the values within a preferable interval, for example between $\alpha = 80$ mg/dL and $\beta = 140$ mg/dL. By enforcing these constraints, the method ensures that the new forecast remains both clinically safe and practically meaningful. To achieve this, our method modifies the exogenous variables $\hat{\mathbf{X}}$ predicted for the future horizon. These exogenous variables, shown in the lower half, for example insulin, carbohydrate intake or exercise intensity, are adjusted into new trajectories $\mathbf{X}^*$ (purple curves). By using $\mathbf{X}^*$ for the forecasting, we obtain an alternative forecast $\mathbf{y}^*$ (green dotted line) that stays within the safe interval $[\boldsymbol{\alpha}, \boldsymbol{\beta}]$. In other words, the counterfactual trajectory suggests actionable changes in controllable variables that lead to a desirable forecasted outcome. This process illustrates the main idea behind our method: starting from an unsafe forecast, we generate counterfactual versions of the exogenous variables that move the prediction into a safe and clinically meaningful range.

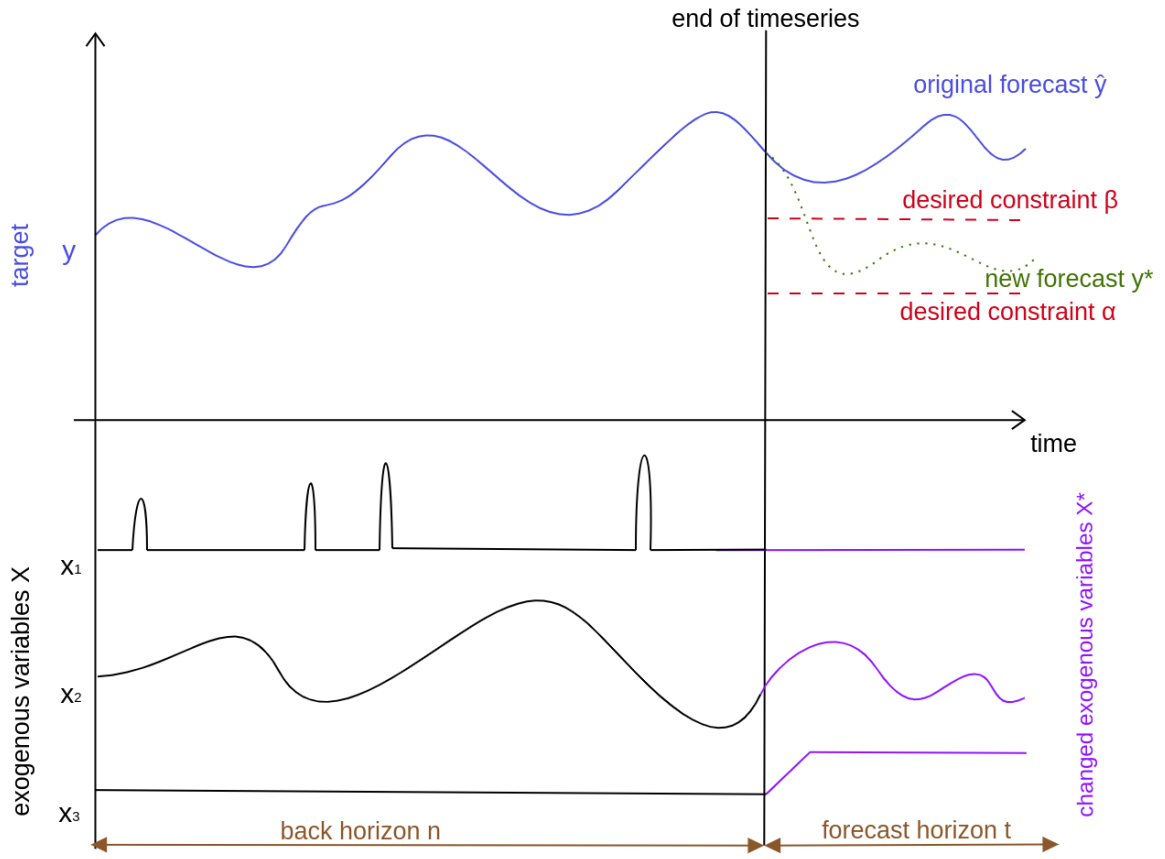


Figure 1: Illustration of the counterfactual forecasting problem. The model first produces an original forecast $\hat{\mathbf{y}}$ for the target variable. If the forecast lies outside of desired bounds $[\boldsymbol{\alpha}, \boldsymbol{\beta}]$, the exogenous forecast $\hat{\mathbf{X}}$ is modified into $\mathbf{X}^*$, yielding a new forecast $\mathbf{y}^*$ that is within the constraints.

Time series definition:
Let $\mathbf{D} := (\mathbf{D}_i)_{i \in \{1,\dots,n\}}$ denote a multivariate time series of length n (back horizon), with

each $d_i \in \mathbb{R}^{m+1}$ composed of the target variable $y_i \in \mathbb{R}$ and the exogenous variables

$$x_i = \begin{pmatrix} x_{1,i} \\ \vdots \\ x_{m,i} \end{pmatrix} \in \mathbb{R}^m.$$

Then $\mathbf{D}$ can be denoted as the combined matrix of the target vector $\mathbf{y} \in \mathbb{R}^{1\times n}$ and the exogenous matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$:

$$\mathbf{D} = \begin{pmatrix} \mathbf{y} \\ \mathbf{X} \end{pmatrix} := \begin{pmatrix} y_1 & \dots & y_n \\ x_{1,1} & \dots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \dots & x_{m,n} \end{pmatrix}.$$

The relationship between $\mathbf{y}$ and $\mathbf{X}$ can be described by the function:

$$c : \mathbb{R}^{m\times n} \to \mathbb{R}^{1\times n}$$

which calculates the target vector from the exogenous matrix:

$$c(\mathbf{X}) = \mathbf{y}.$$

Forecasting model:
Given a multivariate time series forecasting model f that predicts the next t values (forecasting horizon) of $\mathbf{D}$, we define the forecast as:

$$f(\mathbf{D}) = \hat{\mathbf{D}} := (d_{n+i})_{i\in\{1,\dots,t\}}, \quad \hat{\mathbf{D}} = \begin{pmatrix} \hat{\mathbf{y}} \\ \hat{\mathbf{X}} \end{pmatrix}.$$

Here, $\hat{\mathbf{y}}$ is the forecasted target trajectory, and $\hat{\mathbf{X}}$ the forecasted exogenous inputs. There exist various forecasting methods that implement f; we will compare the performance of various of these in Section 6.1.

Constraints and counterfactual objective:
For each step in the forecast horizon, we define lower and upper bounds on the target variable, denoted as:

$$\boldsymbol{\alpha} = (\alpha_{n+i})_{i\in\{1,\dots,t\}}, \boldsymbol{\beta} = (\beta_{n+i})_{i\in\{1,\dots,t\}}.$$

The objective is to generate a counterfactual time series sample $\mathbf{X}^*$, such that $\mathbf{y}^* = c(\mathbf{X}^*)$ satisfies the given bounds:

$$\alpha_i \leq y_i^* \leq \beta_i, \forall y_i^* \in \mathbf{y}^*, i \in \{n+1, \dots, n+t\}.$$

Summarised research objective:
Given a target vector $\mathbf{y}$ affected by the exogenous matrix $\mathbf{X}$, a forecast horizon t, the original forecasted vector $\hat{\mathbf{y}}$ and the original forecasted exogenous matrix $\hat{\mathbf{X}}$, the goal is to modify $\hat{\mathbf{X}}$ to $\mathbf{X}^*$ such that the corresponding target vector $\mathbf{y}^*$ is within constraints $[\boldsymbol{\alpha}, \boldsymbol{\beta}]$. This corresponds to generating actionable interventions on future exogenous variables that lead the target prediction toward desirable outcomes.

2 Background

The global population is both growing and ageing, which is resulting in a rise in chronic diseases. This leads to a demand for more efficient, personalised, and proactive healthcare solutions [HSL+23]. In the past, treatment was typically given only after symptoms appeared, and the chosen treatment plans were often quite general. However, this reactive approach is no longer sufficient today, as patients require more personalised and proactive care. Technological advancements in the field of healthcare have had a significant impact on healthcare providers and patients, with statistical methods being used to predict outcomes. This approach struggles with the complex nature of clinical, demographic and molecular factors that influence the disease progression, leading to machine learning, deep learning, artificial intelligence, and big data analytics becoming increasingly popular fields within the medical and health science domains [ASFWHY+25]. Recent healthcare has been characterised by an increased need for data-driven approaches, with the care process being driven by the flow of data between patients and doctors, and the sharing of decisions, instructions and information amongst care providers. The role that data and information play in decision making and provision of healthcare, has only increased with the growing digitalisation of healthcare. This results in great amounts of data, which enables the implementation of advanced analytical methods, including machine learning and artificial intelligence methods, to obtain valuable and actionable insights. These insights are essential in supporting decision-making processes, ensuring high quality patient care, responding to real-time situations and ultimately reducing mortality. Especially machine learning becomes more and more relevant in healthcare applications, including predictive analysis, treatment optimisation, and patient monitoring [MPS19]. Machine learning algorithms can potentially be used to improve diagnostic accuracy, as well as support early disease detection and prediction. Other applications include, analysing medical imaging data, such as X-rays and MRIs, to detect signs of cancer or neurological disorders, which allows for early diagnosis, as well as a personalised treatment plan for each patient. These uses show the potential that machine learning has for both research and clinical trials, and support the improvement of healthcare overall [RNZ17], [JHS+22].

2.1 Time Series Analysis and Forecasting

Time series analysis uses an ordered sequence of data points recorded over time, usually at regular intervals, to understand patterns, trends, and relationships within the data over time. This time series data also allows for forecasting or predicting future values, based on historical observations. Time series forecasting plays a crucial role in a number of different applications. Some applications include its usage in finance, for predicting stock prices and market trends [MSG14], in meteorology for weather and climate forecasting [KS20], in transportation, for effective traffic flow forecasting [LBF13] and especially recently in healthcare. In healthcare, time series forecasting is increasingly relevant, as patient data such as heart rate, blood pressure, glucose levels, and laboratory results are being collected continuously or in regular intervals. Precise forecasting of these medical time series can allow for early detection of negative outcomes, support the personalisation of treatment plans, and optimise healthcare management overall [KCK+20], [JBS+23].

Generally, time series forecasting consists of finding temporal dependencies and trends in the data. Depending on the nature of the time series, such as for example linear or non-linear, stationary or non-stationary, and univariate or multivariate, requires different forecasting methods. These different forecasting methods can be categorised into three subgroups, statistical models, such as ARIMA and ETS, machine learning models, such as Linear Regression and SVMs, and deep learning models, such as RNNs, LSTMs, GRUs as well as transformer-based models Autoformer or DLinear [MWM24], [Kol24].
Especially the latter models have shown promising results in recent years, achieving very accurate predictions [WLY24]. Despite these recent developments, many deep-learning based forecasting systems are considered black-box models, as it is challenging to interpret and understand both the modelling process and the forecasting outcome. This is particularly problematic in healthcare applications, where particularly interpretability is very important, as clinicians need to understand the reasoning behind the predictions to make informed decisions.

2.2 Counterfactual Explanations

Counterfactual explanations are an emerging technique with the potential to improve the interpretability and explainability of machine learning models. The objective is to identify minimal changes to the input features that would result in a different, and typically more desirable, outcome. More specifically, they provide information such as if an input data point would be $\hat{\mathbf{X}}$ instead of $\mathbf{X}$, then the trained machine learning models prediction would be $\hat{\mathbf{Y}}$ instead of $\mathbf{Y}$, assuming outcome $\hat{\mathbf{Y}}$ would be more favourable [VBH+24]. So counterfactuals are a type of explanation method in machine learning, that might help users to understand the model predictions and decisions better. This is supported by counterfactual explanations being actionable, since they suggest specific changes to alter the outcome, and intuitive for humans, since they align with human reasoning about cause and effect. Often counterfactual explanations are also model-agnostic, meaning they can be applied to a variety of black-box models [Gui24]. A more specific example application of counterfactuals would be a suggestion to slightly reduce a patients systolic blood pressure, which could lead to a prediction of a lower cardiovascular risk. So these kind of insight can not only help clinicians understand a models decisions better, but also assist in timely intervention and treatment planning.
Traditional time series forecasting and counterfactual explanations focus mainly on predicting future values given historical observations, but modifying past values is not feasible in real-world settings, especially in healthcare applications. Instead, a more practical approach is to explore how changing exogenous variables during the forecast horizon could lead to a desired outcome. This approach allows continuous monitoring of patients, making it possible to dynamically adjust treatment plans to lead towards more optimal results.

2.3 Diabetes

Diabetes is one of the most prevalent chronic diseases in the world, with it being a leading cause of death and disability. According, to the World Health Organisation (WHO), around 830 million people worldwide suffer from diabetes, and it is the direct cause of 1.5 million deaths a year. Nowadays, it affects around 14% of adults, this number has doubled since 1990, making it a major public health problem. Diabetes is characterised by elevated

levels of blood glucose, which over time seriously damages the heart, blood vessels, eyes, kidneys and nerves. Type 2 diabetes is much more common and occurs when the body either becomes resistant to insulin or does not make enough insulin. It usually has a later onset and its development can be attributed to factors such as being overweight, not getting enough exercise and genetic predisposition. Type 1 diabetes is a chronic condition, where the pancreas only produces little to no insulin by itself. It is caused by the autoimmune destruction of pancreatic β-cells and affects 5-10% of the diabetes patients [Dia], [EMA+24a]. For patients with conditions such as type 1 diabetes mellitus (T1DM), closely tracking their glucose levels is a necessity. To reduce the risk of complications such as hyperglycaemia (high blood sugar) or hypoglycaemia (low blood sugar), these patients often rely on continuous glucose monitoring (CGM) devices and automated insulin delivery. Hyperglycaemia can lead to serious long-term complications like nerve damage, kidney failure, and cardiovascular issues if not properly managed. Hypoglycaemia, on the other hand, can cause immediate dangers such as dizziness, confusion, seizures, or even loss of consciousness. By continuously tracking glucose levels, CGM devices provide real-time data that helps patients and clinicians maintain blood sugar within a safe range. Integrating machine learning models with CGM data can further enhance diabetes management by predicting abnormal glucose events before they occur. This predictive capability allows for timely adjustments in insulin dosage, improving overall glycaemic control and reducing the risks associated with both hyper- and hypoglycaemia. Eventually, machine learning-driven insights can support more personalised and effective insulin delivery plans, contributing to better patient outcomes. By also incorporating variables such as insulin intake, carbohydrate consumption, and physical activity, a predictive model can allow timely interventions through the generation of actionable recommendations for patients or healthcare providers. This allows for more dynamic treatment based on these forecasted expected glucose trends, which can reduce the long-term risk of diabetes-related complications [EMA+24b].

2.4 Heart Failure With Preserved Ejection Fraction

Heart failure with preserved ejection fraction (HFpEF) is a prevalent and severe cardiovascular condition, where the heart's pumping strength is normal but the heart muscle does not relax properly. This can lead to the heart not filling properly, leading to hospitalisation and in the worst cases to death [BP10]. Heart failure (HF) is classified based on the left ventricular ejection fraction (LVEF) and can be split into three subcategories. Heart failure with reduced ejection fraction (HFrEF) with LVEF $\leq$40%, heart failure with mildly reduced ejection fraction (HFmrEF) with LVEF = 41-49%, and heart failure with preserved ejection fraction (HFpEF) with LVEF $\geq$50% [PVA+16]. Studies have shown that the mortality within one year is around 29%, [OHH+06], [SBA+24], with increased mortality for patients with previous heart failure hospitalisations and other comorbidities [MGL+19]. This makes HFpEF a very serious condition, where early diagnosis is key. So another possible application of medical time series forecasting could be to try identifying early warning signs of heart failure or more specific HFpEF to suggest either lifestyle or treatment changes. Using the vital signs, as for example the heart rate and blood pressure, of a patient as well as other factors like gender and possible comorbidities, allows specific and personal monitoring of disease progression. This way, early warning signs of worsening heart condition can be identified and personalised modifications to lifestyle or medication

can be suggested. Such proactive monitoring can help reduce hospital readmissions and improve patient outcomes.

3 Related Work

The following sections introduce some related work in three different fields. First, related methods in counterfactual forecasting are presented, focusing on the implementation of counterfactuals to improve explainability and interpretability. Then, existing approaches in forecasting blood glucose levels and predicting mortality for patients with HFpEF are analysed, to highlight the necessity for our approach.

3.1 Counterfactual Forecasting Models

Recent research has explored various deep learning models for time series forecasting, including recurrent neural network (RNN)-based models such as gated recurrent units (GRU) and long short-term memory (LSTM), as well as attention-based architectures like transformers [Kol24]. Transformer-based models, including Autoformer and Informer, have demonstrated strong performance in both univariate and multivariate forecasting tasks by capturing long-range dependencies more effectively than traditional RNN approaches. In the clinical domain, deep learning models have been applied extensively to glucose forecasting. For instance, Deep Multi-Output Forecasting [FAJ+18] introduced a multi-step forecasting framework that explicitly models the distribution of future glucose values over a prediction horizon using a multi-output deep architecture. Similarly, WaveNet has been adapted for glucose forecasting by leveraging dilated convolutional neural networks (CNNs) to model long-term dependencies [ZLH+18]. In addition, transfer learning techniques have been employed to enhance predictive performance by fine-tuning pre-trained models on patient-specific data while incorporating exogenous covariates such as insulin dosage and carbohydrate intake [MB20].

Beyond predictive performance, explainability remains a critical challenge in deep learning-based forecasting models. Traditional statistical models, such as ARIMAX and VARIMAX, are able to quantify relationships between exogenous factors and the target variable, but their forecasting accuracy is often outperformed by deep learning approaches [PTJ+22]. Recent research has focused on integrating explainability into forecasting models to combine the strengths of both interpretability and predictive performance. For example, N-BEATSx extends the N-BEATS method by incorporating future exogenous variables into its deep architecture, enabling a more structured decomposition of trend and seasonality. However, its interpretability remains static and does not fully capture the dynamic nature of forecasting outcomes and requires future exogenous values as input [SM23], [OCM+22].

To address the need for explainability, counterfactual explanations have gained traction in time series analysis. Initial efforts focused on time series classification, where counterfactuals were generated through instance-based modifications and gradient-based perturbations [AALC20]. This was done by introducing a framework for generating counterfactual explanations for multivariate time series classification, identifying minimal input modifications needed to alter the model's decision, providing interpretability for high-dimensional time series models.

More recently, counterfactual explanations have been extended to time series forecasting. ForecastCF [WMSP23] proposed a deep learning-based method for generating counterfactuals in time series forecasting by identifying minimal input changes required to achieve desired prediction outcomes. Building on this, COMET [WSMP24] extended counterfac-

tual explanations to multivariate time series forecasting, focusing on modifying exogenous variables, such as insulin, carbohydrates, and exercise, to generate actionable recommendations for glucose management.
Despite these advances, counterfactual explanations for multivariate time series analysis remain an emerging research area. While existing methods demonstrate the feasibility of generating counterfactuals for univariate forecasts, their generalisation to multivariate forecasting and real-world clinical applications remains limited. This work aims to extend on these existing methods by integrating counterfactual reasoning with multivariate forecasting models, focusing on modifying exogenous variables within the prediction horizon to provide actionable and interpretable interventions.

3.2 Blood Glucose Level Prediction

Forecasting physiological indicators such as blood glucose levels, is crucial for managing diabetes. Time series forecasting and model explainability are becoming increasingly important in this field of medical prediction tasks, as in many others. Recent contributions to this area of research highlight the growing use of multivariate machine learning models to improve predictive accuracy and personalisation. This shows the growing need for interpretable and actionable insights, which aligns closely with the goals of this thesis.
Recent research highlights the growing development of multivariate and deep learning-based models for predicting glucose levels. For example, Kalita and Mirza [KM25] proposed a model that combines multi-head attention layers with neural basis expansion networks, capturing complex temporal and cross-feature dependencies in glucose data. Similarly, Benaida et al. [BAI25] demonstrated the effectiveness of deep learning architectures for both single- and multi-step glucose forecasting, emphasising the importance of long-term prediction capabilities in real-world applications. These multivariate models are consistent with the focus of this thesis on leveraging multiple signals, such as past glucose levels, physiological parameters, and contextual variables, for accurate forecasting.
Personalisation has emerged as a key factor in clinical forecasting settings, as patient diversity affects model performance. Shen and Kleinberg [SK25a] addressed this issue by using incrementally retrained LSTM networks that adapt to each individual's glucose dynamics. This improves performance, even when the CGM data is limited. Lara-Abelenda et al. [LACMPC+25] introduced large language models to model personal glucose trends, highlighting the capacity of foundation models to generalise across individuals while retaining patient-specific nuances. These methods emphasise the importance of adaptive and context-aware forecasting.
Several works have also incorporated physiological signals beyond glucose levels to support multivariate forecasting. For example, Giancotti et al. [GBV+24] explored the utility of heart rate as a predictor of forecasting glucose levels in patients with type 1 diabetes, which demonstrates that multimodal data can significantly enhance predictive accuracy. Similarly, Rodríguez-Rodríguez et al. [RRCVR23] utilised data, such as physical activity and diet logs, to enable more holistic and personalised glycaemic forecasting.
Interpretability remains a major challenge for deep learning-based forecasting models, especially in critical fields such as medicine. In response to this, Sun and Kosmas [SK25b] combined a Bayesian forecasting method with expert medical knowledge to model CGM values in type 2 diabetes patients. Their framework improves both uncertainty quantification and clinician interpretability, which is an essential consideration in Healthcare

AI. The need for model transparency directly motivates the use of counterfactual explanations to improve the explainability and actionability of predictive models in medical applications. While many current studies emphasise predictive accuracy, fewer address how predictions can be explained and acted upon by clinicians or patients.
Taken together, these studies reflect a shift towards data-driven, multivariate, and personalised models for medical forecasting. However, there remains a clear gap in integrating these powerful models with robust, interpretable explanations. This thesis aims to bridge this gap by combining multivariate forecasting approaches with counterfactual reasoning, to provide accurate predictions and actionable, understandable explanations, which are essential components for supporting medical decision-making and patient self-management.

3.3 HFpEF Mortality Prediction

Heart Failure with Preserved Ejection Fraction (HFpEF) is a complex and heterogeneous condition, characterised by diagnostic and prognostic uncertainty. This makes it a compelling use case for machine learning in clinical decision support. Recent research has applied various machine learning techniques to improve diagnosis, predict outcomes such as hospitalisation and mortality, and guide individualised management strategies. These efforts emphasise the increasing relevance of multivariate forecasting and the growing demand for explainable models, which are central to the objectives of this thesis.
A significant amount of research has focused on prognostic modelling using structured clinical data. For example, Hu et al. [HMH+25] developed and validated a machine learning model to predict the risk of readmission within one year for HFpEF patients, demonstrating the utility of routinely collected electronic health records (EHRs) in anticipating adverse outcomes. Similarly, McDowell et al. [MKT+24] constructed models for predicting both mortality and morbidity in HFpEF patients, showing that complex risk factors, including comorbidities and laboratory values, can be effectively integrated into predictive models. These studies emphasise the importance of leveraging multivariate data sources to forecast long-term patient outcomes.
Short-term outcome prediction has also been explored, particularly in the context of the early identification of high-risk patients. Another study [SBA+24] used machine learning to predict short-term mortality, which is essential for planning acute care. Other models are focusing on hospitalisation prediction, using historical patient trajectories to anticipate future events. These forecasting tasks not only require accurate time series modelling but also benefit from interpretability to inform clinical decisions.
The diagnosis of HFpEF remains a challenging area due to its symptomatic overlap with other heart failure subtypes. Kavas et al. [KBB23] developed an machine learning-based decision support system using photoplethysmography (PPG) signals to differentiate between HFpEF and HFrEF (Heart Failure with reserverd Ejection Fracion), demonstrating the potential of non-invasive, sensor-based diagnostics. Liao and Hung [CLH24] further extended this approach by incorporating data from a wearable patch device to enhance diagnostic precision. These works highlight the growing role of physiological signal data in heart failure classification, which directly supports multivariate modelling approaches by introducing continuous and high-frequency signals into prediction tasks.
Genomic and molecular data have also been used to support precision medicine approaches in HFpEF. Zhou et al. [ZGW+21] utilised gene expression profiles to build machine learning models capable of risk stratification in HFpEF patients, adding a layer of biological

interpretability to purely clinical models. Although these models are powerful, they are often perceived as black boxes, emphasising the need for explainability techniques such as counterfactual explanations to bridge the gap between model prediction and clinical insight.

Across these studies, however, the challenge of model transparency and interpretability remains largely unaddressed. Most existing models prioritise predictive performance without offering sufficient explanations for individual predictions, which is a critical issue in medical contexts where understanding why a prediction was made is often as important as the prediction itself. This thesis aims to bridge this gap by integrating counterfactual reasoning into multivariate forecasting models, offering clinicians not just a forecast, but a clear explanation of the factors driving the prediction and the minimal changes that could possibly alter an adverse outcome.

In summary, the current research in HFpEF prediction demonstrates the power of machine learning to handle complex, multivariate data across diagnostic and prognostic applications. However, a lack of interpretability limits clinical adoption. This thesis tries to contribute to the field by combining accurate time series forecasting with interpretable, counterfactual explanations, thereby supporting more transparent and actionable decision-making in heart failure care.

4 Counterfactual Hybrid Forecasting

We propose an algorithm to generate counterfactual exogenous inputs $\mathbf{X}^*$ that guide the forecasted target $\mathbf{y}^*$ toward a desired outcome within a multivariate time series context. The algorithm integrates a differentiable forecasting model f and a constrained counterfactual optimisation procedure, ensuring the realism of the generated counterfactuals. It takes as input a multivariate time series $\mathbf{D} = (\mathbf{y}, \mathbf{X})$, where $\mathbf{y}$ is the target vector and $\mathbf{X}$ the matrix of exogenous variables. Hyperparameters include the learning rate η, clipping range (ρ, ϕ), maximum number of optimisation iterations max_iter, and, when applicable, target bounds $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. The differentiable models for forecasting $f(\cdot)$ and counterfactual generation $c(\cdot)$ are provided, where f predicts the future values $\hat{\mathbf{D}}$ and c guides updates on $\mathbf{X}$ to achieve a feasible counterfactual.
The approach is structured into three main stages:

- Multivariate Forecasting: A base forecasting model f is trained to predict the future trajectory of both the target $\hat{\mathbf{y}}$ and the exogenous variables $\hat{\mathbf{X}}$ given historical observations $\mathbf{X}$.
- Desired Bound Generation: Bounds $[\boldsymbol{\alpha}, \boldsymbol{\beta}]$ are specified for the target forecast $\hat{\mathbf{y}}$, reflecting clinically or contextually desired outcomes.
- Counterfactual Optimisation: The predicted exogenous trajectory $\hat{\mathbf{X}}$ is iteratively perturbed into $\mathbf{X}^*$ using gradient-based optimisation, guided by a constrained loss function. The constraints ensure that the resulting counterfactuals remain clinically feasible, temporally realistic, and statistically aligned with observed trajectories.

This algorithmic approach provides an adjustable and interpretable method for generating actionable interventions in time series forecasting tasks. By ensuring that counterfactuals are both feasible and effective, it is particularly suited for medical prediction tasks where unrealistic or unsafe interventions should be avoided. For tasks with binary outcomes, the optimisation is adapted by omitting bound generation and focusing directly on the desirable outcome. Algorithm 1 shows the pseudocode for the method.

Algorithm 1 Counterfactual Hybrid Forecasting

Input: Time series data: target $\mathbf{y}$, exogenous variables $\mathbf{X}$, learning rate η, desired bounds $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, clipping range (ρ, ϕ), max iterations max_iter, differentiable forecaster $f(\cdot)$, differentiable counterfactual generator $c(\cdot)$, weight w, historical values $\mathcal{G}$
Output: Counterfactual $\mathbf{X}^*$ with desired outcome $\mathbf{y}^*$
$(\hat{\mathbf{y}}, \hat{\mathbf{X}}) \leftarrow f(\mathbf{y}, \mathbf{X})$
$\mathcal{S} \leftarrow$ SelectTestSamples
$[\boldsymbol{\alpha}, \boldsymbol{\beta}] \leftarrow$ GenerateBounds($\mathcal{S}$)
$C \leftarrow$ ActivityTemporalConstraint
loss $\leftarrow L((\hat{\mathbf{y}}, \hat{\mathbf{X}}), \boldsymbol{\alpha}, \boldsymbol{\beta}, (\mathbf{y}, \mathbf{X}), C)$
time $\leftarrow 0$
while $(\mathbf{y}^* > \boldsymbol{\beta}$ **or** $\mathbf{y}^* < \boldsymbol{\alpha}) \wedge$ (time $<$ max_iter) **do**
    $\mathbf{X}^* \leftarrow$ AdamOptimize($\mathbf{X}^*, loss, \eta$)
    $\mathbf{X}^* \leftarrow$ Clip($\mathbf{X}^*, \rho, \phi$)
    $\mathbf{y}^* \leftarrow c(\mathbf{X}^*)$
    $C \leftarrow$ HistValueConstraint($\mathbf{X}^*, \mathcal{G}$)
    loss $\leftarrow L(\mathbf{X}^*, w, \boldsymbol{\alpha}, \boldsymbol{\beta}, (\hat{\mathbf{y}}, \hat{\mathbf{X}}), C)$
    time $\leftarrow$ time $+ 1$
end while
return $(\mathbf{y}^*, \mathbf{X}^*)$

4.1 Desired Bound Generation

To encourage counterfactual generation toward realistic and safe outcomes, we define personalised bounds on the target variable. These bounds function as constraints for the optimisation, ensuring that predicted counterfactual targets remain within clinically realistic ranges.

Formally, for each forecast step $i \in n+1, \ldots, n+t$, we define a lower bound α_i and an upper bound β_i on the target variable y_i. In applications such as glucose forecasting, these bounds are based on the patient's current state, allowing for a smooth and feasible transition toward the desired target value. Let y_{n+1} denote the first predicted target value in the forecast horizon, and let y^* represent the desired target after S steps. A polynomial transition function of order p defines a target trajectory:

$$b(i) = y_{n+1} + (y^* - y_{n+1}) \cdot \left(\frac{i-n}{S} \right)^p, \quad i = n+1, \ldots, n+S$$

This trajectory ensures a gradual adjustment from the current value to the target. For steps beyond S, the bounds remain flat at y^*. To account for natural variability in the time series, we add a margin proportional to the standard deviation σ of the input series:

$$\alpha_i = b(i) - \lambda \cdot \sigma, \quad \beta_i = b(i) + \lambda \cdot \sigma, \quad \lambda \in [0, 1]$$

Here, λ is a tunable hyperparameter controlling the width of the bounds, typically set to 0.5. All calculations are performed in normalised space using patient-specific scalers derived from the training data.

For the HFpEF mortality prediction task, the target is binary (survival vs. non-survival). In this case, no bounds are generated, and the optimisation instead focuses on achieving the desirable outcome (survival) directly through the counterfactual perturbations.

4.2 Constraints

To ensure that the generated counterfactual trajectories effectively alter predicted outcomes, while remaining clinically feasible and realistic, we introduce a set of constraint mechanisms directly into the optimisation process. These are the clipping constraint, the historical value constraint, and the activity temporal constraint. Each of these constraints plays a distinct role in maintaining the interpretability, trustworthiness, and clinical plausibility of the counterfactuals.
The constraints are applied iteratively during optimisation, guiding the perturbation of exogenous variables such that the resulting trajectories adhere to domain-specific boundaries and realistic temporal patterns.

4.2.1 Clipping Constraint

The clipping constraint ensures that the perturbed exogenous variables $\mathbf{X}^*$ remain within physically and clinically plausible ranges. Let (ρ, ϕ) denote the minimum and maximum allowable values for each feature, chosen based on clinical knowledge or prior observations. After each optimisation step, elements of $\mathbf{X}^*$ are clipped:

$$X^*_{k,j} = \min(\phi_k, \max(\rho_k, X^*_{k,j})), \quad k \in \{1, \ldots, m\},\ j \in \{1, \ldots, t\}$$

This projection ensures that counterfactuals remain within trusted ranges of realistic values.

4.2.2 Historical Value Constraint

To maintain plausibility, counterfactual inputs should resemble historically observed patterns. Let $\mathcal{G}$ be a dataset of historical exogenous trajectories. The historical value constraint penalises deviations from the closest historical sequence using a Manhattan distance:

$$C_{\text{hist}}(\mathbf{X}^*) = \lambda_{\text{hist}} \cdot \min_{X \in \mathcal{G}} \sum_{k,j} \left| X^*_{k,j} - X_{k,j} \right|$$

where λ_{hist} controls the strength of the constraint. This encourages the generated sequences to remain close to real-world data.

4.2.3 Activity Temporal Constraint

The activity temporal constraint restricts perturbations to clinically actionable time steps. Define a binary vector $C \in \{0, 1\}^t$, where $C_j = 0$ allows a change at time step j, and $C_j = 1$ discourages it. For example, interventions such as medication, exercise, or meals occur at specific times, and changes outside these windows are penalised:

$$C_j = \begin{cases} 0 & \text{if intervention allowed at time } j \\ 1 & \text{otherwise} \end{cases}$$

The corresponding penalty for temporal misalignment is:

$$C_{\text{act}}(\mathbf{X}^*, \hat{\mathbf{X}}) = \sum_j C_j \cdot \sum_k \left| \frac{X^*_{k,j} - X'_{k,j}}{X'_{k,j} + \epsilon} \right|$$

where ϵ is a small constant to avoid division by zero. This ensures that changes primarily occur at relevant, clinically meaningful times.

4.2.4 Combined Loss Function

The optimisation objective integrates the forecasting goal with all constraints. Let $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ denote the desired bounds for the target $\mathbf{y}^*$. The total loss is defined as:

$$L = \text{ForecastLoss}(\mathbf{y}^*, \boldsymbol{\alpha}, \boldsymbol{\beta}) + \lambda_{\text{hist}} \cdot C_{\text{hist}}(\mathbf{X}^*) + \lambda_{\text{act}} \cdot C_{\text{act}}(\mathbf{X}^*, \hat{\mathbf{X}})$$

where λ_{hist} and λ_{act} tune the importance of each constraint. The clipping constraint is enforced via projection and does not appear explicitly in the loss.
These three constraints work together to guide the counterfactual generation process as the clipping constraint ensures that variable values remain in a trusted clinical range, the historical value constraint keeps the counterfactuals close to real, observed trajectories, and the activity temporal constraint encourages changes to occur at appropriate, clinically meaningful time steps. By embedding these constraints into the optimisation loop, the generated counterfactuals are not only effective but also interpretable, and aligned with realistic clinical dynamics.

5 Experimental Setup

5.1 Data

While the proposed model could be broadly applied across domains beyond healthcare, this study focuses exclusively on medical use cases. In particular, we investigate the models usefulness in optimising treatment plans for two conditions: type 1 diabetes and heart failure with preserved ejection fraction (HFpEF). The datasets used in these experiments contain physiological measurements and treatment-related variables, allowing for personalised forecasts and counterfactual intervention generation. To prepare the data for modelling, several preprocessing steps have been applied. First, the data is split into training, validation and test sets and normalised using min-max scaling to ensure stable and consistent model training. The data is then separated into target variables and exogenous inputs. Then, a sequence generator is employed to segment the time series into overlapping windows comprising a back horizon (historical input) and a prediction horizon (target output). This is done both for the target variable alone and for sequences that include both the target and exogenous features. Any sequences containing missing values are discarded to ensure data integrity.

5.1.1 SimGlucose

The SimGlucose dataset is generated using the FDA-approved UVA/PADOVA type 1 diabetes simulator [Xie18], a Python-based tool that models the physiological responses of individuals with type 1 diabetes. The simulator includes 30 virtual patients, comprising of 10 adults, 10 adolescents, and 10 children and produces continuous glucose monitoring (CGM) measurements along with insulin dosages and carbohydrate intake events. The dataset is generated with a predefined CGM sampling frequency and insulin pump settings based on the algorithm developed in [DMML+14]. For this study, simulated data was generated for ten adult patients over a one-week period. The blood glucose levels serve as the primary target variable, while the insulin dosage and carbohydrate intake are used as exogenous variables influencing the glucose levels. After generation, the data undergoes preprocessing steps as described above to prepare it for the forecasting and counterfactual generation. An example of the generated data is shown in Figure 2. In this example, BG denotes blood glucose levels, CHO indicates carbohydrate intake, and Insulin reflects the administered dosage. The risk index illustrates periods of hyperglycaemic or hypoglycaemic risk. The green band in the blood glucose trace highlights the target glucose range (70–180 mg/dL), while red regions denote values that fall outside this range, corresponding to hypo- or hyperglycaemia.

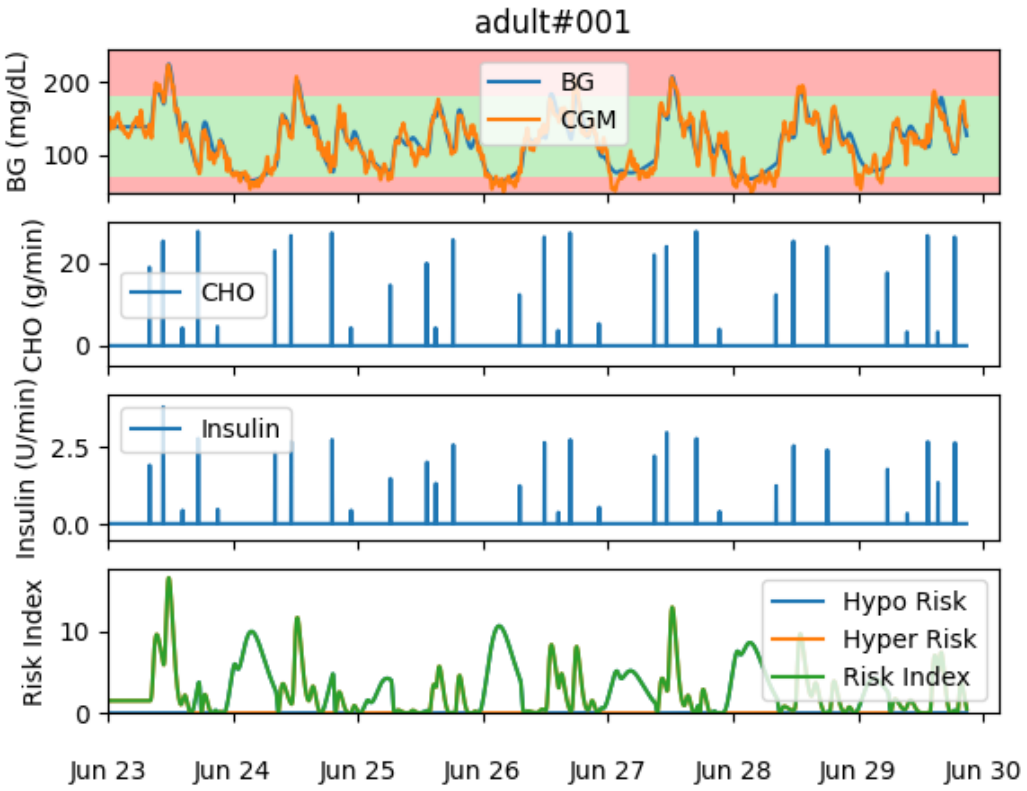


Figure 2: Simulation of an adult patient over the span of one week.

5.1.2 OhioT1DM:

The OhioT1DM dataset [MB20] contains real-world glucose monitoring data collected from 12 individuals with type 1 diabetes over an eight-week period by Ohio University. Following prior research [CHN+21], [WSMP24], we extracted the most clinically relevant features for forecasting: continuous glucose monitoring (CGM) measurements, basal insulin, bolus insulin, carbohydrate intake, and physical activity. Compared to the SimGlucose dataset, OhioT1DM includes a more varied set of exogenous variables, particularly basal and bolus insulin administration, dietary intake, and physical activity data. As a real-world dataset, it presents additional challenges such as missing values and irregular sampling intervals. These are addressed using interpolation and resampling techniques to ensure consistency in the input sequences. The inclusion of diverse exogenous variables enables the development of more nuanced counterfactual interventions and supports improved forecasting performance. Figure 3 illustrates a 24-hour time window for one patient, showing the temporal relationship between glucose levels, insulin administration, and carbohydrate consumption. The blue dotted line represents CGM-based blood glucose levels, while the black line represents the basal insulin, reflecting its slow-acting, sustained delivery throughout the day. Orange dots indicate the amount of bolus insulin, while small blue boxes indicate meals. Spikes in blood glucose often correspond to meal times (carbohydrate intake), followed by bolus insulin doses that aim to bring glucose back into the target range. This visualisation exemplifies the complex dynamics and temporal dependencies that the model must capture to enable accurate and personalised glucose forecasting.

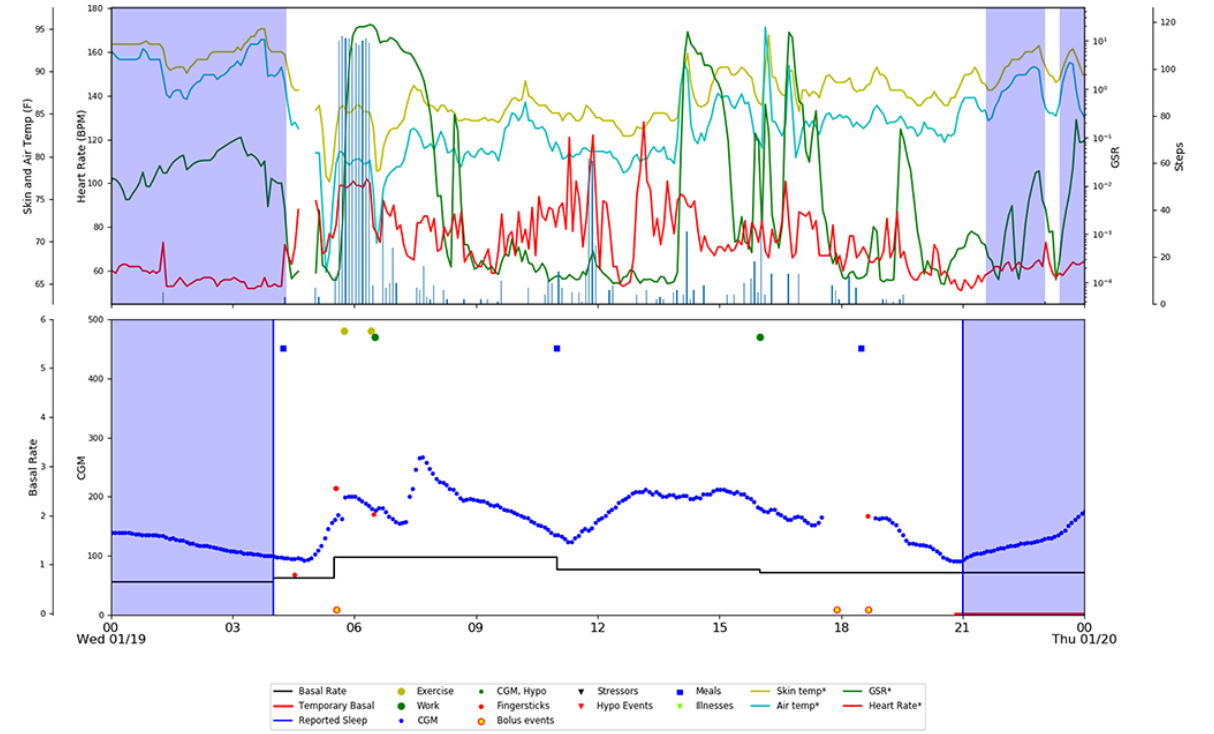


Figure 3: 24-hour measurements of one patient from the OhioT1DM Viewer [MB20].

5.1.3 MIMIC-IV:

The proposed model aims to generalise beyond diabetes forecasting to other medical applications, such as predicting disease progression in HFpEF patients. The MIMIC-IV dataset [JBS+23] contains de-identified electronic health records (EHRs) of ICU patients, including vital signs (heart rate, blood pressure, oxygen saturation, etc.), medication records, and laboratory results. For this study, a subset of MIMIC-IV focusing on cardiovascular patients is used. The target variable, mortality risk, was divided into two groups: death within 30 days and death within one year. The exogenous variables include vital signs and laboratory values, while sex and comorbidities are used to split the data into multiple cohorts. By analysing the data in different cohorts, it is possible to get more specific and accurate counterfactual interventions.

5.1.3.1 Preprocessing

Following a preceding study [SBA+24], the International Classification of Diseases (ICD) codes were used for the initial preprocessing. The ICD-codes are a standardised international classification system used for the categorisation and encoding of diseases, symptoms, and associated health-related conditions. By using the appropriate ICD-9 and ICD-10 codes, as outlined in Table 1, the hospital admissions involving patients aged ≥ 18 with HFpEF as a primary diagnosis have been identified. Given that the diagnosis was based on ICD codes, the clinical notes were analysed in order to validate the selection of patients. This was achieved through filtering the clinical notes on mentions of the left ventricular ejection fraction (LVEF) value, with a value of 50 and above being counted as a normal LVEF value. Some clinical notes only mentioned a normal or preserved LVEF value with out a measured LVEF. These were also counted as normal LVEF values. Table 1 also shows the number of hospital admissions per diagnosis. The study sample consisted of

16122 individual hospitalisations with a suspected diagnosis of HFpEF. We had access to clinical notes for 11720 (72.7%) hospital admissions of which 4458 (38%) had an LVEF measurement reported. Of these, 3798 (85.2%) had an LVEF value $\geq 50\%$, and 400 (9%) had an LVEF $< 50\%$. An additional 260 (5.8%) admissions had mention of a normal or preserved LVEF. For these 4058 admissions, vital signs and laboratory values were available for 2432 (60%). It should be noted that in this instance, only the most recent admission of a patient who had previously been admitted with a similar diagnosis was taken into consideration. Prior admissions were incorporated into the analysis as a co-morbidity, and after adding all laboratory values, vital signs and comorbidities, resulting in 2113 hospital admissions of 1845 unique patients.

Diagnosis	**ICD Code**	**Frequency**
Unspecified diastolic (congestive) heart failure	I5030	38
Diastolic heart failure, unspecified	42830	69
Acute diastolic (congestive) heart failure	I5031	84
Acute diastolic heart failure	42381	180
Chronic diastolic (congestive) heart failure	I5032	256
Chronic diastolic heart failure	42382	453
Acute on chronic diastolic (congestive) heart failure	I5033	426
Acute on chronic diastolic heart failure	42833	623

Table 1: The corresponding Diagnosis and ICD-9 and ICD-10 codes for Heart Failure with preserved Ejection Fraction.

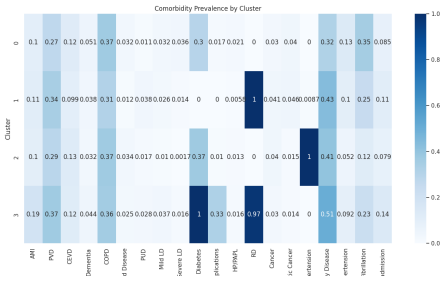
The extracted features are listed in table 2, split into four categories. These are the targets (Death within 30 days, Death within 1 year) as well as vital signs and laboratory values, comorbidities, and sex of the patients. Prior admission was included in the list of comorbidities, since it is here used as a comorbidity for the clustering and not the forecasting.

Feature types	**Specifics**	**Occurrences**
Targets		
	Death within 30 days	136
	Death within 1 year	150
Vital signs and laboratory values		
	BMI	1845
	Heart Rate	1845
	SpO2	1845
	Diastolic BP	1837
	Systolic BP	1837
	Temperature	1829
	Creatinine	1825
	Sodium	1825
	Bicarbonate	1823
	Hemoglobin	1814
	WBC Count	1814
	Platelet Count	1812
	Troponin	958
	NT-proBNP	76
Comorbidities		
	Diabetes	798
	RD	768
	Coronary Artery Disease	761
	COPD	658
	Hypertension	599
	PVD	577
	Atrial Fibrillation	424
	AMI	230
	CEVD	218
	Prior Admission	188
	Pulmonary Hypertension	170
	Diabetes + Complications	156
	Dementia	75
	Cancer	65
	Metastatic Cancer	50
	Rheumatoid Disease	50
	Mild LD	46
	PUD	40
	Moderate/Severe LD	30
	HP/PAPL	27
Sex		
	Male	773
	Female	1072

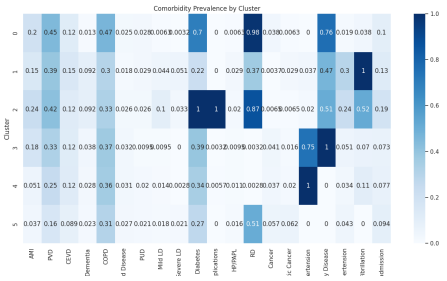
Table 2: All features of the MIMIC data divided by feature types: target, vital signs and laboratory values, comorbidities, and sex.

5.1.3.2 Clustering

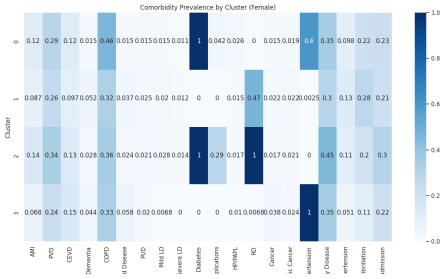
The MIMIC data does not only include a wide range of features, but also a great number of comorbidities. Since these comorbidities can influence the chances of HFpEF, it is important to include these in the analysis. In this study, the comorbidities were used to cluster the patients, to get more specific counterfactuals. For this, the data was split according to patients having similar comorbidities to get factual cohorts. Another split was made by dividing the data on sex, since HFpEF is more prevalent in female patients and might need different treatment. Figure 4 shows the prevalence of different comorbidities of the MIMIC patients in the clusters, divided into four sub figures depending on the clustering coefficient and the sex.



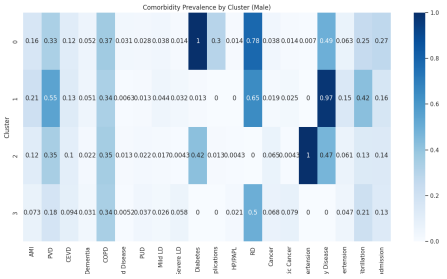
(a) Comorbidities in 4 clusters.



(b) Comorbidities in 6 clusters.



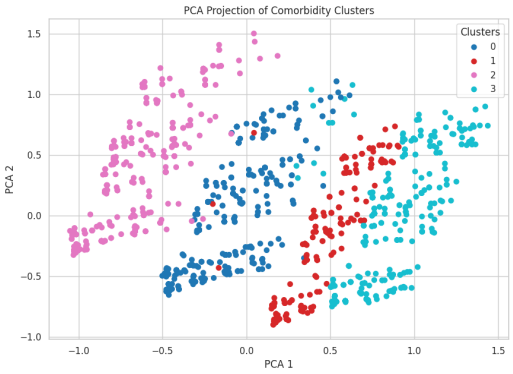
(c) Comorbidities for female patients in 4 clusters.



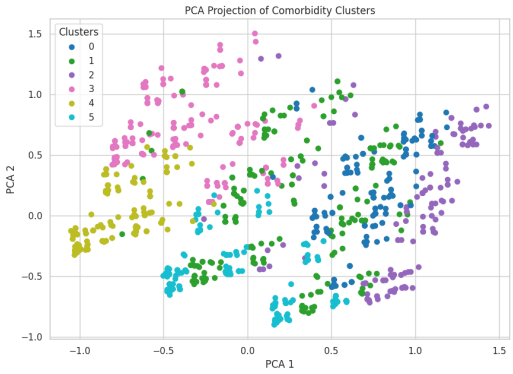
(d) Comorbidities for male patients in 4 clusters.

Figure 4: Prevalence of different comorbidities of the MIMIC patients in the clusters. There are 4 subgroups, clustering by comorbidity with $k = 4$, clustering by comorbidity with $k = 6$, clustering only the female patients with $k = 4$ and clustering only the male patients with $k = 4$.

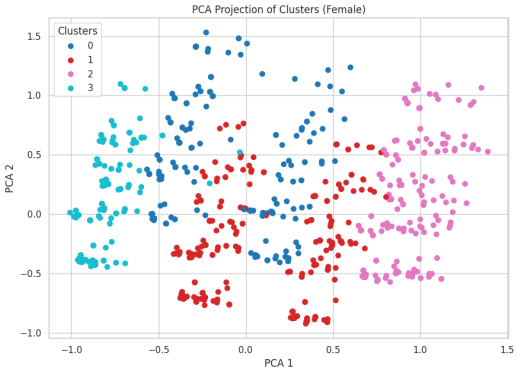
Figure 5 shows the Principal Component Analysis (PCA) of the different comorbidities, again divided into the same subgroups.



(a) PCA clustering in 4 clusters.



(b) PCA clustering in 6 clusters.



(c) PCA clustering for female patients in 4 clusters.

(d) PCA clustering for male patients in 4 clusters.

Figure 5: Principal component analysis (PCA) clustering of different comorbidities of the MIMIC patients. There are 4 subgroups, clustering by comorbidity with $k = 4$, clustering by comorbidity with $k = 6$, clustering only the female patients with $k = 4$ and clustering only the male patients with $k = 4$.

Table 3 shows the prevalence of the different comorbidities of the four general clusters, that were later used for the prediction task.

Comorbidity	**Cluster 0**	**Cluster 1**	**Cluster 2**	**Cluster 3**
COPD	37.0	31.3	36.9	35.9
Atrial Fibrillation	35.3	24.9	12.1	23.0
Coronary Artery Disease	31.7	43.5	40.6	50.7
Diabetes	30.0	0.0	37.4	100.0
PVD	27.0	34.5	28.5	37.1
Pulmonary Hypertension	13.4	10.4	5.2	9.2
CEVD	12.1	9.9	12.6	12.0
AMI	10.0	11.3	10.2	19.1
Prior Admission	8.5	11.3	7.9	14.3
Dementia	5.1	3.8	3.2	4.4
Metastatic Cancer	4.0	4.6	1.5	1.4
Moderate/Severe LD	3.6	1.4	0.2	1.6
Mild LD	3.2	2.6	1.0	3.7
Rheumatoid Disease	3.2	1.2	3.4	2.5
Cancer	3.0	4.1	4.0	3.0
HP/PAPL	2.1	0.6	1.3	1.6
Diabetes + Complications	1.7	0.0	1.0	32.7
PUD	1.1	3.8	1.7	2.8
RD	0.0	100.0	0.0	97.5
Hypertension	0.0	0.9	100.0	0.0

Table 3: Prevalence of Comorbidities by Cluster in %

5.2 Experiments

The model is split into two main parts, the forecasting and the counterfactual generation. Initially a multivariate forecasting model is used to make a first forecast for both the target variable and the exogenous variables. This forecast is then used for the second part, where different regression models are used to change the exogenous and target variable to get the desired outcome. For the multivariate forecasting we used GRU and N-BEATS and for the counterfactual generation we used four different kinds of models, like a statistical (SARIMAX), a regression based (OLS), and two different deep learning based (GRU and N-BEATS) models.

5.2.1 Multivariate Forecasting

For the multivariate forecasting task, two deep learning architectures were implemented and evaluated: (1) a 2-layer Gated Recurrent Unit (GRU) model, and (2) a 4-layer Neural Basis Expansion Analysis for Time Series (N-BEATS) model. The GRU model consisted of two stacked layers, each comprising 200 hidden units, followed by a linear output layer to produce the forecast. The N-BEATS model was configured with four fully connected layers, integrating both backcast and forecast blocks, and concluded with a linear output head to reconstruct future values. This design allows the model to capture both short-term patterns and longer-range temporal dependencies effectively. For the application of the

method to the HFpEF data, a multi-head approach was used, since the target is binary, while the exogenous data is continuous. By training the target with a sigmoid layer and the exogenous data with a linear layer, the model was able to work with this kind of data as well.
To prevent overfitting, early stopping was employed in both architectures with a patience of 10 epochs to ensure enough passes through the training data, and a fixed learning rate of 0.001 was used for training. The models were trained on varying back horizons and forecast horizons, allowing a thorough investigation of how different historical contexts influenced future prediction accuracy. For the back horizon, 24 timesteps and 12 timesteps were chosen for the OhioT1DM data, which correspond to 2 hours or 1 hour respectively, with 6 timesteps being a prediction window of 30 minutes. For the SimGlucose dataset, a similar prediction window was chosen, with 40 timesteps being 2 hours, and the forecasting window of 5 timesteps also representing 30 minutes. The HFpEF data from the MIMIC-IV dataset, was restricted to a 24-hour window, to allow for a magnitude of vital signs and laboratory data. In this window a back horizon of 12 hours was chosen with a horizon of 6 hours.
Model performance was evaluated using two standard error metrics: symmetric Mean Absolute Percentage Error (sMAPE) and Root Mean Squared Error (RMSE). These metrics were selected to balance sensitivity to outliers (RMSE) with scale-invariant accuracy (sMAPE), ensuring a robust evaluation of predictive performance. Lower values for both metrics indicated better forecasting quality. Model performance was assessed using four complementary metrics: Symmetric Mean Absolute Percentage Error (sMAPE), Root Mean Squared Error (RMSE), Accuracy, and F1 Score.

- Symmetric Mean Absolute Percentage Error (sMAPE): sMAPE measures the relative accuracy of forecasts by comparing the absolute difference between predicted and actual values to their average magnitude. It is scale-invariant, making it suitable for comparing performance across different ranges. Lower sMAPE values indicate higher predictive accuracy.

- Root Mean Squared Error (RMSE): RMSE quantifies the square root of the average squared differences between predicted and actual values. It is sensitive to large errors (outliers) and provides insight into the magnitude of typical prediction errors. Lower RMSE values reflect better model performance.

- Accuracy: Accuracy represents the proportion of correctly predicted instances (both positives and negatives) out of all predictions. It provides a straightforward measure of overall model correctness, particularly relevant in binary classification tasks.

- F1 Score: The F1 score is the harmonic mean of precision and recall. It balances the trade-off between false positives and false negatives, offering a robust evaluation for imbalanced classification scenarios. A higher F1 score indicates better balance and reliability in predicting the positive class.

These four metrics were chosen to provide a comprehensive evaluation framework. To assess the continuos forecasting performance as necessary for the OhioT1DM and SimGlucose datasets, sMAPE and RMSE are used. Accuracy and F1 Score specifically evaluate classification performance, particularly important for binary outcomes as found in the

MIMIC dataset. In all cases, lower sMAPE and RMSE values, along with higher Accuracy and F1 scores, indicate better model quality.
After thorough evaluation across multiple configurations, the model with the lowest average error on all validation sets was selected as the baseline for subsequent counterfactual analysis. This selection ensured that counterfactual explanations were derived from the most reliable and accurate forecasting model available.

5.2.2 Counterfactual Generation

To evaluate our counterfactual generation approach, we conducted experiments on two clinical prediction tasks: glycaemic forecasting for type 1 diabetes and survival prediction for Heart Failure with Preserved Ejection Fraction (HFpEF).
The models used for the counterfactual generation span both traditional statistical methods and deep learning architectures, such as SARIMAX, OLS, GRU, and N-BEATS. Each model was adapted to support input optimisation using either gradient-based backpropagation (for GRU and N-BEATS) or finite-difference approximation (for SARIMAX and OLS).

5.2.2.1 SARIMAX

We used the Seasonal Autoregressive Integrated Moving Average with Exogenous Variables (SARIMAX) [SAR] model to capture seasonal patterns and time dependencies in glucose data, while including outside factors like insulin doses and carb intake. For each data sample, we fit a SARIMAX with autoregressive order $p = 1$, differencing degree $d = 0$, and moving average order $q = 0$ using maximum likelihood estimation via the statsmodels library. When generating counterfactuals, we adjusted input exogenous variables iteratively to keep the predicted glucose within set bounds or push the mortality target towards survival. We calculated gradients approximately using finite differences on the forecast function of the fitted model. To make changes more realistic, we applied custom weighting for the glucose prediction focused on meal-related features during optimisation.

5.2.2.2 OLS

We also applied Ordinary Least Squares (OLS) [OLS] regression similarly to SARIMAX, modeling the linear relationship between lagged inputs (including interaction terms like insulin and carbs) and glucose levels. Each sample had its own OLS model. Counterfactual optimisation followed the same process as SARIMAX, by using finite difference gradients to guide input changes, constrained by feature-specific limits and for glucose prediction weighted towards meal-related inputs to keep the results physiologically plausible.

5.2.2.3 GRU

The Gated Recurrent Unit (GRU) [GRU] model had two recurrent layers with 100 units each, followed by a dense layer outputting predictions. It was trained on multiple input signals like carbs, insulin, and physical activity, with dropout of 0.2 to avoid overfitting. Training used the Adam optimizer with a learning rate of 0.001, and early stopping based on validation loss. For counterfactuals, gradient perturbation to get exact gradients

of outputs with respect to inputs. This allowed efficient gradient-based optimisation of input features under clipping limits and with custom weighting focused on meal times. We stopped the optimisation once predicted glucose stayed within desired ranges or after a max number of steps.

5.2.2.4 N-BEATS

N-BEATS [N-B] is a model that can explicitly incorporate external inputs like carbs, insulin, and activity into its trend and seasonal components. It used fully connected residual blocks with 128 hidden units and was trained to minimise mean absolute error. Training used the Adam optimizer with a 0.001 learning rate. Like the GRU, counterfactual generation was done with gradient-based optimisation, constrained by clipping and for glucose prediction guided by meal-focused weighting. This allowed the model to capture complex non-linear patterns while making it easier to interpret how exogenous inputs affect predictions during counterfactual analysis.

5.2.2.5 Parameters for the Diabetes Data (OhioT1DM, SimGlucose)

For the diabetes datasets, test samples were split into subgroups based on predicted glucose levels over a 30-minute forecast horizon:

- Hyperglycemia group: samples with any predicted value $\mathbf{y}_{n+i} \geq 180$ mg/dL.
- Hypoglycemia group: samples with any predicted value $\mathbf{y}_{n+i} \leq 70$ mg/dL.

For each group, we randomly selected 100 test samples for counterfactual generation. The same process was repeated across datasets with dataset-specific bounds: for SimGlucose, we used $[80, 160]$ mg/dL, since it is an artificial dataset; for OhioT1DM, $[70, 180]$ mg/dL.

5.2.2.6 Parameters for the HFpEF Survival Prediction

In the HFpEF dataset, binary predictions over a 24-step horizon were used to classify samples as high-risk (non-survival) if more than 50% of the time steps predicted a non-survival event (label 1). Due to limited availability of non-survival cases, it was not possible to select 100 samples from each group. Instead, samples were selected as follows:

- Non-survival group: $\sum_{i=1}^{t} \mathbf{y}_{n+i} \geq \frac{t}{2}$, all available samples meeting this criterion were included.
- Survival group: $\sum_{i=1}^{t} \mathbf{y}_{n+i} < \frac{t}{2}$, randomly selected, for 100 samples in total.

5.2.2.7 Counterfactual Generation Procedure

We followed a consistent three-step pipeline:

- Bound Generation: For continuous targets (OhioT1DM and SimGlucose), we constructed soft upper and lower bounds using a polynomial interpolation function transitioning from the current forecast to a desired safe target over a fixed number of time steps ($S = 24$ for OhioT1DM, $S = 20$ for SimGlucose). For HFpEF, since the target is binary, no such bounds were generated.

- Optimisation: Exogenous inputs (for example carbohydrate intake, insulin dose, clinical variables) were perturbed to minimise a custom loss function. For the differentiable models (GRU, N-BEATS), gradients were obtained via backpropagation. For non-differentiable models (SARIMAX, OLS), finite-difference approximations were used. For HFpEF, optimisation aimed to flip the predicted class while minimising input changes, without any bound constraints.

- Loss: The loss balanced moving predictions toward the target class or within bounds for continuous targets and minimising the magnitude of input perturbations, where feature-specific weights emphasised intervention-relevant variables (for example insulin, carbohydrates for glucose, and clinical features for HFpEF) to encourage realistic and meaningful counterfactuals.

All inputs and targets were scaled using per-patient normalisers to preserve physiological interpretability. Historical context, such as meal history, past glucose trends, was retained to ensure the counterfactuals remained plausible within each individual's recent clinical trajectory.

5.3 Evaluation Metrics

To evaluate the quality of the counterfactual interventions, multiple evaluation metrics were implemented. First, traditional forecasting performance was measured using Root Mean Squared Error (RMSE) and Symmetric Mean Absolute Percentage Error (sMAPE), giving a quantitative assessment of prediction accuracy across the forecast horizon for the diabetes data. For the HFpEF data, Accuracy and F1 score was used, to work with the binary target data. For the evaluation of the generated counterfactuals, several additional metrics were introduced. These have been applied to ensure that the newly generated data for the exogenous variables is realistic, plausible and applicable.

5.3.1 Magnitude and Sparsity of Exogenous Variable Changes

The average value of change calculates the mean absolute difference between original and counterfactual exogenous variables, averaged over all samples, time steps, and features. It reflects the overall magnitude of adjustments required. By determining the proportion of exogenous input entries that underwent modification, we assess the sparsity of the intervention, providing insight into whether changes are targeted or widespread across the input profile.

5.3.2 Severity of Change

The severity metric measures changes normalised by the original variability of each exogenous feature, enabling assessment of whether modifications are within a physiologically reasonable range. For this, we employ Local Outlier Factor (LOF) scores to identify whether the counterfactual exogenous profiles exhibit atypical or extreme deviations from the original data distribution, highlighting potential concerns about the realism of the suggested interventions.

5.3.3 Fitting of Predictions within Bounds

The extent to which counterfactual blood glucose predictions fall within the desired range is quantified by the percentage of time points inside the bounds and the root mean squared error (RMSE) of any violations outside these limits. This metric directly evaluates the success of the optimisation in achieving clinically safe target blood glucose levels.

5.3.4 Comparison to Healthy Exogenous Profiles

To ensure biological and clinical plausibility, the optimised exogenous variables are compared against reference exogenous profiles derived from healthy patient data. Both raw and normalised Euclidean distances, as well as cosine similarity metrics, quantify the alignment of counterfactual inputs with typical healthy patterns, supporting qualitative evaluation of intervention realism.

5.3.5 Interpretation

This multi-faceted evaluation framework provides a detailed and nuanced understanding of how effectively the optimisation generates actionable and safe counterfactual interventions. The ideal outcomes achieve a careful balance between several important, sometimes competing, goals. First, they bring predicted blood glucose levels within the desired target range without causing the forecasts to deviate excessively from the original values, ensuring the adjustments remain realistic and consistent with the patient's typical glucose patterns. At the same time, the changes to exogenous variables, such as insulin dosages or carbohydrate intake, are kept as small and infrequent as possible, since large or frequent modifications might be impractical or even unsafe in real-world settings. Additionally, the interventions maintain physiological plausibility by limiting the intensity of changes and avoiding outlier values that could indicate unrealistic or extreme behaviours. Clinical safety is also prioritised by strictly enforcing blood glucose predictions to stay within safe bounds, reducing the risk of hypo- or hyperglycaemia. Finally, the adjusted exogenous inputs are aligned closely with patterns observed in healthy patients, making the suggested interventions not only effective but also clinically meaningful and believable. By assessing performance across all these dimensions, this evaluation approach highlights both the strengths and limitations of the counterfactual generation process, guiding future improvements and enhancing its practical value in diabetes management.

6 Results

6.1 Multivariate Forecasting

To evaluate the performance of the multivariate forecasting and to ensure a realistic basis for counterfactual generation, both the GRU and the N-BEATS models were tested across multiple datasets and forecasting configurations. Models were trained on historical segments of the data and evaluated on subsequent future segments, allowing for a robust comparison between predicted and actual values. The forecasting quality for the two diabetes datasets was assessed using two standard metrics: Symmetric Mean Absolute Percentage Error (sMAPE) and Root Mean Squared Error (RMSE). sMAPE was chosen due to its scale-invariant properties, allowing for fair comparison across variables with different magnitudes. RMSE was included to capture sensitivity to larger errors, penalising substantial deviations between predictions and true values. Together, these metrics provide a balanced view of model performance, addressing both relative and absolute error considerations. For the MIMIC-IV dataset, which involves binary survival prediction (death vs. survival), classification metrics were more appropriate given the nature of the target variable. Specifically, Accuracy and F1-Score were employed. Accuracy provides a general measure of how often the model's predictions match the true outcomes. However, due to the severe class imbalance in the MIMIC-IV dataset, where death events are comparatively rare, F1-Score was also reported to better capture the trade-off between precision and recall for the minority class. This combination of metrics ensures both the overall predictive reliability and the ability to identify critical but infrequent events are adequately evaluated.

Dataset	Back Horizon	Horizon	Model	sMAPE	RMSE
OhioT1DM	12	6	GRU	6.2816	14.8471
			N-BEATS	**6.0328**	**14.1500**
	24	6	GRU	6.1983	14.6856
			N-BEATS	**5.8003**	**13.7677**
SimGlucose	20	5	GRU	0.3792	1.6770
			N-BEATS	**0.3867**	**0.8271**
	40	10	GRU	1.5423	5.5572
			N-BEATS	**1.4303**	**3.0461**

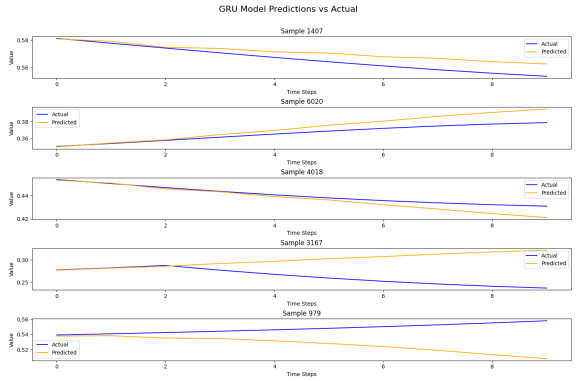
Table 4: Multivariate forecasting training metrics.

Table 4 summarises the results across the two diabetes datasets, prediction horizons, and back horizons, the results for the HFpEF dataset can be found in section 6.1.3. For the OhioT1DM dataset, N-BEATS consistently outperformed the GRU model in both sMAPE and RMSE across all configurations. Notably, with a back horizon of 12 and a forecast horizon of 6, N-BEATS achieved a sMAPE of 6.03 compared to GRU's 6.28, and a lower RMSE of 14.15 versus 14.85 for GRU. The same pattern was observed for back horizon 6 and 24, with N-BEATS slightly outperforming GRU in all cases. For the SimGlucose dataset, N-BEATS again generally showed improved forecasting performance. While GRU slightly outperformed N-BEATS in sMAPE for the 20-step back, 5-step horizon configuration (0.3792 vs. 0.3867), N-BEATS achieved a substantially lower RMSE (0.8271 vs. 1.6770), indicating more accurate absolute predictions. This performance gap widened in

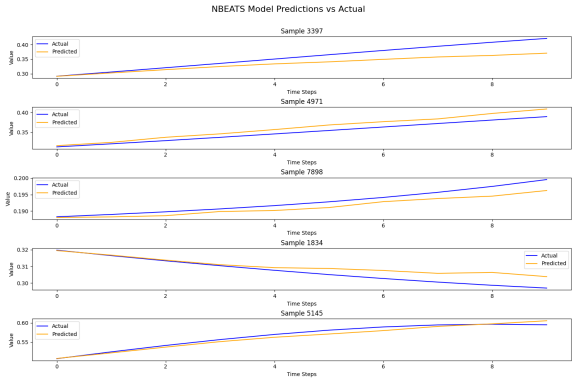
the 40-back, 10-horizon configuration, where N-BEATS clearly surpassed GRU on both metrics.
Overall, these results demonstrate that the N-BEATS model generally offers superior performance across datasets and configurations, especially in terms of RMSE, making it a strong candidate for generating accurate multivariate forecasts and serving as a foundation for counterfactual analysis.
To complement the quantitative evaluation of the multivariate forecasting models, several visualisations were created to illustrate how well the N-BEATS and GRU models predict future values compared to actual ones. Figure 6 and 7 compare the predicted time series against the true values for randomly selected test samples from the datasets. Each subplot presents a single test sample, with time steps on the x-axis and the corresponding variable value on the y-axis. The actual values are shown in blue, while the predicted values are shown in orange. These plots offer a qualitative insight into the temporal alignment and amplitude accuracy of the forecasts, beyond what is captured by metrics like sMAPE or RMSE.

6.1.1 SimGlucose

For a back horizon of 40 timesteps and a horizon of 10 timesteps, the forecast horizon is increased, making the prediction task more difficult. Figure 6 shows the N-BEATS model's ability to handle long-range dependencies. Although small prediction lags and amplitude mismatches occur, the model captures the overall progression well, preserving directionality in most sequences. When looking at the GRU predictions under the same setup, the model struggles a bit more with extended forecasts, often showing divergence from actual trajectories, particularly toward the final time steps. These deviations are consistent with the slightly higher RMSE reported for this configuration. The results for the back horizon of 20 timesteps and horizon of 5 timesteps are quite similar, the corresponding figure can be found in Appendix A.



(a) Results of the multivariate forecasting using GRU.

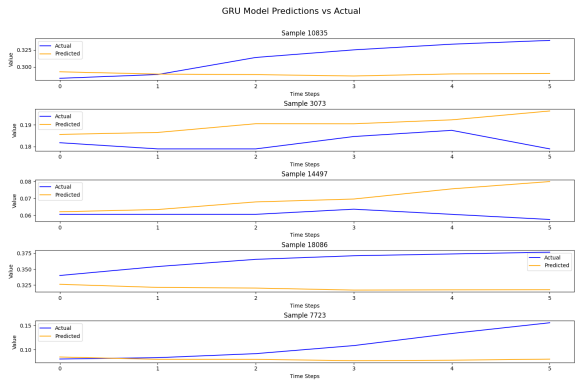


(b) Results of the multivariate forecasting using N-BEATS.

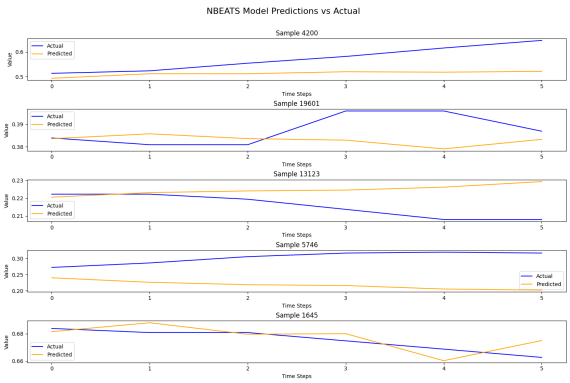
Figure 6: Results of the multivariate forecasting for the SimGlucose dataset with back horizon = 40 and forecast horizon = 10, showing the accuracy of the forecasting.

6.1.2 OhioT1DM

Figure 7 depicts the performance of N-BEATS and GRU, with a back horizon of 24 timesteps and 6 future timesteps, The figure illustrates that N-BEATS maintains robust trend prediction, with only modest lag or dampening effects in some samples. The model adapts well to both upward and downward trajectories. The GRU model performs more variably under this longer back horizon. While the general trajectory is still often captured, the forecasts occasionally overreact or smooth out variations, resulting in lower precision at the end of the forecast horizon.



(a) Results of the multivariate forecasting using GRU.



(b) Results of the multivariate forecasting using N-BEATS.

Figure 7: Results of the multivariate forecasting for the OhioT1DM dataset with back horizon = 24 and forecast horizon = 6, showing the accuracy of the forecasting.

6.1.3 MIMIC-IV

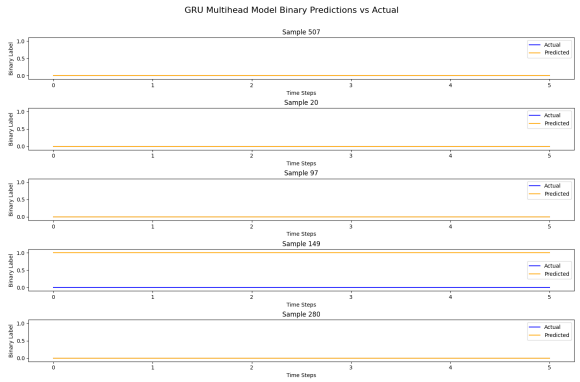
We evaluated the classification performance of GRU and N-BEATS models on the MIMIC-IV dataset, focusing on female patients with heart failure with preserved ejection fraction (HFpEF). Class-wise metrics for both 30-day and 1-year mortality prediction tasks are shown in Table 27.

While both models achieved high overall accuracy, over 91% for 30-day and 90% for 1-year predictions, they failed to correctly identify any instances of the non-survival class. Precision, recall, and F1-score for this minority class were 0.00 across all settings. This indicates a strong bias toward the majority survival class, likely due to class imbalance. These results highlight a key limitation, where high overall accuracy does not reflect clinically meaningful performance when models systematically miss rare but critical outcomes. Alternative strategies, such as resampling, class weighting, or specialized loss functions, may be needed to improve minority class detection in imbalanced clinical datasets.

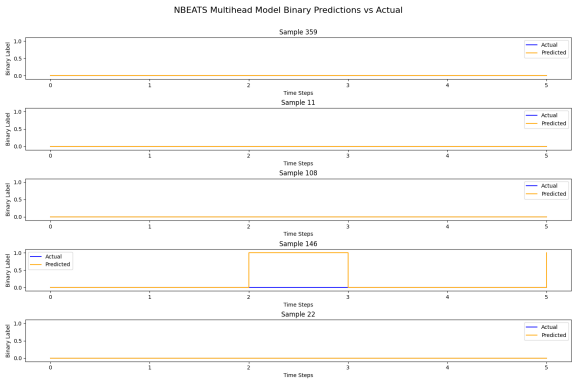
Model	Target	Class	Precision	Recall	F1-Score
GRU	30 days	Survived	0.94	1.00	0.97
		Died	0.00	0.00	0.00
		Overall Accuracy		0.9414	
	1 year	Survived	0.94	1.00	0.97
		Died	0.00	0.00	0.00
		Overall Accuracy		0.9369	
N-BEATS	30 days	Survived	0.94	1.00	0.97
		Died	0.00	0.00	0.00
		Overall Accuracy		0.9398	
	1 year	Survived	0.94	1.00	0.97
		Died	0.00	0.00	0.00
		Overall Accuracy		0.9362	

Table 5: Classification Metrics for GRU and N-BEATS Models for Female Patients

Figure 8 illustrates the predicted 30-day mortality over time using a back horizon of 12 timesteps and a forecast window of 6. Both GRU and N-BEATS show a strong bias toward the majority (survival) class. Even when true labels are death, the model outputs remain consistently zero (survival), indicating a lack of sensitivity to the non-survival class. Predictions are largely flat across timesteps and do not adapt to actual class transitions, further showing how the model does not react properly to critical events.
This behaviour confirms the impact of class imbalance: although accuracy appears to be high, the models do not perform well in identifying patients at risk of mortality, which is precisely the group where accurate predictions matter most.



(a) Results of the multivariate forecasting using GRU.



(b) Results of the multivariate forecasting using N-BEATS.

Figure 8: Results of the multivariate forecasting for the MIMIC-IV dataset with 30-day mortality as target variable, back horizon = 12, and forecast horizon = 6, showing the accuracy of the forecasting.

Across all clusters and subgroups, the results were quite similar, with the N-BEATS model usually outperforming the GRU model. The complete results can be found in Appendix B.

6.2 Counterfactuals

6.2.1 OhioT1DM

We evaluated the performance of four models, (GRU, SARIMAX, OLS, and N-BEATS) for generating counterfactuals, using a comprehensive set of metrics encompassing predictive accuracy, constraint adherence, and intervention efficiency. The results can be found in Table 6.

Metric	GRU	SARIMAX	OLS	N-BEATS
Forecast Accuracy				
MAE	**50.67**	111.25	116.61	52.41
RMSE	**50.94**	111.48	116.85	56.39
Max Deviation	**203.47**	812.03	797.12	234.97
MAPE (%)	53.35	116.53	123.57	**40.65**
Constraint Adherence				
In-Bound (%)	**43.5**	2.7	5.7	0.8
Mean Violation	**5.59**	87.42	86.26	80.60
Total Violation Area	**2944.51**	52421.28	51689.86	48170.33
Intervention Cost (Mean % Change)				
Basal Insulin	**6.3**	211.7	200.5	16.7
Bolus Insulin	**6.2**	210.9	199.6	14.3
Carbohydrates	**5.8**	211.8	200.5	15.0
Exercise Intensity	**6.3**	210.7	199.4	15.4

Table 6: Comparison of the Counterfactual Generation Methods across Evaluation Metrics

Regarding forecasting accuracy, GRU achieved the best performance in terms of Mean Absolute Error (MAE = 50.67), Root Mean Squared Error (RMSE = 50.94), and maximum deviation (203.47), indicating its strength in both average and extreme prediction errors. N-BEATS, while slightly behind in MAE and RMSE, achieved the lowest Mean Absolute Percentage Error (MAPE = 40.65), suggesting better relative error control compared to the others. SARIMAX and OLS performed similarly and worse than both GRU and N-BEATS across all accuracy metrics, with SARIMAX showing particularly high errors (MAE = 111.25, RMSE = 111.48, MAPE = 116.53). To assess the feasibility and domain realism of the counterfactuals, we examined constraint adherence metrics, including the proportion of predictions remaining within bounds, mean violation, and total violation area. GRU again outperformed all other models, with the highest in-bound percentage (43.5%) and the lowest mean violation (5.59) and total violation area (2944.51), indicating superior constraint compliance. In contrast, N-BEATS had the poorest performance in this regard, with only 0.8% of predictions within bounds and a total violation area of 48170.33. SARIMAX and OLS showed slightly better adherence than N-BEATS but

were still significantly worse than GRU. We further evaluated the intervention cost by measuring the mean percentage change required in key input variables. GRU required the smallest changes across all intervention variables (ranging from 5.8% to 6.3%), highlighting its ability to generate realistic and efficient counterfactuals. N-BEATS incurred moderate intervention costs (14.3%–16.7%), whereas SARIMAX and OLS demanded excessive changes (approximately 200%–211%), rendering their counterfactuals largely impractical for real-world application. In summary, GRU achieved the best balance across predictive accuracy, constraint adherence, and minimal intervention, making it the most suitable model for counterfactual generation in this setting. While N-BEATS demonstrated relatively strong predictive performance, its poor constraint adherence and higher intervention costs limit its applicability. SARIMAX and OLS consistently underperformed across all evaluated dimensions.

Figure 9 and 10 present an example of generated counterfactuals for the OhioT1DM dataset. They show a short sequence of 6 timesteps where the predicted blood glucose level is approximately 140 mg/dL, which is higher than the very low original range of 80 to 85 mg/dL. This new blood glucose level falls within the predefined desired bounds, indicating a plausible improvement. Notably, the generated exogenous variables, such as insulin doses, carbohydrate intake, and exercise intensity, differ considerably from the original values. These differences suggest actionable changes that may help achieve improved glycaemic control.

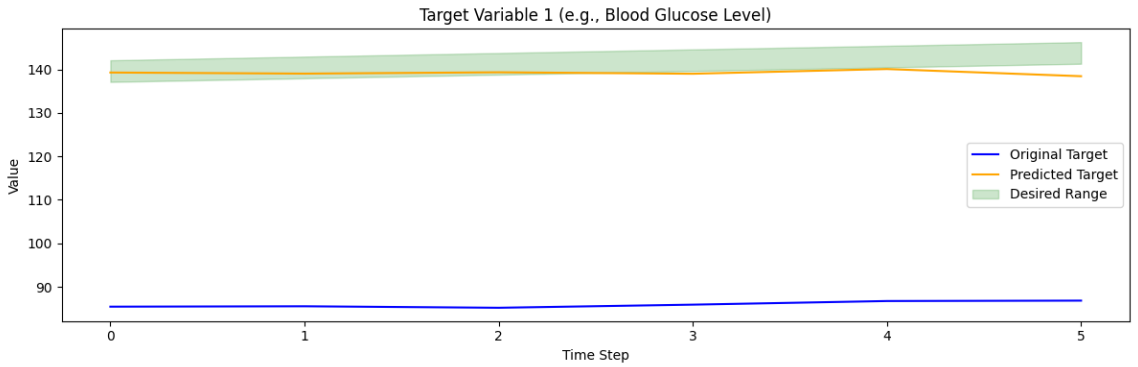


Figure 9: Example of the target generated using GRU for the OhioT1DM dataset. Top: Blood glucose levels for the original and counterfactual samples (original in blue, counterfactual in yellow, bounds in green).

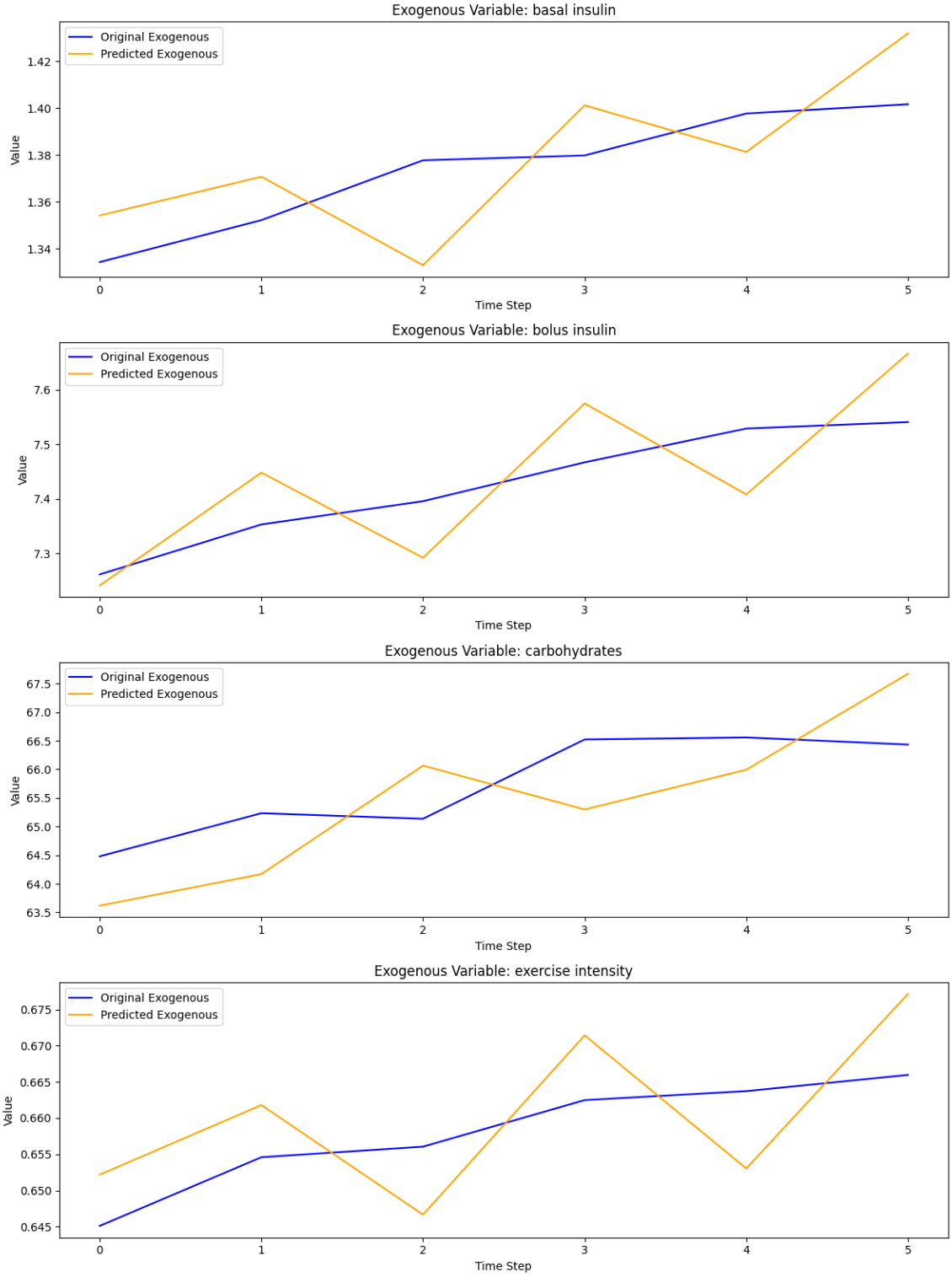


Figure 10: Example of the counterfactuals generated using GRU for the OhioT1DM dataset. Comparison of exogenous variables (original in blue, counterfactual in yellow).

6.2.2 SimGlucose

Table 7 presents a comparison of the four counterfactual generation methods, GRU, SARIMAX, OLS, and N-BEATS, evaluated for forecast accuracy, and constraint adherence. Among all models, SARIMAX consistently achieved the best predictive performance

across all accuracy metrics. It obtained the lowest Mean Absolute Error (MAE) of 3.33 and Root Mean Squared Error (RMSE) of 3.58, significantly outperforming the next best model. Additionally, SARIMAX had the lowest Maximum Deviation (15.36) and Mean Absolute Percentage Error (MAPE) at just 2.98%. In contrast, the OLS and N-BEATS models showed substantially higher errors, with MAEs above 68 and MAPEs exceeding 50%, suggesting poor forecasting fidelity. The GRU model showed moderate performance with an MAE of 29.15 and MAPE of 28.19%. Constraint adherence was generally poor across all models, with none achieving substantial in-bound prediction rates. SARIMAX achieved a minimal in-bound percentage of 0.8%, while all other models had 0.0%. Despite this, GRU had a relatively lower mean violation (31.80) compared to SARIMAX (44.30), though SARIMAX yielded the highest total violation area (44,266.23), suggesting widespread deviations from desired bounds. N-BEATS demonstrated the lowest total violation area (17,257.68), indicating it may produce more conservative but consistently bounded outputs, despite its poor accuracy. Overall, SARIMAX stands out in terms of forecast accuracy, but its high constraint violations leads to the assumption that the counterfactuals are not realistic. GRU provides a more balanced trade-off between moderate accuracy and moderate constraint violations, while OLS underperforms in both dimensions. N-BEATS offers conservative outputs with smaller violations but lacks the predictive precision required for reliable counterfactual generation.

Metric	**GRU**	**SARIMAX**	**OLS**	**N-BEATS**
Forecast Accuracy				
MAE	29.15	**3.33**	72.12	68.47
RMSE	29.39	**3.58**	72.15	68.72
Max Deviation	76.91	**15.36**	129.43	134.68
MAPE (%)	28.19	**2.98**	57.65	54.25
Constraint Adherence				
In-Bound (%)	0.0%	**0.8%**	0.0%	0.0%
Mean Violation	31.80	**44.30**	21.09	17.26
Total Violation Area	31796.35	44266.23	21091.24	**17257.68**

Table 7: Comparison of Counterfactual Generation Methods Across Evaluation Metrics

Figure 11 presents the results of counterfactual forecasting for the SimGlucose dataset using a GRU-based multivariate forecasting model. The aim was to identify alternative exogenous variable trajectories, specifically carbohydrate intake and insulin administration, that lead to improved blood glucose levels within a clinically desirable range. The figure shows that the model made quite drastic changes to the exogenous variables, while the predicted target is not within the bounds, but closer than it originally was. Both counterfactuals for carbohydrate intake and insulin administration are noticeably reduced compared to the original values. These reductions can also be found in the blood glucose levels, which suggests that the model does identify the cohesion between the exogenous variables and the target.

Overall, these results demonstrate that the model has difficulties working with the more streamlined and artificial SimGlucose data. Additionally the data only includes two exogenous variables, which could lead to there not being enough variance in the data, as well as the influence of the exogenous data on the target not being measurable or clear enough.

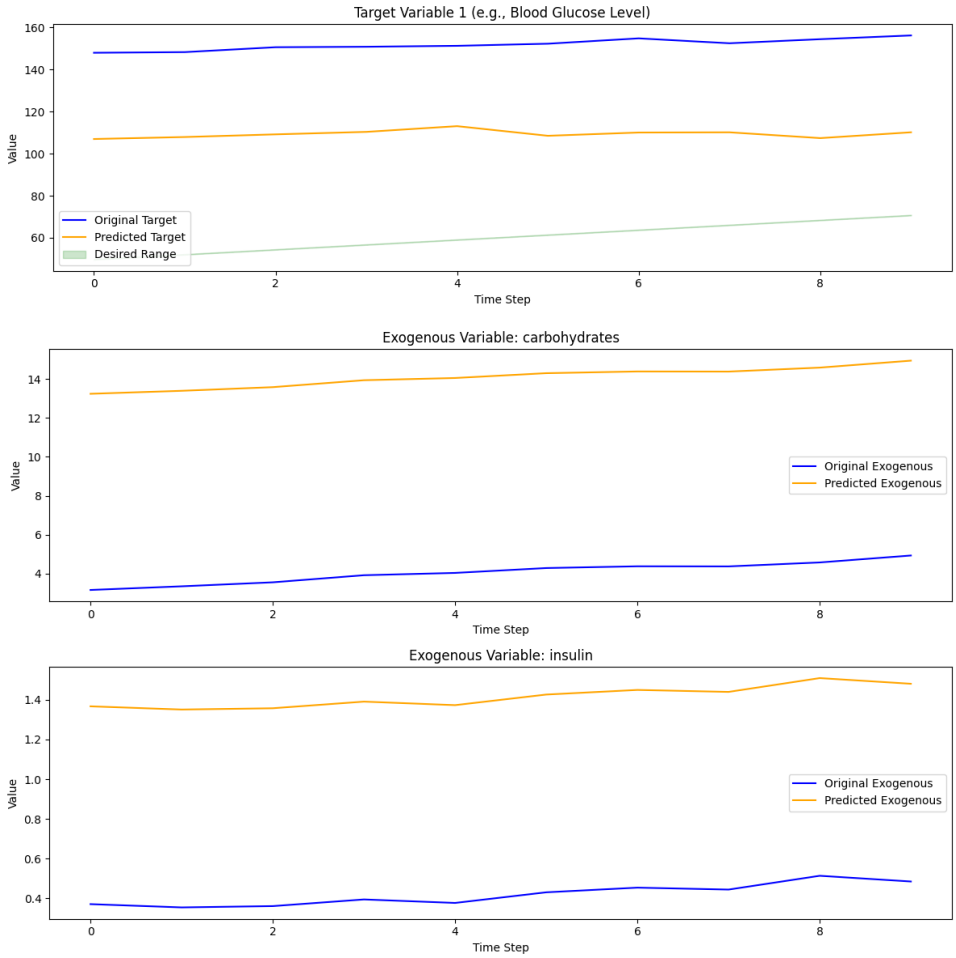


Figure 11: Example of the counterfactuals generated for the SimGlucose dataset, with the blood glucose level closer to the desired bounds, as well as the different exogenous variables, with the original values in blue and the predicted in yellow.

6.2.3 MIMIC-IV

To evaluate how well our model can generate counterfactuals for clinical time series data, we tested several modelling approaches on the MIMIC-IV dataset, focusing specifically on patients with heart failure with preserved ejection fraction (HFpEF). The goal was to create realistic counterfactuals that shift a patient's predicted outcome from death to survival (binary target = 0), by making minimal and interpretable changes to exogenous clinical variables.

Table 8 presents the results for female HFpEF patients, while Table 9 summarizes outcomes for male patients. All models were able to successfully shift a number of predicted outcomes from death to survival without erroneously flipping any survival predictions to

death. In the female cohort, all four models converted 14 predictions to survival. However, GRU and N-BEATS showed higher target prediction deviations (MAE and RMSE) compared to SARIMAX and OLS, which had nearly negligible deviations. GRU exhibited the most substantial changes in clinical variables, with several features, such as Heart Rate, Haemoglobin, and WBC count, modified by over 250%, showing a lack of realism. N-BEATS also made sizeable but more moderate changes (for example SpO2 with 52.3% and Bicarbonate with 46.4%). In contrast, SARIMAX and OLS made minimal adjustments (mostly below 5%), suggesting a preference for interpretability but potentially oversimplified interventions. GRU's broader feature shifts may reflect a deeper modelling of non-linear physiological relationships, offering a more flexible mechanism for achieving counterfactual success at the cost of interpretability.

Metric	**GRU**	**SARIMAX**	**OLS**	**N-BEATS**
Target Prediction Deviation				
MAE	0.09	0.02	0.02	0.18
RMSE	0.11	0.03	0.03	0.25
Max Deviation	1.24	1.00	1.00	1.62
MAPE (%)	7.06M	33.2K	33.2K	15.4M
Survival Changes				
Changes to Survived	14	14	14	14
Changes to Dead	0	0	0	0
Mean % Change in Exogenous Features				
Heart Rate	366.5%	1.2%	1.2%	30.1%
Systolic BP	97.5%	2.3%	2.3%	36.6%
Diastolic BP	68.3%	2.8%	2.8%	40.5%
SpO2	100.1%	3.5%	3.5%	52.3%
Temperature	52.7%	0.5%	0.5%	9.5%
BMI	81.5%	0.3%	0.3%	8.4%
Bicarbonate	78.0%	2.7%	2.7%	46.4%
Creatinine	77.5%	0.9%	0.9%	23.8%
Hemoglobin	368.4%	1.5%	1.5%	25.5%
Platelet Count	76.3%	0.4%	0.4%	9.3%
WBC Count	289.4%	4.0%	4.0%	61.6%
Sodium	158.1%	0.1%	0.1%	27.4%
NT-proBNP	73.2%	0.0%	0.0%	9.2%
Troponin T	175.0%	3.5%	3.5%	50.9%

Table 8: Comparison of counterfactual generation performance across GRU, SARIMAX, OLS, and N-BEATS on female HFpEF patients. Metrics include prediction deviation, number of successful outcome changes, and average percent change in key clinical features. Lower deviations and smaller, targeted feature changes are desirable.

In the male cohort, GRU and N-BEATS achieved far greater impact in terms of survival changes (396 successful conversions) compared to SARIMAX and OLS, which only managed two each. Again, GRU demonstrated a favourable balance, with relatively low prediction deviation (MAE = 0.66) and modest changes to most clinical variables. Notably, GRU made near-zero changes to SpO2, Temperature, and Sodium, while introducing targeted adjustments to variables like Creatinine (12.6%) and Troponin T (39.5%). N-BEATS, while also effective in outcome changes, applied more aggressive modifications in several features, such as Systolic BP (152.1%) and Troponin T (168.6%), raising concerns about the clinical plausibility of such interventions.

Metric	**GRU**	**SARIMAX**	**OLS**	**N-BEATS**
Target Prediction Deviation				
MAE	0.66	0.15	0.15	0.74
RMSE	0.66	0.15	0.15	0.77
Max Deviation	1.01	0.63	0.63	1.45
MAPE (%)	56025.6	178.0	14.9	5169912.5
Survival Changes				
Changes to Survived	396	2	2	396
Changs to Died	0	0	0	0
Mean % Change in Exogenous Features				
Heart Rate	0.1%	10.2%	10.2%	1.8%
Systolic BP	91.6%	67.1%	67.1%	152.1%
Diastolic BP	84.8%	102.1%	102.1%	176.6%
SpO_2	0.1%	3.1%	3.1%	0.5%
Temperature	0.2%	4.1%	4.1%	0.7%
BMI	1.1%	18.8%	18.8%	2.9%
Bicarbonate	0.2%	9.6%	9.6%	1.6%
Creatinine	12.6%	44.2%	44.2%	19.4%
Hemoglobin	0.2%	8.2%	8.2%	1.4%
Platelet Count	8.1%	33.9%	33.9%	13.5%
WBC Count	0.9%	32.4%	32.4%	6.6%
Sodium	0.1%	1.7%	1.7%	0.3%
NT-proBNP	4.6%	23.4%	23.4%	10.0%
Troponin T	39.5%	57.5%	57.5%	168.6%

Table 9: Comparison of counterfactual generation performance across GRU, SARIMAX, OLS, and N-BEATS on male HFpEF patients. Metrics include prediction deviation, number of successful outcome changes, and average percent change in key clinical features. Lower deviations and smaller, targeted feature changes are desirable.

Overall, GRU consistently achieved the best trade-off between effectiveness, as seen in the number of outcomes changed, realism, such as the magnitude of feature changes, and

predictive fidelity, especially in the male patient cohort. SARIMAX and OLS made interpretable but minimal changes that often failed to alter predicted outcomes, highlighting their limited flexibility for counterfactual tasks. While N-BEATS was also effective, its larger feature shifts may reduce its clinical applicability. Additional results by comorbidity cluster are provided in Appendix C.2.

Figures 12 through 15 show examples of counterfactuals generated using four different models for male patients: SARIMAX, Ordinary Least Squares (OLS), GRU, and N-BEATS. In each case, the original values are plotted in blue, and the counterfactual values, those modified to achieve a survival outcome, are shown in yellow, across six time steps. The results for male patients can also be found in Appendix C.2.

Each model approaches the counterfactual generation task differently, and while all share the goal of flipping the predicted outcome, they vary significantly in the magnitude and plausibility of their feature changes.

SARIMAX (Figure 12) often introduces large, unrealistic shifts in variables like Diastolic BP, Troponin T, and Bicarbonate. While the model adheres to temporal smoothness due to its structure, its interventions are extreme and clinically implausible, frequently overshooting physiological limits. This behaviour suggests that SARIMAX lacks an effective internal mechanism to constrain feature values within realistic bounds.

Figure 12: Example of the counterfactuals generated for the MIMIC dataset using SARIMAX, with survival. The different exogenous variables, with the original values in blue and the predicted in yellow.

OLS (Figure 13), by contrast, fails to make meaningful changes to the input features. In many cases, the counterfactuals generated are identical to the originals, suggesting that the linear nature of OLS is too rigid or underpowered for producing actionable edits in time-dependent clinical data. Despite its simplicity, OLS does not effectively adapt the model's prediction, making it the least useful among the methods evaluated.

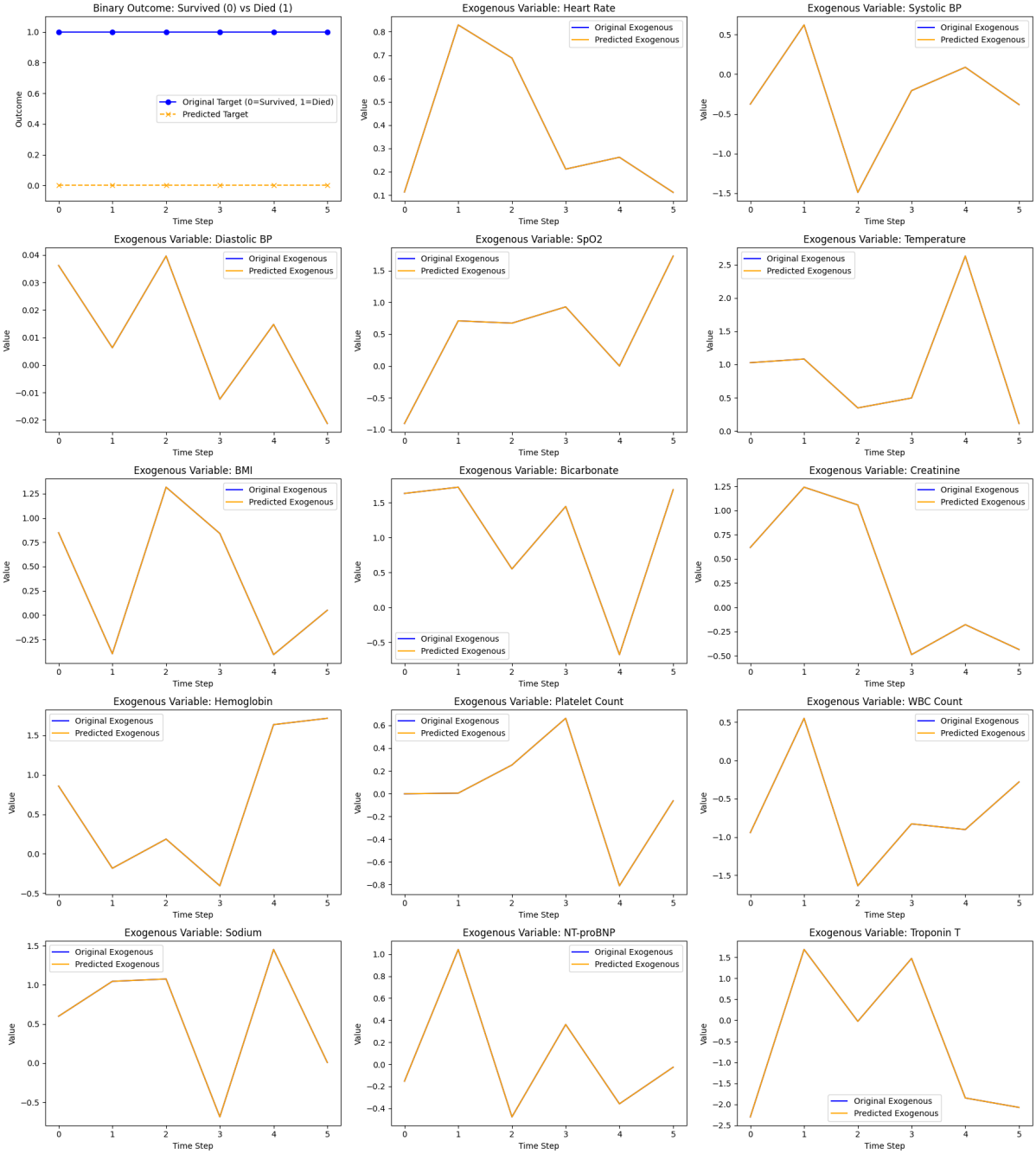


Figure 13: Example of the counterfactuals generated for the MIMIC dataset using OLS, with survival. The different exogenous variables, with the original values in blue and the predicted in yellow.

GRU (Figure 14) produces subtle and targeted changes. Thanks to its ability to model temporal dependencies, it modifies variables like Hemoglobin, Temperature, and Heart Rate just enough to alter the predicted outcome, while maintaining plausible trajectories and staying well within feature bounds. GRU strikes a balance between flexibility and constraint, making it highly effective for clinical counterfactual generation.

Figure 14: Example of the counterfactuals generated for the MIMIC dataset using GRU, with survival. The different exogenous variables, with the original values in blue and the predicted in yellow.

N-BEATS (Figure 15) also achieves strong results, introducing moderate and smooth changes to clinically important variables such as NT-proBNP and Systolic Blood Pressure. While its counterfactuals are generally realistic, it sometimes modifies features more than necessary, and on occasionally generates less plausible clinical values. Nevertheless, its hierarchical structure lends itself well to learning both short- and long-range patterns in the data.

Figure 15: Example of the counterfactuals generated for the MIMIC dataset using N-BEATS, with survival. The different exogenous variables, with the original values in blue and the predicted in yellow.

In summary, while all four models aim to generate plausible counterfactuals, they vary in how effectively they do so. SARIMAX alters key variables too aggressively, often breaching plausibility, while OLS fails to generate impactful counterfactuals at all. GRU introduces small, controlled changes that maintain realism and interpretability, and N-BEATS produces meaningful edits with good temporal coherence but with occasional overreach.

6.3 Evaluation Metrics

6.3.1 Model Comparison

Table 10 presents the average prediction error metrics across all samples and time steps for each evaluated model on the example of the OhioT1DM dataset. The GRU and N-BEATS models significantly outperform SARIMAX and OLS in all prediction accuracy metrics. N-BEATS achieves the lowest MAPE (40.65%), indicating better proportional accuracy, while GRU attains the lowest MAE and RMSE.

Model	MAE	RMSE	Max Dev	MAPE (%)
GRU	**50.67**	**50.94**	**203.47**	53.35
N-BEATS	52.41	56.39	234.97	**40.65**
SARIMAX	111.25	111.48	812.03	116.53
OLS	116.61	116.85	797.12	123.57

Table 10: Prediction Deviation Metrics Across Models

Table 11 summarises each model's ability to generate predictions within the desired blood glucose bounds. GRU again performs best with 43.5% of predictions in range and minimal violation area. Despite the low MAPE of the N-BEATS method, its bound compliance is particularly poor (0.8%), with a large violation area, indicating that low absolute error does not necessarily translate into physiological safety or effectiveness.

Model	In Bound (%)	Mean Violation	Violation Area
GRU	**43.5**	**5.59**	**2944.51**
N-BEATS	0.8	80.60	48170.33
SARIMAX	2.7	87.42	52421.28
OLS	5.7	86.26	51689.86

Table 11: Bound Compliance and Violation Metrics

Table 12 compares the magnitude of changes to exogenous inputs required by each model. GRU introduces the smallest average interventions across all features, with changes remaining under 7%. In contrast, SARIMAX and OLS generate drastic changes, exceeding 200% on average, suggesting poor optimisation stability or unrealistic control strategies. N-BEATS strikes a balance, introducing moderate exogenous adjustments (around 15%).

Model	Basal Insulin	Bolus Insulin	Carbs	Exercise
GRU	**6.3%**	**6.2%**	**5.8%**	**6.3%**
N-BEATS	16.7%	14.3%	15.0%	15.4%
SARIMAX	211.7%	210.9%	211.8%	210.7%
OLS	200.5%	199.6%	200.5%	199.4%

Table 12: Mean Percentage Change in Exogenous Inputs

6.3.2 Feature-wise Euclidean Distances

Table 13 summarises the average Euclidean distances between original and counterfactual feature values, presented both in raw and normalised forms to account for feature scale differences.

Feature	Avg Euclidean Distance	Normalised Euclidean Distance
Basal Insulin	1.2356	4.2353
Bolus Insulin	7.1287	4.4934
Carbohydrates	59.2930	4.2498
Exercise Intensity	0.5816	4.1659

Table 13: Average Euclidean Distances per Feature

As observed, carbohydrates exhibit the largest raw magnitude of change, while normalised distances indicate comparable scale-adjusted changes across all features.

6.3.3 Average Change Per Feature

Table 14 shows the average magnitude of modification applied to each feature in the counterfactual generation process.

Feature	Avg Change
Basal Insulin	0.4400
Bolus Insulin	2.5774
Carbohydrates	21.0900
Exercise Intensity	0.2076

Table 14: Average Change per Feature

The carbohydrates feature undergoes the largest average adjustment, suggesting it plays a key role in driving counterfactual changes.

6.3.4 Healthy Patient Comparison

The following Table 15 presents detailed Euclidean distances between selected original patient samples and their closest healthy patient counterfactuals, for both raw and normalised values, per feature.

Table 15: Euclidean Distances for Healthy Patient Comparisons

Pair	Target Euclidean	Basal Insulin		Bolus Insulin		Carbohydrates		Exercise Intensity	
		Act.	Norm.	Act.	Norm.	Act.	Norm.	Act.	Norm.
1	1.7209	0.8084	2.7711	7.6929	4.8490	48.4378	3.4718	0.3876	2.7764
2	2.1164	0.7179	2.4608	7.7761	4.9014	49.4722	3.5459	0.3764	2.6960
3	2.2299	1.5801	5.4161	11.6026	7.3134	75.4905	5.4108	0.9714	6.9587
4	2.3459	1.9827	6.7964	7.9109	4.9864	89.2984	6.4004	0.4430	3.1732
5	2.4201	1.4544	4.9852	5.0520	3.1844	63.4633	4.5487	0.3887	2.7841

Figure 16 shows the corresponding plots for pair 5, to visualise how the counterfactuals differ from a comparative patient sample.

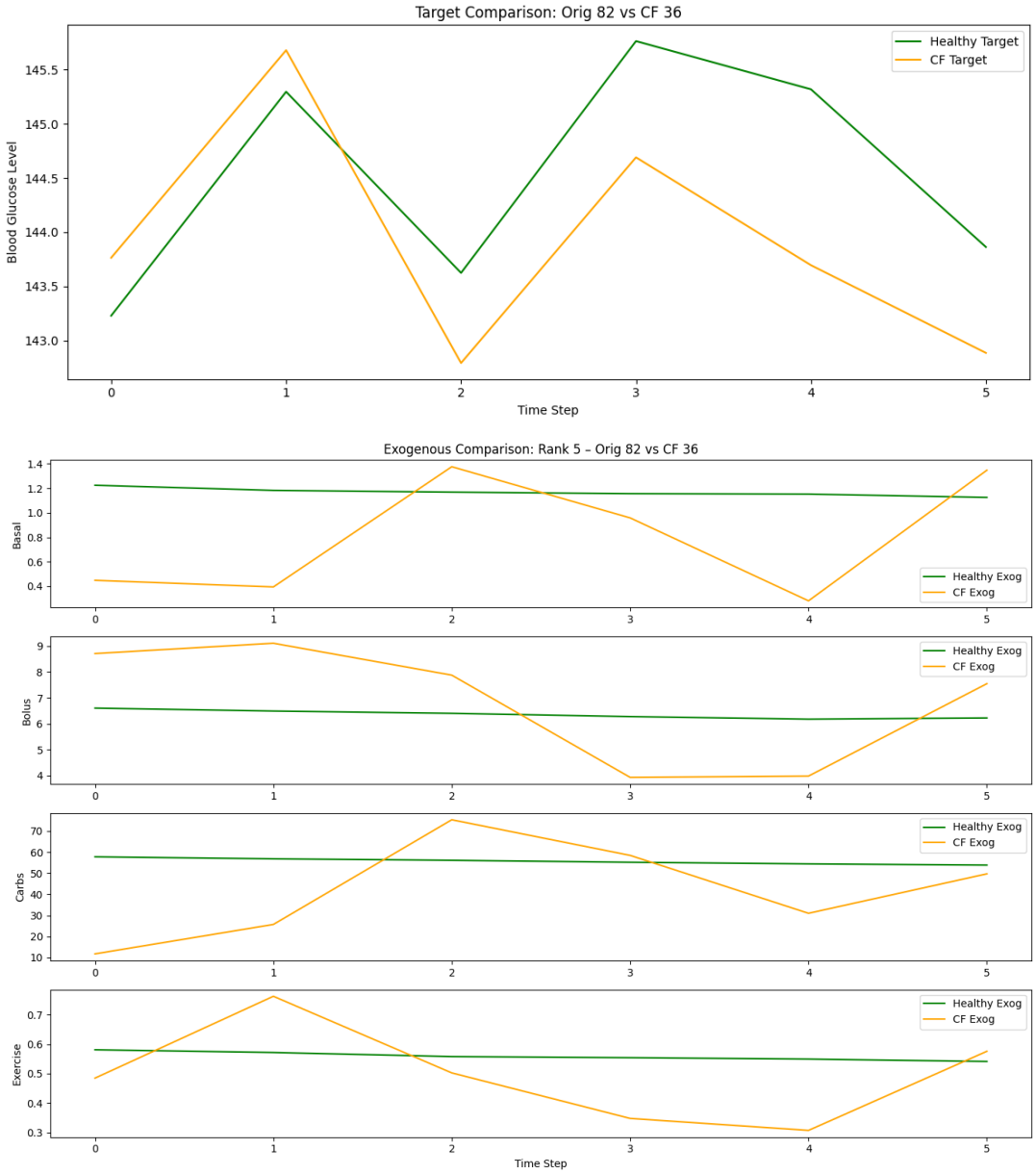


Figure 16: Example of the comparison between generated counterfactuals and similar original data for the OhioT1DM dataset. The target and the different exogenous variables, with the original values in green and the predicted in yellow.

Across all pairs, bolus insulin and carbohydrates exhibit the largest raw Euclidean distances, which is consistent with the trends observed in the previous results. Although exercise intensity shows smaller changes in raw magnitude, its normalised distances reveal that these changes are still significant relative to the feature's scale. The target Euclidean distances between the pairs range from 1.72 to 2.42, indicating a relatively close alignment between the original patient samples and their predicted healthy counterparts. Addition-

ally, there is noticeable variability in the normalised distances across features, reflecting different adjustment requirements for each patient-feature combination.

6.3.5 Overall Counterfactual Evaluation Metrics

Table 16 compiles the key global metrics assessing the counterfactual set's properties, feasibility, and similarity to healthy samples.

Table 16: Global Counterfactual Evaluation Metrics

Metric	Value
Fraction of Values Changed	0.9979
Average Z-Score Change	37.9923
Average Local Outlier Factor (LOF)	-1.0000
Percent Within Bounds	41.6667%
Violation Root Mean Square Error	3.5656
Mean Absolute Difference to Healthy	6.8727
Cosine Similarity to Healthy	0.9085

6.3.5.1 Interpretation

Nearly all feature values are changed (99.8%), indicating extensive counterfactual modifications. The large average z-score change, approximately 38, signifies substantial deviations relative to the underlying feature distributions. The LOF score of -1.0 suggests that the counterfactuals lie within the natural data manifold, meaning they are considered inliers. However, only 41.7% of the counterfactuals satisfy the imposed bounds, which points to potential feasibility concerns. The violation RMSE quantifies the extent of these constraint violations and highlights areas where improvements are needed. Despite this, the high cosine similarity of 0.91 indicates that the counterfactuals maintain a close directional alignment with the healthy samples. Finally, the mean absolute difference of 6.87 shows a moderate absolute distance from the healthy reference points.
The evaluation reveals that carbohydrates and bolus insulin features undergo the most significant changes in producing counterfactuals. While the counterfactuals largely reside within the data manifold, only a portion satisfy all constraints, suggesting room for improvement in enforcing feasibility. Overall, the counterfactuals demonstrate promising similarity to healthy samples but warrant careful clinical validation due to the scale of feature modifications.

6.3.6 Blood Glucose Prediction Performance

The blood glucose prediction deviations were evaluated over 100 samples and 6 time steps. Table 17 summarises the MAE, RMSE, and MAPE metrics per time step. The Mean Absolute Error (MAE) remained relatively stable across the time horizon, ranging from 49.94 to 51.49, with the highest error observed at Time 4. Root Mean Squared Error (RMSE) values followed a similar trend, varying between 64.88 and 66.91. The Mean Absolute Percentage Error (MAPE) was consistently high, between 52.22% and 54.37%,

indicating substantial relative errors despite moderate absolute deviations. The maximum observed deviation across all predictions was 203.47, suggesting occasional large outliers.

Table 17: Blood Glucose Prediction Deviation Metrics (per time step)

Time Step	MAE	RMSE	MAPE (%)
0	50.29	65.56	52.81
1	50.70	65.99	53.46
2	50.99	66.40	53.66
3	50.61	66.12	53.59
4	51.49	66.91	54.37
5	49.94	64.88	52.22
Overall Max Dev		203.47	

6.3.7 Compliance with Desired Blood Glucose Bounds

Table 18 details the percentage of predictions within desired blood glucose bounds, mean violation magnitude, and total violation area per time step. The percentage of predictions within the clinically desired bounds varied considerably over time. Early and late time points (timesteps 0, 3, 4, and 5) exhibited lower compliance, with in-bound percentages around 34 to 36%, whereas the middle time steps (timesteps 1 and 2) showed improved compliance, reaching up to 69% at timestep 2. Mean violation magnitudes were highest at timesteps 1 and 2, reflecting that when predictions fell outside bounds during these steps, the deviations were more severe. Total violation areas, representing cumulative out-of-bound magnitudes, were largest at timestep 0 and gradually decreased through timestep 4, before a slight increase at timestep 5. These results indicate temporal variability in the model's ability to adhere to target glucose constraints, with better performance mid-horizon but challenges at the boundaries.

Table 18: Blood Glucose vs Desired Bounds (per time step)

Time Step	In Bound (%)	Mean Violation	Total Violation Area
0	35.0	11.93	775.71
1	52.0	12.10	580.70
2	69.0	13.50	418.36
3	34.0	5.90	389.17
4	36.0	5.43	347.79
5	35.0	6.66	432.79

6.3.8 Exogenous Variable Adjustments Over Time

The exogenous variables, basal insulin, bolus insulin, carbohydrates, and exercise intensity, show differing magnitudes of adjustment throughout the time series, are summarised in Tables 19 to 22. For the basal insulin, the adjustments were minimal and stable across time, with MAE values between 0.06 and 0.08 and mean percentage changes consistently

around 5 to 7%. Maximum deviations remained low (approximately 0.20 to 0.25 units), suggesting small but steady basal insulin modulations.

Table 19: Basal Insulin Changes (per time step)

Time Step	MAE	Max Dev	Mean % Change
0	0.06	0.25	5.04
1	0.07	0.24	6.25
2	0.07	0.20	6.69
3	0.07	0.22	6.43
4	0.08	0.24	6.73
5	0.07	0.21	6.47

In comparison, the bolus insulin values exhibited larger changes than basal insulin, with MAE increasing from 0.30 to 0.42 over time and percentage changes around 5 to 7%. Max deviations up to 1.09 units indicate moderate but significant bolus insulin adjustments.

Table 20: Bolus Insulin Changes (per time step)

Time Step	MAE	Max Dev	Mean % Change
0	0.30	0.86	5.24
1	0.34	1.09	5.79
2	0.37	1.07	6.00
3	0.37	1.06	6.32
4	0.42	0.92	6.81
5	0.42	1.05	6.97

The largest exogenous changes occurred in carbohydrate intake, with MAE values rising from 2.33 to 3.75 and maximum deviations exceeding 11 units. The mean percentage change also increased over time, peaking at 7.15%, reflecting substantial dietary interventions, particularly in later time steps.

Table 21: Carbohydrates Changes (per time step)

Time Step	MAE	Max Dev	Mean % Change
0	2.33	11.11	4.22
1	2.96	10.10	5.55
2	2.98	10.74	5.57
3	3.37	11.34	6.44
4	3.75	9.89	7.15
5	3.20	8.98	6.08

For exercise intensity, the changes were minimal and consistent, with MAE around 0.03 to 0.04 and max deviations below 0.12. Mean percentage changes hovered near 6%, indicating stable but small modifications in exercise intensity.

Table 22: Exercise Intensity Changes (per time step)

Time Step	MAE	Max Dev	Mean % Change
0	0.03	0.10	6.09
1	0.03	0.10	6.53
2	0.03	0.12	6.25
3	0.03	0.09	5.85
4	0.04	0.09	6.91
5	0.03	0.09	6.39

The prediction errors as well as the percentages of the targets that are in the bounds over time, can be seen in figure 17.

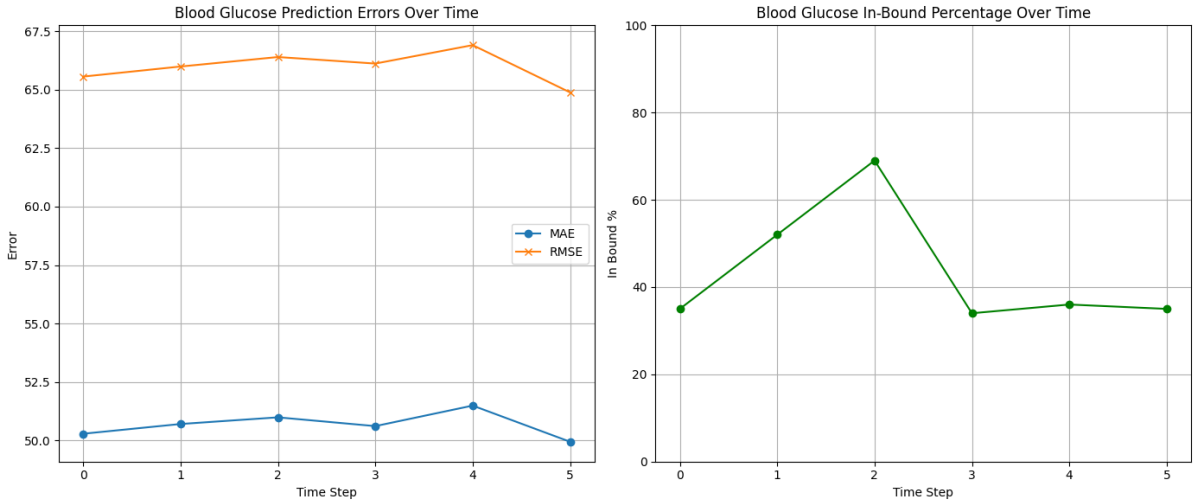


Figure 17: Prediction Errors and In-Bound percentage over time.

7 Discussion

7.1 Multivariate Forecasting

The multivariate forecasting experiments show the clear differences in model behaviour across datasets and configurations. N-BEATS consistently outperforms GRU, particularly in terms of RMSE, indicating better accuracy in capturing the changes in the clinical data. This seems to hold true across both real-world (OhioT1DM) and synthetic (SimGlucose) datasets, with N-BEATS maintaining more stable accuracy as the forecast horizon increases.
While GRU performs reasonably well in detecting general data trends, it struggles with precisely matching the test data, especially with longer horizons. The visualisations further support this, as the N-BEATS forecasts align more closely with observed values, particularly around small changes, whereas GRU tends to smooth over such transitions. These findings support the use of N-BEATS as the underlying forecasting component in counterfactual generation, ensuring that the proposed interventions are based on realistic and consistent predictions.

While these findings strongly support N-BEATS for multivariate forecasting, the experiments on the MIMIC-IV classification task introduce a significant limitation. Despite achieving high overall accuracy in predicting 30-day and 1-year mortality, both N-BEATS and GRU failed to identify any positive cases of the minority non-survival class, resulting in zero precision, recall, and F1-score for that outcome. This problem highlights the challenge of applying these sequence models, that are designed primarily for regression, to severely imbalanced classification problems.
This observation emphasises that model suitability is highly context-dependent: while N-BEATS is well-aligned with continuous time-series forecasting, specialised techniques, such as class rebalancing, may be necessary to adapt N-BEATS for classification tasks, particularly in these healthcare scenarios where identifying rare events is critical. The inconsistency also underlines the need to carefully select metrics and training strategies that support task-specific goals.
In conclusion, the experimental evidence supports the usage of N-BEATS as the preferred forecasting model within the counterfactual generation method. Its robustness across datasets and input configurations ensures that recommendations are based on realistic, accurate projections. At the same time, the MIMIC-IV results show that forecasting strengths do not necessarily translate to classification accuracy in all contexts, underscoring the need for context-specific model design.

7.1.1 Limitations and Future Work

While the GRU and N-BEATS models demonstrate strong performance for the diabetes datasets, a critical limitation was observed in their inability to detect the minority non-survival class for the MIMIC-IV dataset. Both models focused on the majority survival class, which skewed performance metrics such as accuracy and mean squared error. This imbalance is especially problematic in clinical prediction settings where identifying high-risk patients is most important. The observed class imbalance basically biases the learning process, causing the models to overlook lesser occurring but clinically significant outcomes.

This limitation suggests that relying only on conventional evaluation metrics may provide an incomplete picture of model utility in imbalanced clinical datasets. The results highlight the need for incorporating alternative strategies, such as resampling techniques, class weighting, or cost-sensitive learning, to ensure more unbiased predictive performance across all outcome classes. Without such interventions, these models may fail to provide actionable insights in real-world healthcare applications, where missing minority class instances can have serious implications.

Addressing the limitations outlined above should be a central focus of future research. Several promising strategies could be explored to reduce the effects of class imbalance and improve minority class detection. One approach involves implementing class weighting during model training, adjusting the loss function so that misclassifying instances of the minority non-survival class receive a higher penalty. This adjustment would lead to the models paying greater attention to these critical cases.
In conclusion, while the GRU and N-BEATS models show promising results on continuous predictive tasks, realising their full clinical utility requires the implementation of targeted strategies to enhance minority class performance. Addressing class imbalance through these techniques will be essential for developing predictive models that are not only statistically robust but also practically valuable in healthcare decision-making contexts.

7.2 Counterfactuals

The results show that generating multivariate counterfactuals for time series forecasting is not only possible but also practically useful. By adjusting exogenous variables while preserving the general shape of the data, the counterfactual sequences successfully reach desired target outcomes, such as keeping key variables within certain clinical ranges, without introducing unrealistic values. For example, in the diabetes datasets, the method often suggests reducing the predicted glucose level by adjusting exogenous variables such as insulin or carbohydrate intake. In practice, this translates into actionable recommendations like a slightly higher insulin dose at mealtime or moderating carbohydrate consumption. Similarly, in the HFpEF use case, the counterfactuals may involve improving blood pressure, oxygen saturation, or electrolyte levels.
Both visual and quantitative evaluations suggest that the counterfactuals stay within the bounds of what we would expect from real patient data. When compared with naturally occurring trajectories that lead to similar outcomes, the counterfactuals often follow different trajectories using different combinations of features. This reflects a main characteristic of multivariate time series, where multiple input patterns can lead to similar predictions due to adaptability or redundancy in the system.
We also observed that the difference between counterfactuals and nearby real samples helps highlight how features can work together to get a particular outcome. Instead of relying on a single-variable change, the models tend to adjust several features at once, much like how real clinical decisions often involve multiple interventions happening at the same time.
Taken together, these findings support the idea that multivariate counterfactual generation can be a valuable tool for time series forecasting tasks where interpretability and actionable insight are important. The following sections discuss the findings for the dif-

ferent datasets in more detail, highlighting the different strengths and weaknesses of the various methods and datasets.

7.2.1 OhioT1DM

The evaluation of counterfactual generation methods on the OhioT1DM dataset highlights a key problem in clinical time series modelling, which is the trade-off between predictive accuracy and real-world usability. While models like N-BEATS achieve strong forecast accuracy, their lack of constraint adherence and high intervention costs limit their practical applicability in managing blood glucose levels. This reinforces a critical point, that numerical accuracy alone is not sufficiant for clinical decision-making if the suggested interventions are not clinically plausible.
GRU emerged as the most balanced model, offering strong performance across accuracy, constraint satisfaction, and low intervention cost. Its ability to suggest small, targeted changes, particularly in insulin and carbohydrate inputs, makes it well-suited for actionable recommendations in diabetes management.
The OLS model, though interpretable, required extreme changes of the exogenous variables to get the desired predictions, making them unsuitable for intervention planning. Similarly, SARIMAX, despite having moderate accuracy, also lacked plausibility due to high interventions. These findings underscore that interpretability alone does not guarantee practical utility.
The analysis also revealed that key features, especially bolus insulin and carbohydrates, consistently drive changes toward healthy glucose predictions. While effective, these features changes lead to significant interventions, which requires clinical oversight to implement safely. Normalised distance metrics showed that even small raw changes in features like exercise can represent meaningful deviations, highlighting the need for scale-aware evaluation when assessing these counterfactuals.
Comparisons with healthy patient data confirmed the biological plausibility of the generated counterfactuals. Metrics such as cosine similarity and low outlier scores indicate that the recommendations stay within a realistic data range and align closely with healthy data trajectories, even when requiring substantial changes across multiple inputs.
In summary, GRU-based counterfactuals demonstrate a promising balance between predictive ability and clinical plausibility. However, this evaluation also reveals opportunities to improve constraint satisfaction and reduce the necessary amount of intervention. Future work should explore adaptive strategies, such as personalised constraints or cost-aware optimization, to further align counterfactuals with clinical feasibility and individual patient needs.

7.2.2 SimGlucose

In the SimGlucose setting, the counterfactual sequences take a slightly different approach. The model tends to propose conservative strategies, minimising both carbohydrate intake and insulin doses to improve blood glucose levels. In many cases, insulin is reduced to zero, and carb intake is either flat or only slightly increased in later time steps. These subtle adjustments lead to the blood glucose predictions often not falling within the desired range, but staying closer to the original glucose range. This behaviour shows that, while there is quite some potential, the SimGlucose dataset needs a more specific approach to be able to generate realistic and meaningful counterfactuals.

7.2.3 MIMIC

This study explored the use of counterfactual generation in clinical time series data, focusing on patients with heart failure with preserved ejection fraction (HFpEF) in the MIMIC-IV dataset. The goal was to generate alternate trajectories for exogenous clinical variables that could change a predicted outcome from death to survival, in a way that remains realistic and interpretable.

The results show that all four modelling approaches (SARIMAX, OLS, GRU, and N-BEATS) successfully changed the predicted outcomes to survival without introducing new death predictions. However, the extent of the feature adjustments varied significantly across models. GRU achieved the best trade-off between predictive accuracy and counterfactual plausibility, with relatively low prediction deviation, with MAE = 0.09 for females, 0.66 for males, and meaningful, targeted feature changes. N-BEATS also showed strong adaptability but introduced larger adjustments, such as a Heart Rate change of 30.1% for females, which could make clinical implementation and interpretability difficult.

In contrast, SARIMAX and OLS applied minimal modifications (for example a Heart Rate change of 1.2% for females), preserving interpretability but potentially oversimplifying the underlying dynamics of HFpEF progression. This careful behaviour is reflected in their low target prediction deviation (MAE = 0.02 for females and 0.15 for males), but the subtle changes may not always represent realistic interventions.

A closer look at the exogenous features changes shows that the models selectively modified clinically relevant variables. For example, Heart Rate, Systolic BP, and Troponin T exhibited the largest changes under GRU and N-BEATS, particularly in female patients, such as a Heart Rate change of 366.5%. Conversely, features like BMI and NT-proBNP remained relatively stable across all models, suggesting that these variables were deemed less critical for changing survival outcomes. This selective behaviour strengthens the interpretability of the counterfactuals, but shows again the lack of realism of the interventions.

Interestingly, male patients required fewer drastic adjustments compared to female patients for the GRU model. This could be explained by the fact, that the initial percentage of females surviving HFpEF was much larger, but it also suggests possible sex-specific differences in how clinical trajectories influence predicted outcomes, a finding which would be interesting to further explore in a clinical research setting.

SARIMAX, as a classical statistical model, produced gradual and trend-consistent counterfactuals, but also introduced some very abrupt changes. This could be explained by the restricted adaptability due to its linear architecture. OLS, really struggled with the counterfactual generation, often predicting a changed outcome, without adapting any of the features. This makes OLS a less trustworthy method for this problem in clinical decision-making.

On the other hand, the deep learning models, GRU and N-BEATS, were better at modelling the non-linear dependencies and complex temporal patterns of the data. GRU generated counterfactuals that were both targeted and consistent, making it the most clinically promising among the models tested. N-BEATS often required larger shifts, that might be difficult to align with realistic clinical interventions.

The findings underscore the potential of counterfactual modelling to provide "what-if" scenarios for critical care. By demonstrating how slight adjustments to key clinical variables might alter predicted outcomes, these models can support treatment planning and decision support. However, the results also reveal model-specific trade-offs between flexibility

and interpretability.
Overall, GRU emerged as the most effective approach, balancing accuracy, interpretability, and realistic feature adjustments, while SARIMAX and OLS provided valuable baselines due to their transparency. N-BEATS demonstrated strong pattern modelling capabilities but requires careful evaluation to avoid over-adjusting the variables. These findings set a foundation for building clinically actionable counterfactual models that enhance transparency and trust in time series forecasting for healthcare.

7.3 Limitations and Future Work

While our evaluation framework provides valuable insights into the behaviour and plausibility of the generated counterfactual interventions, there are several important limitations to note. Firstly, although the counterfactuals often align with clinical norms and appear visually and quantitatively plausible, our evaluation remains limited by the absence of clinical domain experts on the research team. As demonstrated in the OhioT1DM and MIMIC analyses, the models sometimes suggest changes, such as insulin adjustments or large heart rate shifts, that may seem numerically reasonable but seem to be clinically unrealistic or unsafe. Collaborating with a diverse group of healthcare professionals, is essential to validate the realism, safety, and applicability of these suggestions in actual medical practice.
Another limitation is, that our counterfactual generation is guided primarily by mathematical objectives, such as euclidean distances, z-scores, and LOF scores, rather than by patient-centric considerations such as behavioural adherence, lifestyle constraints, or individual variability in insulin sensitivity. Consequently, the suggested interventions may overlook factors like meal timing, stress, or simultaneous medications, which can strongly influence the blood glucose levels. For example, in SimGlucose, the model often reduces insulin to zero, which may not reflect practical or safe treatment trajectories. Similarly, in OhioT1DM, while small input shifts (for example in carbohydrates or bolus insulin) led to improved glycaemic trajectories, some interventions represented substantial perturbations that might not align with real-world adherence patterns or safety thresholds. Future work should embed clinical rules, lifestyle patterns, and personalised behaviour models to guide counterfactual generation more realistically.
Our current approach computes counterfactual perturbations using feature-wise distances normalised by global variance, assuming independence between variables. However, in clinical contexts, features often interact in non-linear ways, as seen in the differential impacts of insulin and carbohydrate intake in OhioT1DM, or sex-specific variation in MIMIC. Future work should move toward joint and conditional distance measures, multi-objective-based constraints, and domain-informed regularisation strategies that better capture multivariate dependencies. Additionally, our findings show that counterfactual quality and realism vary across datasets and model types. In SimGlucose, for example, overly conservative counterfactuals often fail to move the predicted glucose level into the target range, indicating that further adjusting the approach to simulator-based or synthetic environments would be necessary. This also suggests the need for adaptive strategies that adjust the amount of interventions and more detailed implementation of constraints depending on dataset characteristics, clinical context, and model behaviour.
To address these limitations, several solutions could be investigated. Firstly, as already mentioned above, including clinicians and healthcare professionals will be essential to re-

fine constraint design, define safe intervention ranges, and assess model outputs in pilot or retrospective clinical studies. This will improve the clinical relevance and safety of the counterfactuals, especially in sensitive healthcare domains. Another future implementation could incorporate more detailed patient metadata, such as age, comorbidities, insulin sensitivity profiles, and lifestyle habits, which could allow for truly individualised counterfactuals. Integrating behavioural models to estimate adherence likelihood, and employing multi-objective optimisation to jointly optimise for outcome, plausibility, and patient load, could significantly enhance model usability as well. Finally, as shown by the irregular behaviour of GRU, N-BEATS, OLS, and SARIMAX across datasets, different models offer different strengths. Future work could explore ensemble or hybrid approaches that combine the interpretability of linear models with the expressiveness of neural networks, while dynamically adjusting the intervention intensity. By addressing these limitations, future iterations of this work can move towards developing trustworthy, interpretable, and clinically actionable counterfactual explanations for time series forecasting in healthcare.

8 Conclusion

This thesis presents a new method for counterfactual time series forecasting that focuses on modifying exogenous variables within the forecast horizon to achieve desired outcomes, rather than altering historical data. This approach attempts to fill a critical gap in current research, offering a more actionable and interpretable alternative for clinical decision support. By learning the relationship between forecasted targets and exogenous inputs using models such as SARIMAX, OLS, GRU, and N-BEATS, we enable the generation of realistic, constrained counterfactuals that respect the temporal dynamics of healthcare data.

Comprehensive experiments across two distinct healthcare contexts, blood glucose forecasting using OhioT1DM and SimGlucose datasets, and mortality prediction in HFpEF patients with MIMIC-IV, demonstrate both the strengths and limitations of the proposed approach. In forecasting tasks, N-BEATS consistently outperformed GRU in terms of accuracy and temporal stability, particularly over longer horizons, establishing it as the most effective model for high-accuracy prediction. However, GRU proved to be more adept at generating clinically plausible and constrained counterfactuals, especially when intervention feasibility and physiological adherence were critical, as seen in the OhioT1DM case. In contrast, OLS and SARIMAX, while interpretable, often failed to suggest realistic or effective interventions, either requiring impractical input changes or neglecting necessary adjustments. The counterfactuals generated using GRU and N-BEATS often mirrored realistic multi-variable interventions, supporting their potential role in personalised treatment planning.

Importantly, the analysis revealed that generating meaningful interventions often involves coordinated adjustments across multiple variables, reflecting real-world clinical strategies. However, limitations such as poor minority class detection for the MIMIC-IV dataset and the absence of clinical domain input highlight the need for further refinement. Moreover, while the models produce mathematically and physiologically coherent counterfactuals, their real-world clinical utility remains uncertain due to the lack of domain-expert input and patient-specific behavioural considerations. The proposed interventions, though promising, may not fully reflect constraints like treatment adherence, lifestyle feasibility, or individual variability in response.

Future work should involve collaboration with clinicians to evaluate the realism and safety of suggested interventions, and integrate patient metadata and behavioural modelling to enhance personalisation. Additionally, optimisation techniques that account for adherence, safety, and ethical considerations, could further improve the interpretability and practicality of the proposed method.

In summary, this work demonstrates that counterfactual generation for time series forecasting is both feasible and clinically relevant. It provides a foundation for building transparent, adaptive, and patient-centred decision support systems capable of suggesting personalised, data-driven interventions based on realistic future scenarios.

References

[AALC20] Emre Ates, Burak Aksar, Vitus Leung, and Kaan Coşkun. Counterfactual Explanations for Machine Learning on Multivariate Time Series Data. *CoRR*, 2020.

[ASFWHY+25] Nur Farah Afifah Ahmad Sukri, Wan Mohd Amir Fazamin Wan Hamzah, Mohd Kamir Yusof, Ismahafezi Ismail, Harmy Mohamed Yusoff, and Azliza Yacob. A Systematic Literature Review on Machine Learning in Healthcare Prediction. *International Journal of Online & Biomedical Engineering (iJOE)*, 21(6):155–177, 2025.

[BAI25] Mamoune Benaida, Ibtissam Abnane, and Ali Idri. Deep learning based one step and multi-steps ahead forecasting blood glucose level. *Expert Systems: The Journal of Knowledge Engineering*, 42(1), 2025.

[BP10] Barry Borlaug and James Paulus. Heart failure with preserved ejection fraction: Pathophysiology, diagnosis, and treatment. *European Heart Journal*, 32:670–679, 2010.

[CHN+21] Ran Cui, Chirath Hettiarachchi, Christopher Nolan, Elena Daskalaki, and Hanna Suominen. Personalised Short-Term Glucose Prediction via Recurrent Self-Attention Network. In *34th IEEE International Symposium on Computer-Based Medical Systems, CBMS 2021, Aveiro, Portugal, June 7-9, 2021*, pages 154–159. IEEE, 2021.

[CLH24] Ying Chen, Pei-Hung Liao, and Chung-Lieh Hung. Diagnostic Yield and Model Prediction Using Wearable Patch Device in HFpEF. In *Innovation in Applied Nursing Informatics, 16th International Congress in Nursing Informatics, Manchester, UK, July 28-31, 2024*, volume 315 of *Studies in Health Technology and Informatics*, pages 25–30. IOS Press, 2024.

[Dia] Diabetes. Last Accessed: Jul. 5, 2025. Available: https://www.who.int/health-topics/diabetes.

[DMML+14] Chiara Dalla Man, Francesco Micheletto, Dayu Lv, Marc Breton, Boris Kovatchev, and Claudio Cobelli. The UVA/PADOVA type 1 diabetes simulator: New features. *Journal of diabetes science and technology*, 8:26–34, 2014.

[EMA+24a] Nuha Elsayed, Rozalina McCoy, Grazia Aleppo, Kirthikaa Balapattabi, Elizabeth Beverly, Kathaleen Early, Dennis Bruemmer, Osagie Ebekozien, Justin Echouffo-Tcheugui, Laya Ekhlaspour, Jason Gaglia, Rajesh Garg, Kamlesh Khunti, Rayhan Lal, Ildiko Lingvay, Glenn Matfin, Naushira Pandya, Elizabeth Pekas, Scott Pilla, and Raveendhara Bannuru. 2. Diagnosis and Classification of Diabetes: Standards of Care in Diabetes - 2025. *Diabetes Care*, 48:S27–S49, 2024.

[EMA+24b] Nuha Elsayed, Rozalina McCoy, Grazia Aleppo, Kirthikaa Balapattabi, Elizabeth Beverly, Kathaleen Early, Dennis Bruemmer, Justin Echouffo-Tcheugui, Laya Ekhlaspour, Rajesh Garg, Kamlesh Khunti, Rayhan Lal,

Ildiko Lingvay, Glenn Matfin, Naushira Pandya, Elizabeth Pekas, Scott Pilla, Sarit Polsky, Alissa Segal, and Raveendhara Bannuru. 6. Glycemic Goals and Hypoglycemia: Standards of Care in Diabetes - 2025. *Diabetes Care*, 48:S128–S145, 2024.

[FAJ+18] Ian Fox, Lynn Ang, Mamta Jaiswal, Rodica Pop-Busui, and Jenna Wiens. Deep Multi-Output Forecasting: Learning to Accurately Predict Blood Glucose Trajectories. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 1387–1395. ACM, 2018.

[GBV+24] Raffaele Giancotti, Pietro Bosoni, Patrizia Vizza, Giuseppe Tradigo, Agostino Gnasso, Pietro Guzzi, Riccardo Bellazzi, Concetta Irace, and Pierangelo Veltri. Forecasting glucose values for patients with type 1 diabetes using heart rate data. *Computer Methods and Programs in Biomedicine*, 257:108438, 2024.

[GRU] Gated Recurrent Unit. Last Accessed: Jul. 5, 2025. Available: https://www.tensorflow.org/api_docs/python/tf/keras/layers/GRU.

[Gui24] Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 38(5):2770–2824, 2024.

[HMH+25] Yue Hu, Fanghui Ma, Mengjie Hu, Binbing Shi, Defeng Pan, and Jingjing Ren. Development and validation of a machine learning model to predict the risk of readmission within one year in HFpEF patients. *International Journal of Medical Informatics*, 194:105703, 2025.

[HSL+23] Chenlu Hong, Linjuan Sun, Guangwen Liu, Boyuan Guan, Chengfu Li, and Yanan Luo. Response of Global Health Towards the Challenges Presented by Population Aging. *China CDC Weekly*, 5(39):884–887, 2023.

[JBS+23] Alistair Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei Lehman, Leo Celi, and Roger Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10:1, 2023.

[JHS+22] Mohd Javaid, Abid Haleem, Ravi Pratap Singh, Rajiv Suman, and Shanay Rab. Significance of machine learning in healthcare: Features, pillars and applications. *International Journal of Intelligent Networks*, 3:58–73, 2022.

[KBB23] Pinar Kavas, Mehmet Bozkurt, and Cahit Bilgin. Machine learning-based medical decision support system for diagnosing HFpEF and HFrEF using PPG. *Biomedical Signal Processing and Control*, 79:104164, 2023.

[KCK+20] Shruti Kaushik, Abhinav Choudhury, Pankaj Kumar, Nataraj Dasgupta, Sayee Natarajan, Larry Pickett, and Varun Dutt. AI in Healthcare: Time-Series Forecasting Using Statistical, Neural, and Ensemble Architectures. *Frontiers in Big Data*, 3:4, 2020.

[KM25] Deepjyoti Kalita and Khalid Mirza. Multivariate Glucose Forecasting Using Deep Multihead Attention Layers Inside Neural Basis Expansion Networks. *IEEE Journal of Biomedical and Health Informatics*, 29(5):3654–3663, 2025.

[Kol24] Milind Kolambe. Forecasting the Future: A Comprehensive Review of Time Series Prediction Techniques. *Journal of Electrical Systems*, 20:575–586, 2024.

[KS20] Zahra Karevan and Johan Suykens. Transductive LSTM for time-series prediction: An application to weather forecasting. *Neural Networks*, 125:1–9, 2020.

[LACMPC+25] Francisco Lara Abelenda, David Chushig-Muzo, Pablo Peiro-Corbacho, Ana Wägner, Conceição Granja, and Cristina Soguero Ruiz. Personalized glucose forecasting for people with type 1 diabetes using large language models. *Computer Methods and Programs in Biomedicine*, 265:108737, 2025.

[LBF13] Marco Lippi, Matteo Bertini, and Paolo Frasconi. Short-Term Traffic Flow Forecasting: An Experimental Comparison of Time-Series Analysis and Supervised Learning. *IEEE Transactions on Intelligent Transportation Systems*, 14(2):871–882, 2013.

[MB20] Cindy Marling and Razvan Bunescu. The OhioT1DM Dataset for Blood Glucose Level Prediction: Update 2020. In *Proceedings of the 5th International Workshop on Knowledge Discovery in Healthcare Data co-located with 24th European Conference on Artificial Intelligence, KDH@ECAI 2020, Santiago de Compostela, Spain & Virtually, August 29-30, 2020*, volume 2675 of *CEUR workshop proceedings*, pages 71–74, 2020.

[MGL+19] Awais Malik, Gauravpal Gill, Fahad Lodhi, Lakshmi Tummala, Steven Singh, Charity Morgan, Richard Allman, Gregg Fonarow, and Ali Ahmed. Prior Heart Failure Hospitalization and Outcomes in Patients with Heart Failure with Preserved and Reduced Ejection Fraction. *The American Journal of Medicine*, 133:84–94, 2019.

[MKT+24] Kirsty McDowell, Toru Kondo, Atefeh Talebi, Ken Teh, Erasmus Bachus, Rudolf de Boer, Ross Campbell, Brian Claggett, Ashkay Desai, Kieran Docherty, Adrian Hernandez, Silvio Inzucchi, Mikhail Kosiborod, Carolyn Lam, Felipe Martinez, Joanne Simpson, Muthiah Vaduganathan, Pardeep Jhund, Scott Solomon, and John Mcmurray. Prognostic models for mortality and morbidity in heart failure with preserved ejection fraction. *JAMA cardiology*, 9(5):457–465, 2024.

[MPS19] Nishita Mehta, Anil Pandit, and Sharvari Shukla. Transforming healthcare with big data analytics and artificial intelligence: A systematic mapping study. *Journal of Biomedical Informatics*, 100:103311, 2019.

[MSG14] Prapanna Mondal, Labani Shit, and Saptarsi Goswami. Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices. *International Journal of Computer Science, Engineering and Applications*, 4(2):13–29, 2014.

[MWM24] Kasun Mendis, Manjusri Wickramasinghe, and Pasindu Marasinghe. Multivariate Time Series Forecasting: A Review. In *Proceedings of the 2024 2nd Asia Conference on Computer Vision, Image Processing and Pattern Recognition, CVIPPR 2024, Xiamen, China, April 26-28, 2024*, pages 1–9. ACM, 2024.

[N-B] Neural Basis Expansion Analysis for interpretable Time Series forecasting. Last Accessed: Jul. 5, 2025. Available: https://blog.mlq.ai/time-series-with-tensorflow-n-beats-algorithm/.

[OCM+22] Kin G. Olivares, Cristian Challu, Grzegorz Marcjasz, Rafał Weron, and Artur Dubrawski. Neural basis expansion analysis with exogenous variables: Forecasting electricity prices with NBEATSx. *International Journal of Forecasting*, 39, 2022.

[OHH+06] Theophilus Owan, David Hodge, Regina Herges, Steven Jacobsen, Veronique Roger, and Margaret Redfield. Heart Failure with Preserved Ejection Fraction: Trends in Prevalence and Outcomes. *The New England Journal of Medicine*, 355:251–259, 2006.

[OLS] Ordinary Least Squares. Last Accessed: Jul. 5, 2025. Available: https://www.statsmodels.org/stable/generated/statsmodels.regression.linear_model.OLS.html.

[PTJ+22] Muskaan Pirani, Paurav Thakkar, Pranay Jivrani, Mohammed Bohara, and Dweepna Garg. A Comparative Analysis of ARIMA, GRU, LSTM and BiLSTM on Financial Time Series Forecasting. In *2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*, pages 1–6, 2022.

[PVA+16] Piotr Ponikowski, Adriaan Voors, Stefan Anker, Héctor Bueno, John Cleland, Andrew Coats, Volkmar Falk, José González-Juanatey, Veli-Pekka Harjola, Ewa Jankowska, Mariell Jessup, Cecilia Linde, Petros Nihoyannopoulos, John Parissis, Burkert Pieske, Jillian Riley, Giuseppe Rosano, Luis Ruilope, Frank Ruschitzka, and Peter Meer. 2016 esc guidelines for the diagnosis and treatment of acute and chronic heart failure: The task force for the diagnosis and treatment of acute and chronic heart failure of the european society of cardiology (esc). developed with the special contribution of the heart failure association (hfa) of the esc. *European Heart Journal*, 37:2129–2200, 2016.

[RNZ17] Muhammad Imran Razzak, Saeeda Naz, and Ahmad Zaib. Deep Learning for Medical Image Processing: Overview, Challenges and Future. *Classification in BioApps: Automation of decision making*, pages 323–350, 2017.

[RRCVR23] Ignacio Rodríguez-Rodríguez, María Campo-Valera, and José-Víctor Rodríguez. Forecasting Glycaemia for Type 1 Diabetes Mellitus Patients by Means of IoMT Devices. *Internet of Things*, 24:100945, 2023.

[SAR] Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors model. Last Accessed: Jul. 5, 2025. Available: https://www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html.

[SBA+24] Ikgyu Shin, Nilay Bhatt, Alaa Alashi, Keervani Kandala, and Karthik Murugiah. Predicting 30-Day and 1-Year Mortality in Heart Failure with Preserved Ejection Fraction (HFpEF). *medRxiv : the preprint server for health sciences*, 2024.

[SK25a] Yiheng Shen and Samantha Kleinberg. Personalized Blood Glucose Forecasting From Limited CGM Data Using Incrementally Retrained LSTM. *IEEE Transactions on Biomedical Engineering*, 72(4):1266–1277, 2025.

[SK25b] Yuyang Sun and Panagiotis Kosmas. Integrating Bayesian Approaches and Expert Knowledge for Forecasting Continuous Glucose Monitoring Values in Type 2 Diabetes Mellitus. *IEEE Journal of Biomedical and Health Informatics*, 29(2):1419–1432, 2025.

[SM23] Hugo Souto and Amir Moradi. Introducing NBEATSx to Realized Volatility Forecasting. *Expert Systems with Applications*, 242:122802, 2023.

[VBH+24] Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan Hines, John Dickerson, and Chirag Shah. Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review. *ACM Computing Surveys*, 56, 2024.

[WLY24] Wenhao Wang, Xiuqin Li, and Pengjia Yan. A Multi-Scaler Hybrid Autoformer for Enhanced Time Series Forecasting in Energy Consumption. *IEEE Access*, 12:196347–196363, 2024.

[WMSP23] Zhendong Wang, Ioanna Miliou, Isak Samsten, and Panagiotis Papapetrou. Counterfactual Explanations for Time Series Forecasting. In *IEEE International Conference on Data Mining, ICDM 2023, Shanghai, China, December 1-4, 2023*, pages 1391–1396. IEEE, 2023.

[WSMP24] Zhendong Wang, Isak Samsten, Ioanna Miliou, and Panagiotis Papapetrou. COMET: Constrained Counterfactual Explanations for Patient Glucose Multivariate Forecasting. In *37th IEEE International Symposium on Computer-Based Medical Systems, CBMS 2024, Guadalajara, Mexico, June 26-28, 2024*, pages 502–507. IEEE, 2024.

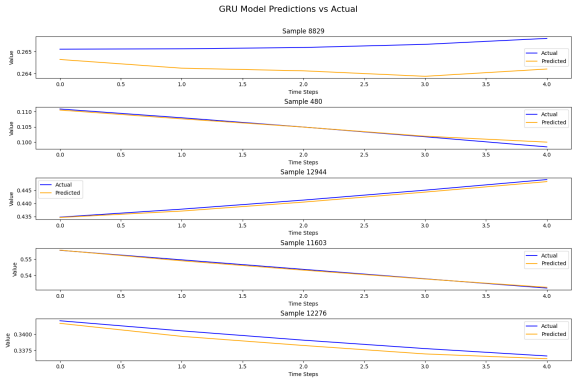
[Xie18] Jinyu Xie. Simglucose v0.2.1, 2018. Accessed: Sep. 16, 2024. Available: https://github.com/jxx123/simglucose.

[ZGW+21] Liye Zhou, Zhifei Guo, Bijue Wang, Yongqing Wu, Zhi Li, Hongmei Yao, Ruiling Fang, Haitao Yang, Hongyan Cao, and Yuehua Cui. Risk Prediction in Patients With Heart Failure With Preserved Ejection Fraction Using Gene Expression Data and Machine Learning. *Frontiers in Genetics*, 12:652315, 2021.

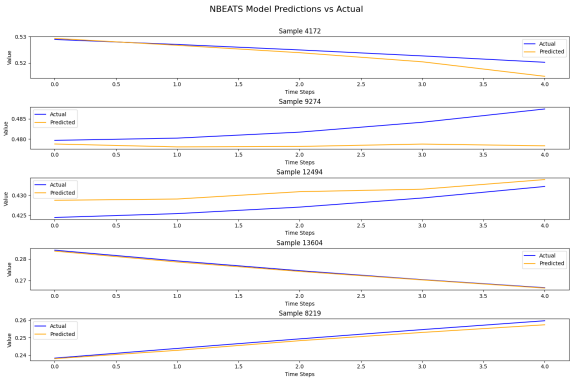
[ZLH+18] Taiyu Zhu, Kezhi Li, Pau Herrero, Jianwei Chen, and Pantelis Georgiou. A Deep Learning Algorithm for Personalized Blood Glucose Prediction. In *Proceedings of the 3rd International Workshop on Knowledge Discovery in Healthcare Data co-located with the 27th International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence (IJCAI-ECAI 2018), Stockholm, Sweden, July 13, 2018*, CEUR Workshop Proceedings, pages 64–78. CEUR-WS.org, 2018.

A Detailed Results of the Multivariate Forecasting

The following figures provide some additional results for the Multivariate Forecasting with different back horizons and horizon for both the SimGlucose and the OhioT1DM dataset.

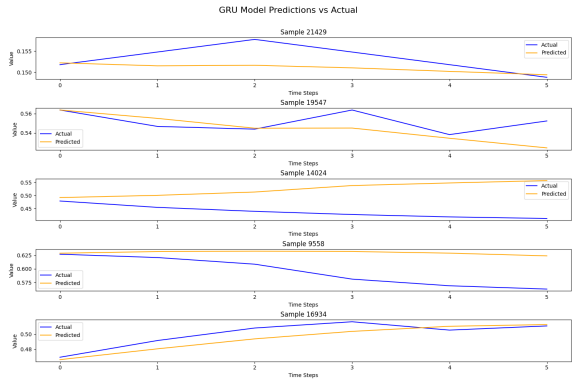


(a) Results of the multivariate forecasting using GRU.

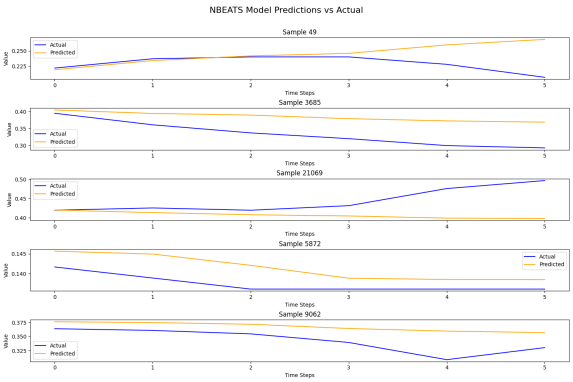


(b) Results of the multivariate forecasting using N-BEATS.

Figure 18: Results of the multivariate forecasting for the SimGlucose dataset with back horizon = 20 and forecast horizon = 5, showing the accuracy of the forecasting.



(a) Results of the multivariate forecasting using GRU.



(b) Results of the multivariate forecasting using N-BEATS.

Figure 19: Results of the multivariate forecasting for the OhioT1DM dataset with back horizon = 12 and forecast horizon = 6, showing the accuracy of the forecasting.

B Detailed Classification Metrics for MIMIC Dataset

The following section includes tables that provide comprehensive classification metrics (precision, recall, F1-score, and overall accuracy) for GRU and N-BEATS models across all patient clusters and gender subgroups in the MIMIC dataset, as well as the corresponding figures visualizing these results.

Model	Target	Class	Precision	Recall	F1-Score
GRU	30 days	Survived	0.91	0.86	0.88
		Died	0.00	0.00	0.00
		Overall Accuracy		0.7882	
	1 year	Survived	0.90	0.91	0.90
		Died	0.00	0.00	0.00
		Overall Accuracy		0.8235	
N-BEATS	30 days	Survived	0.92	0.97	0.94
		Died	0.00	0.00	0.00
		Overall Accuracy		0.8922	
	1 year	Survived	0.90	0.96	0.93
		Died	0.06	0.02	0.03
		Overall Accuracy		0.8706	

Table 23: Classification Metrics for GRU and N-BEATS Models Cluster 0

Model	Target	Class	Precision	Recall	F1-Score
GRU	30 days	Survived	0.80	0.22	0.35
		Died	0.07	0.50	0.12
		Overall Accuracy		0.2486	
	1 year	Survived	1.00	0.05	0.10
		Died	0.10	1.00	0.19
		Overall Accuracy		0.1475	
N-BEATS	30 days	Survived	0.90	0.93	0.91
		Died	0.00	0.00	0.00
		Overall Accuracy		0.8424	
	1 year	Survived	0.91	0.97	0.94
		Died	0.26	0.08	0.13
		Overall Accuracy		0.8871	

Table 24: Classification Metrics for GRU and N-BEATS Models Cluster 1

Model	Target	Class	Precision	Recall	F1-Score
GRU	30 days	Survived	0.90	0.53	0.67
		Died	0.12	0.50	0.19
		Overall Accuracy		0.5291	
	1 year	Survived	0.91	0.61	0.73
		Died	0.14	0.50	0.21
		Overall Accuracy		0.5979	
N-BEATS	30 days	Survived	0.89	0.99	0.94
		Died	0.18	0.02	0.04
		Overall Accuracy		0.8807	
	1 year	Survived	0.90	0.88	0.89
		Died	0.15	0.18	0.16
		Overall Accuracy		0.7982	

Table 25: Classification Metrics for GRU and N-BEATS Models Cluster 2

Model	Target	Class	Precision	Recall	F1-Score
GRU	30 days	Survived	0.95	1.00	0.97
		Died	0.00	0.00	0.00
		Overall Accuracy		0.9487	
	1 year	Survived	0.95	1.00	0.97
		Died	0.00	0.00	0.00
		Overall Accuracy		0.9487	
N-BEATS	30 days	Survived	0.95	1.00	0.97
		Died	0.00	0.00	0.00
		Overall Accuracy		0.9484	
	1 year	Survived	0.95	1.00	0.97
		Died	0.00	0.00	0.00
		Overall Accuracy		0.9487	

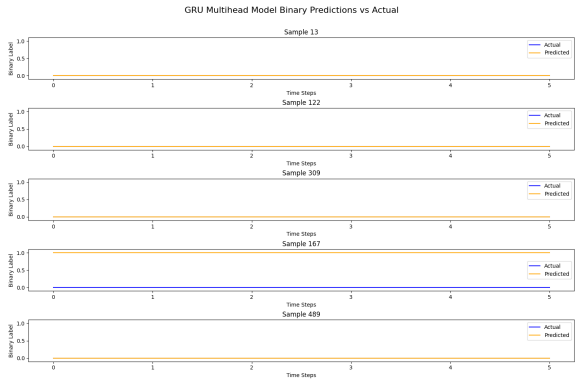
Table 26: Classification Metrics for GRU and N-BEATS Models Cluster 3

Model	Target	Class	Precision	Recall	F1-Score
GRU	30 days	Survived	0.94	1.00	0.97
		Died	0.00	0.00	0.00
		Overall Accuracy		0.9414	
	1 year	Survived	0.94	1.00	0.97
		Died	0.00	0.00	0.00
		Overall Accuracy		0.9369	
N-BEATS	30 days	Survived	0.94	1.00	0.97
		Died	0.00	0.00	0.00
		Overall Accuracy		0.9398	
	1 year	Survived	0.94	1.00	0.97
		Died	0.00	0.00	0.00
		Overall Accuracy		0.9362	

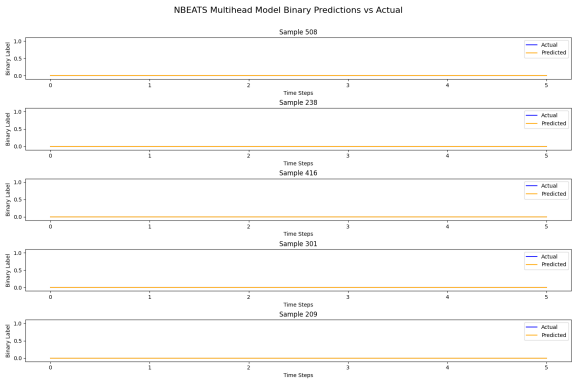
Table 27: Classification Metrics for GRU and N-BEATS Models for Female Patients

Model	Target	Class	Precision	Recall	F1-Score
GRU	30 days	Survived	0.93	1.00	0.96
		Died	0.00	0.00	0.00
		Overall Accuracy		0.9299	
	1 year	Survived	0.93	1.00	0.96
		Died	0.00	0.00	0.00
		Overall Accuracy		0.9299	
N-BEATS	30 days	Survived	0.93	1.00	0.96
		Died	0.00	0.00	0.00
		Overall Accuracy		0.9298	
	1 year	Survived	0.93	1.00	0.96
		Died	0.00	0.00	0.00
		Overall Accuracy		0.9296	

Table 28: Classification Metrics for GRU and N-BEATS Models for Male Patients

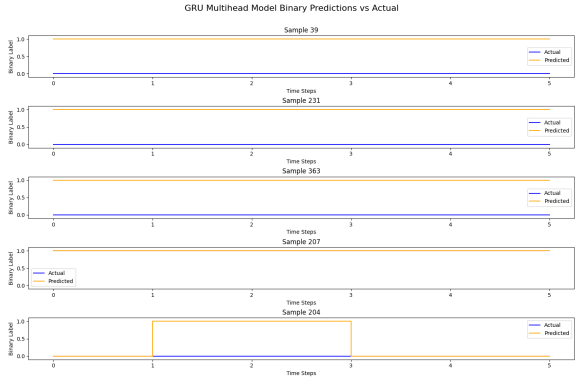


(a) Results of the multivariate forecasting using GRU.

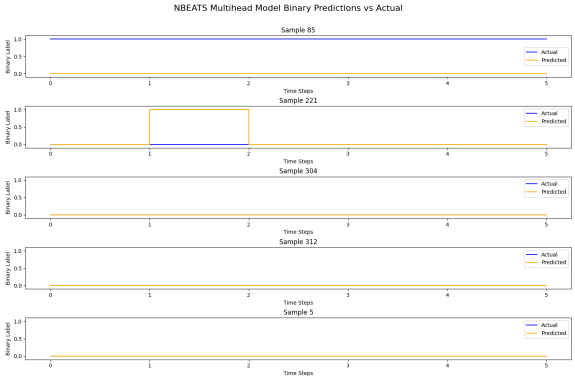


(b) Results of the multivariate forecasting using N-BEATS.

Figure 20: Results of the multivariate forecasting for the MIMIC dataset with 1-year mortality as target variable, back horizon = 12, and forecast horizon = 6, showing the accuracy of the forecasting for cluster 0.

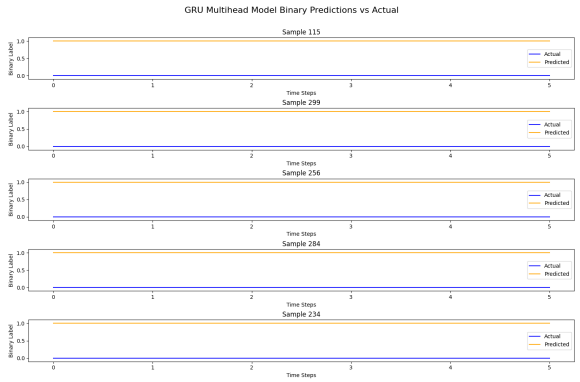


(a) Results of the multivariate forecasting using GRU.

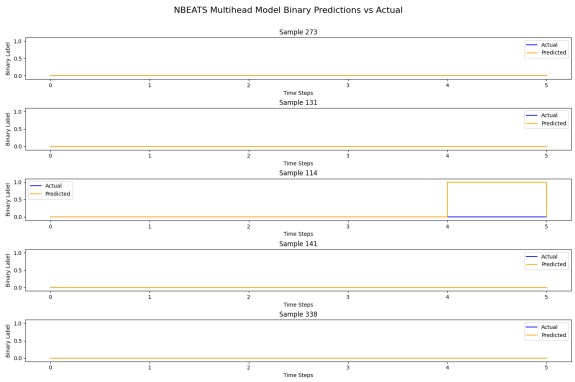


(b) Results of the multivariate forecasting using N-BEATS.

Figure 21: Results of the multivariate forecasting for the MIMIC dataset with 30-day mortality as target variable, back horizon = 12, and forecast horizon = 6, showing the accuracy of the forecasting for cluster 1.

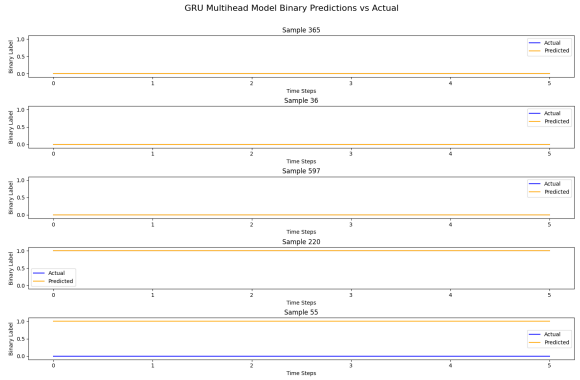


(a) Results of the multivariate forecasting using GRU.

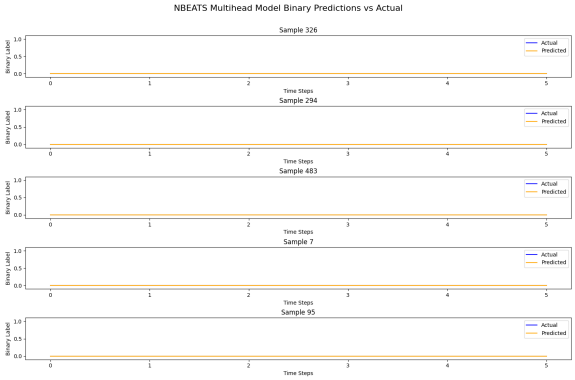


(b) Results of the multivariate forecasting using N-BEATS.

Figure 22: Results of the multivariate forecasting for the MIMIC dataset with 1-year mortality as target variable, back horizon = 12, and forecast horizon = 6, showing the accuracy of the forecasting for cluster 1.

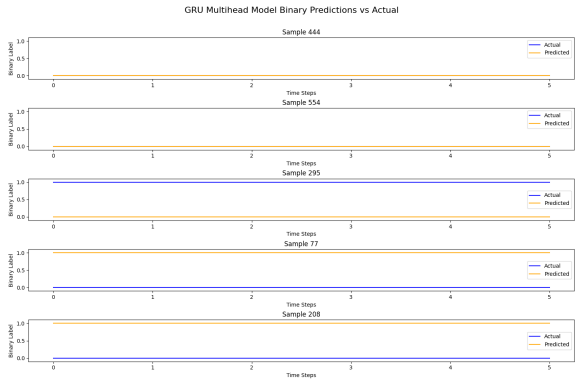


(a) Results of the multivariate forecasting using GRU.

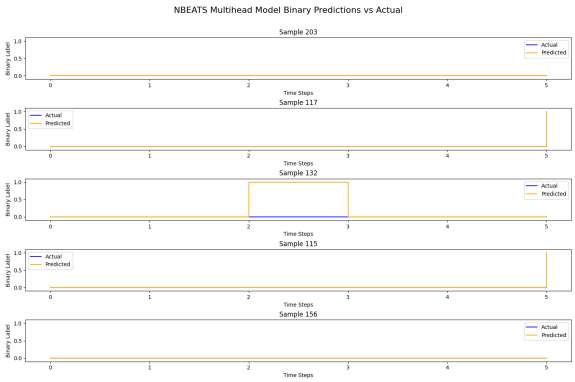


(b) Results of the multivariate forecasting using N-BEATS.

Figure 23: Results of the multivariate forecasting for the MIMIC dataset with 30-day mortality as target variable, back horizon = 12, and forecast horizon = 6, showing the accuracy of the forecasting for cluster 2.

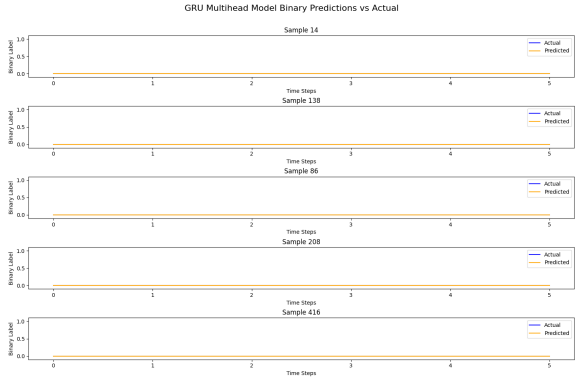


(a) Results of the multivariate forecasting using GRU.

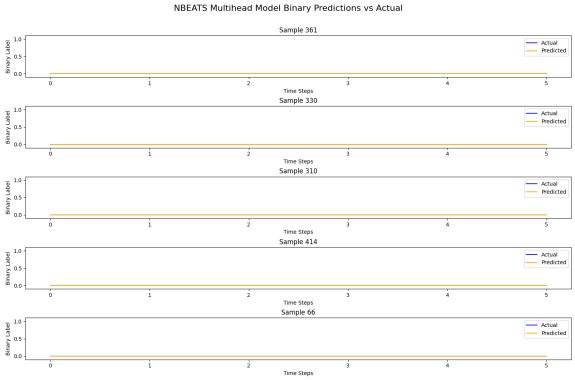


(b) Results of the multivariate forecasting using N-BEATS.

Figure 24: Results of the multivariate forecasting for the MIMIC dataset with 1-year mortality as target variable, back horizon = 12, and forecast horizon = 6, showing the accuracy of the forecasting for cluster 2.

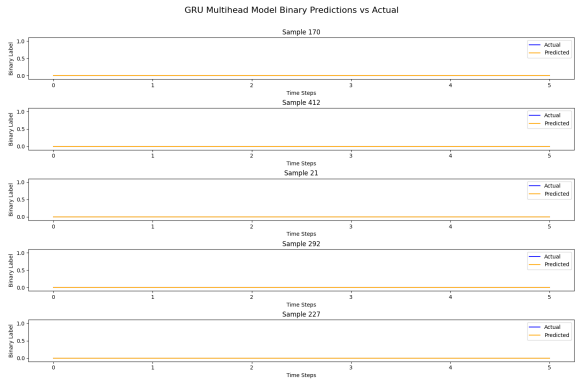


(a) Results of the multivariate forecasting using GRU.

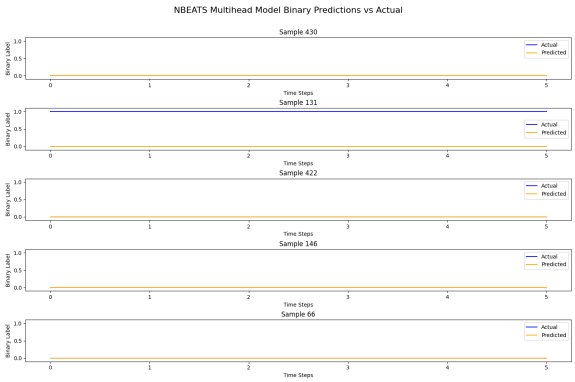


(b) Results of the multivariate forecasting using N-BEATS.

Figure 25: Results of the multivariate forecasting for the MIMIC dataset with 30-day mortality as target variable, back horizon = 12, and forecast horizon = 6, showing the accuracy of the forecasting for cluster 3.

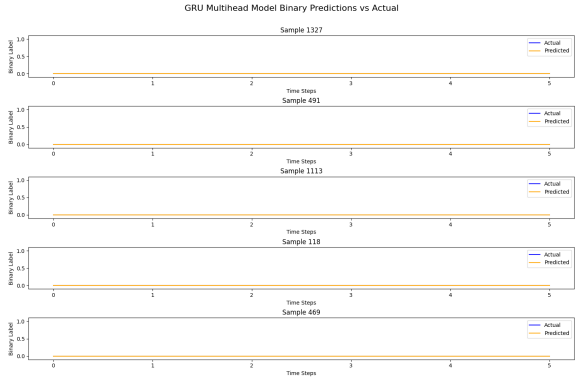


(a) Results of the multivariate forecasting using GRU.

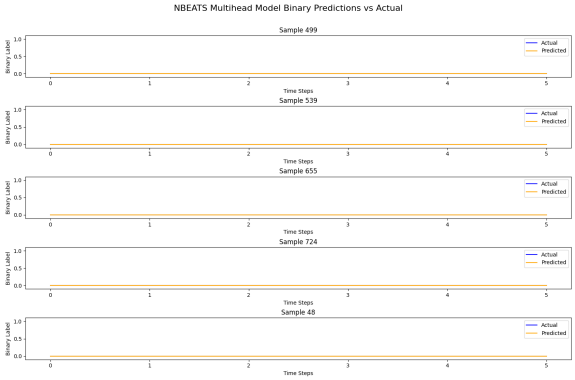


(b) Results of the multivariate forecasting using N-BEATS.

Figure 26: Results of the multivariate forecasting for the MIMIC dataset with 1-year mortality as target variable, back horizon = 12, and forecast horizon = 6, showing the accuracy of the forecasting for cluster 3.

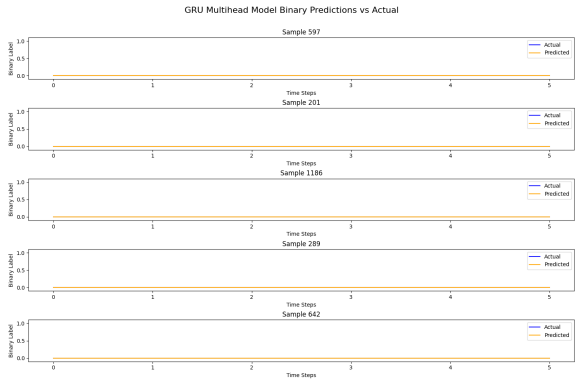


(a) Results of the multivariate forecasting using GRU.

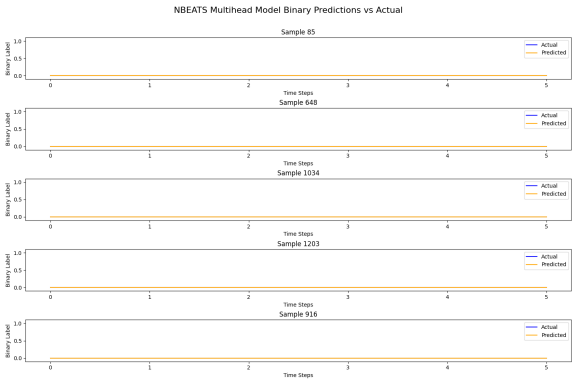


(b) Results of the multivariate forecasting using N-BEATS.

Figure 27: Results of the multivariate forecasting for the MIMIC dataset with 30-day mortality as target variable, back horizon = 12, and forecast horizon = 6, showing the accuracy of the forecasting for the female cluster.

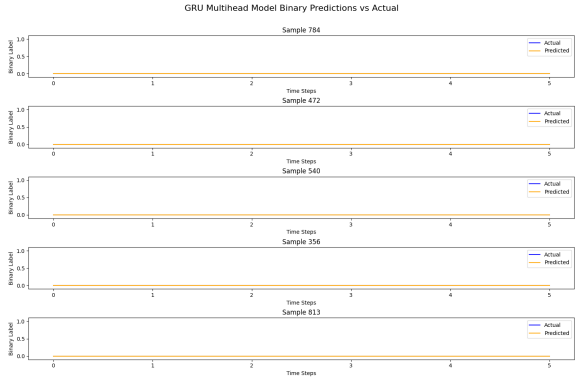


(a) Results of the multivariate forecasting using GRU.

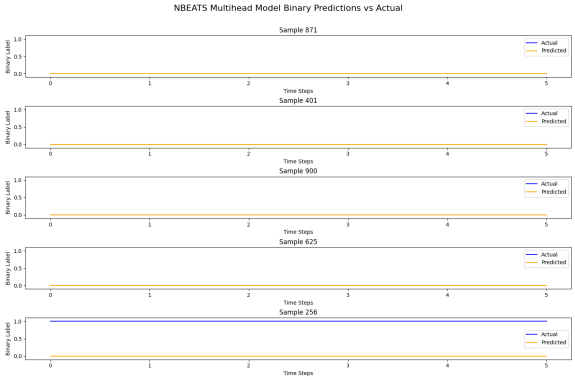


(b) Results of the multivariate forecasting using N-BEATS.

Figure 28: Results of the multivariate forecasting for the MIMIC dataset with 1-year mortality as target variable, back horizon = 12, and forecast horizon = 6, showing the accuracy of the forecasting for the female cluster.

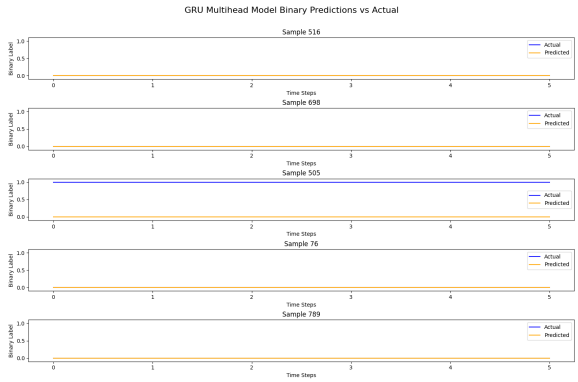


(a) Results of the multivariate forecasting using GRU.

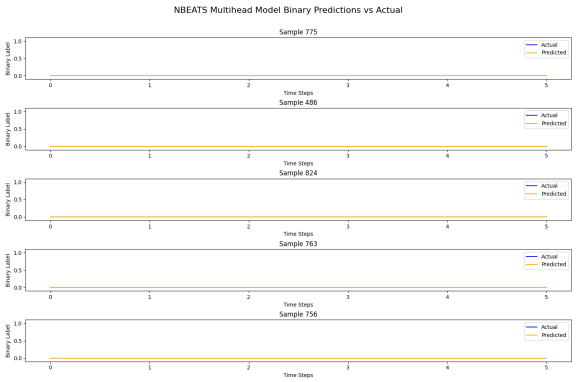


(b) Results of the multivariate forecasting using N-BEATS.

Figure 29: Results of the multivariate forecasting for the MIMIC dataset with 30-day mortality as target variable, back horizon = 12, and forecast horizon = 6, showing the accuracy of the forecasting for the male cluster.



(a) Results of the multivariate forecasting using GRU.



(b) Results of the multivariate forecasting using N-BEATS.

Figure 30: Results of the multivariate forecasting for the MIMIC dataset with 1-year mortality as target variable, back horizon = 12, and forecast horizon = 6, showing the accuracy of the forecasting for the male cluster.

C Detailed Results for the Counterfactual Generation

C.1 OhioT1DM

The following section shows the detailed results for the OhioT1DM counterfactual generation, for all four methods.

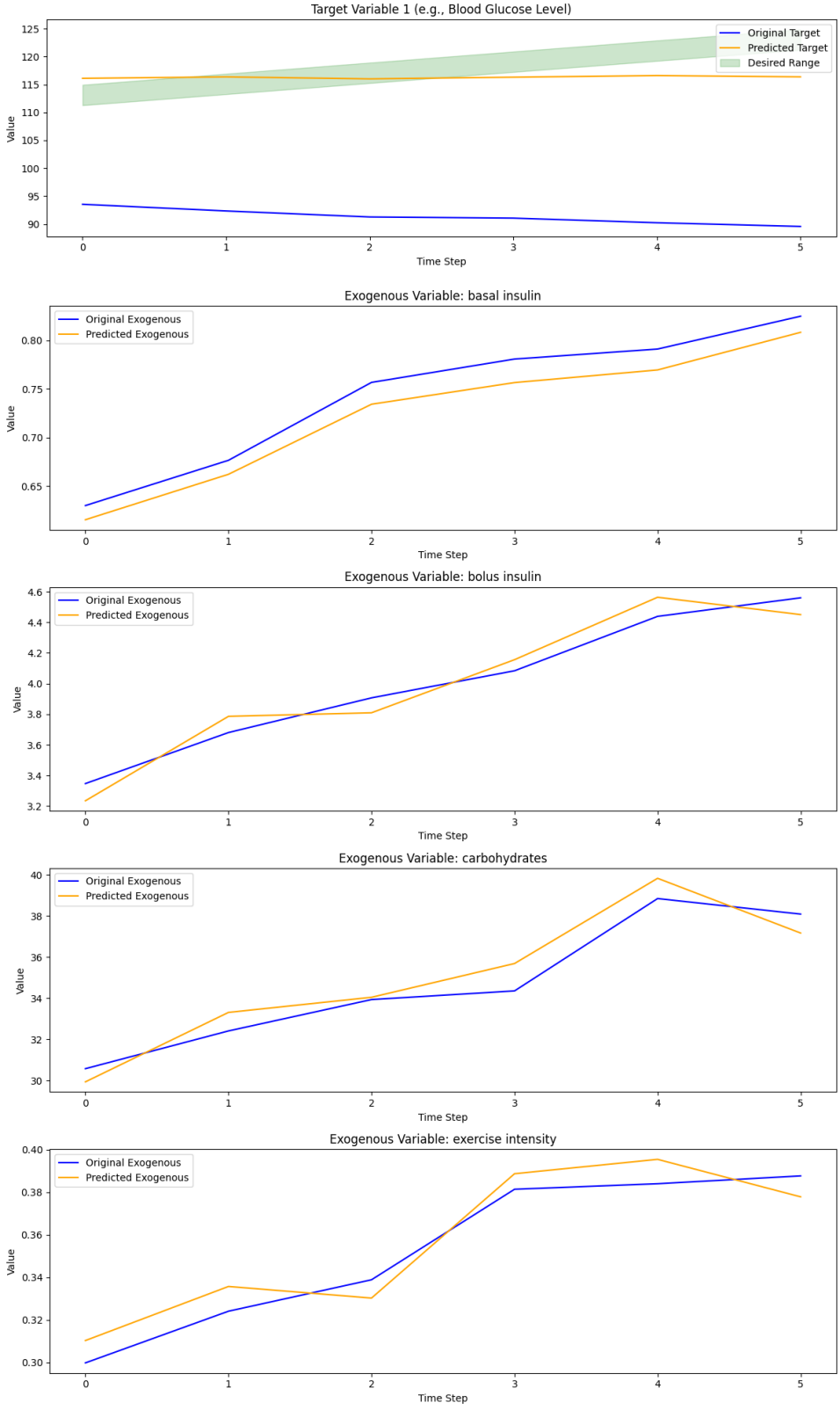


Figure 31: Example of counterfactuals generated using GRU for the OhioT1DM dataset. Top: Blood glucose levels for the original and counterfactual samples. Bottom: Comparison of exogenous variables (original in blue, counterfactual in yellow, bounds in green).

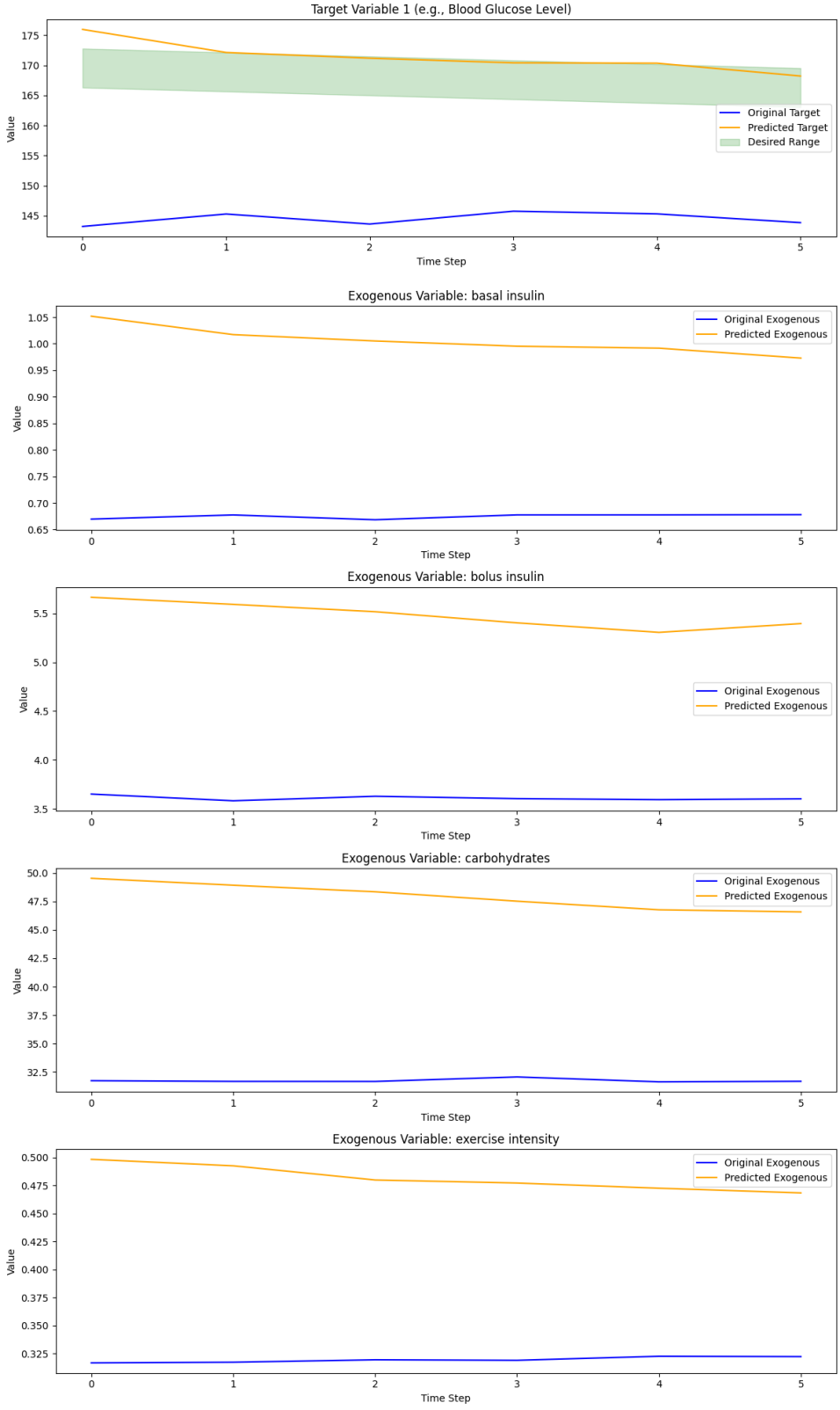


Figure 32: Example of counterfactuals generated using SARIMAX for the OhioT1DM dataset. Top: Blood glucose levels for the original and counterfactual samples. Bottom: Comparison of exogenous variables (original in blue, counterfactual in yellow, bounds in green).

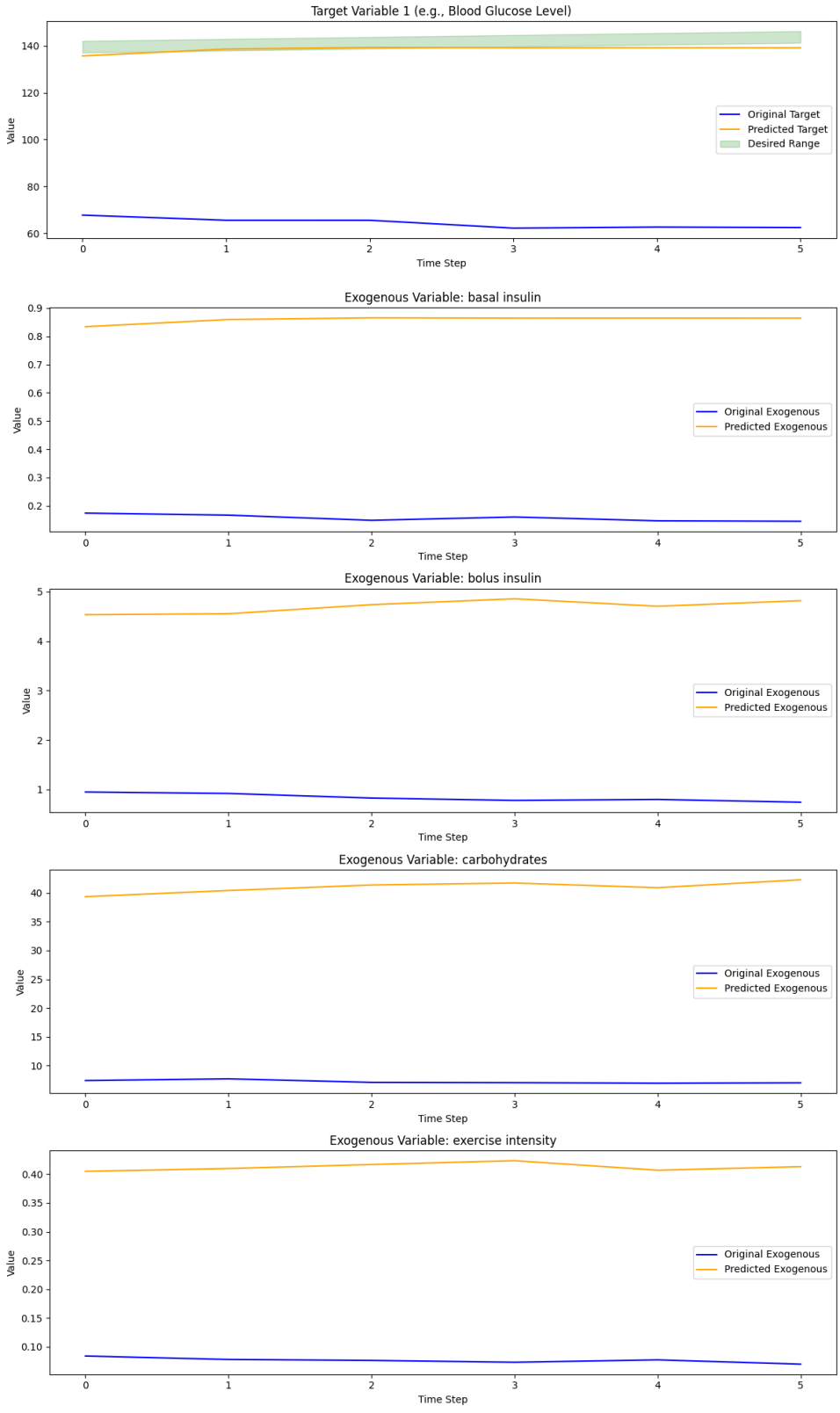


Figure 33: Example of counterfactuals generated using OLS for the OhioT1DM dataset. Top: Blood glucose levels for the original and counterfactual samples. Bottom: Comparison of exogenous variables (original in blue, counterfactual in yellow, bounds in green).

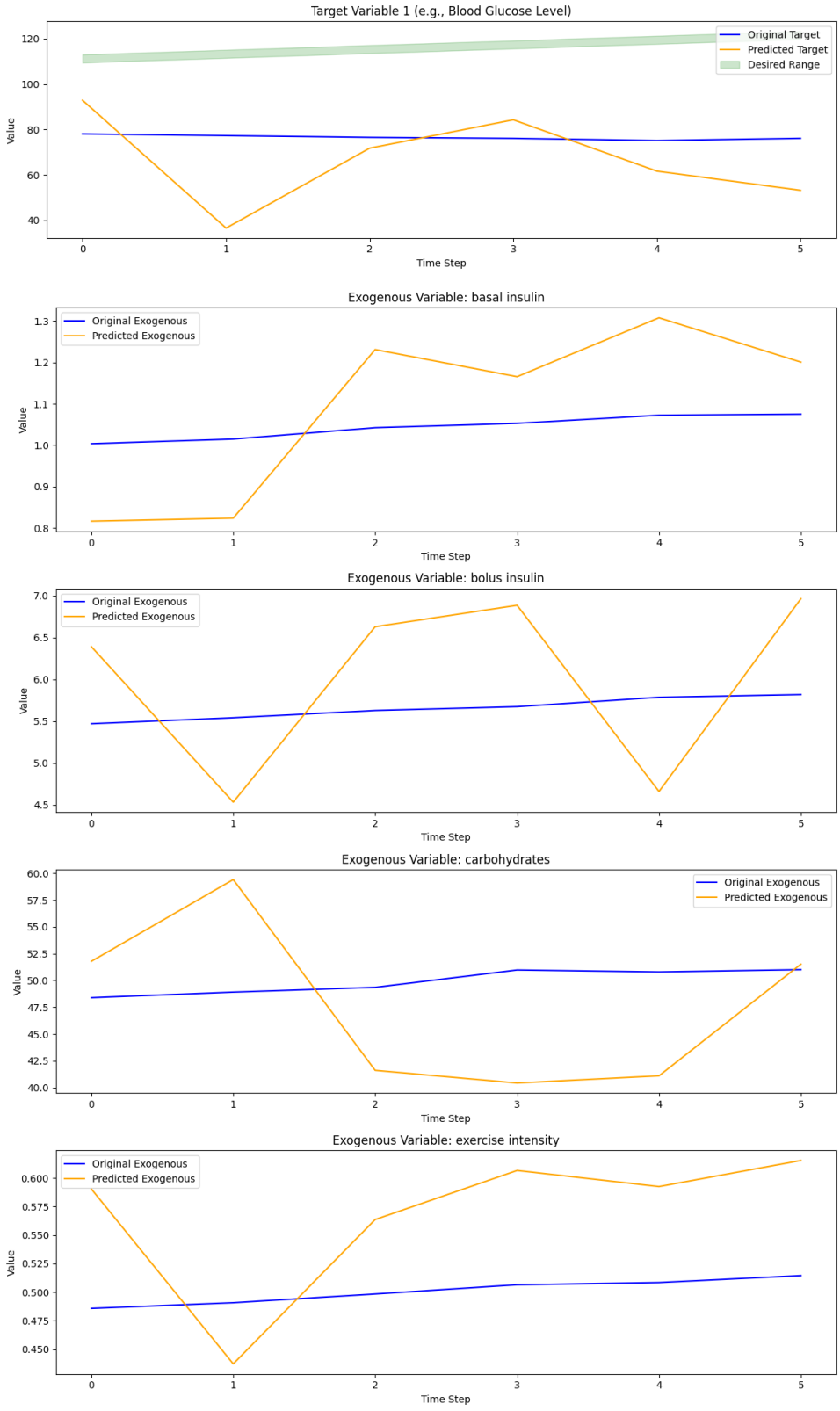


Figure 34: Example of counterfactuals generated using N-BEATS for the OhioT1DM dataset. Top: Blood glucose levels for the original and counterfactual samples. Bottom: Comparison of exogenous variables (original in blue, counterfactual in yellow, bounds in green).

C.2 MIMIC-IV

The following section shows the detailed results for the MIMIC-IV counterfactual generation, for all four methods. This includes some example visualizations for the male cluster, as well as the detailed metrics for the clustering according to comorbidities.

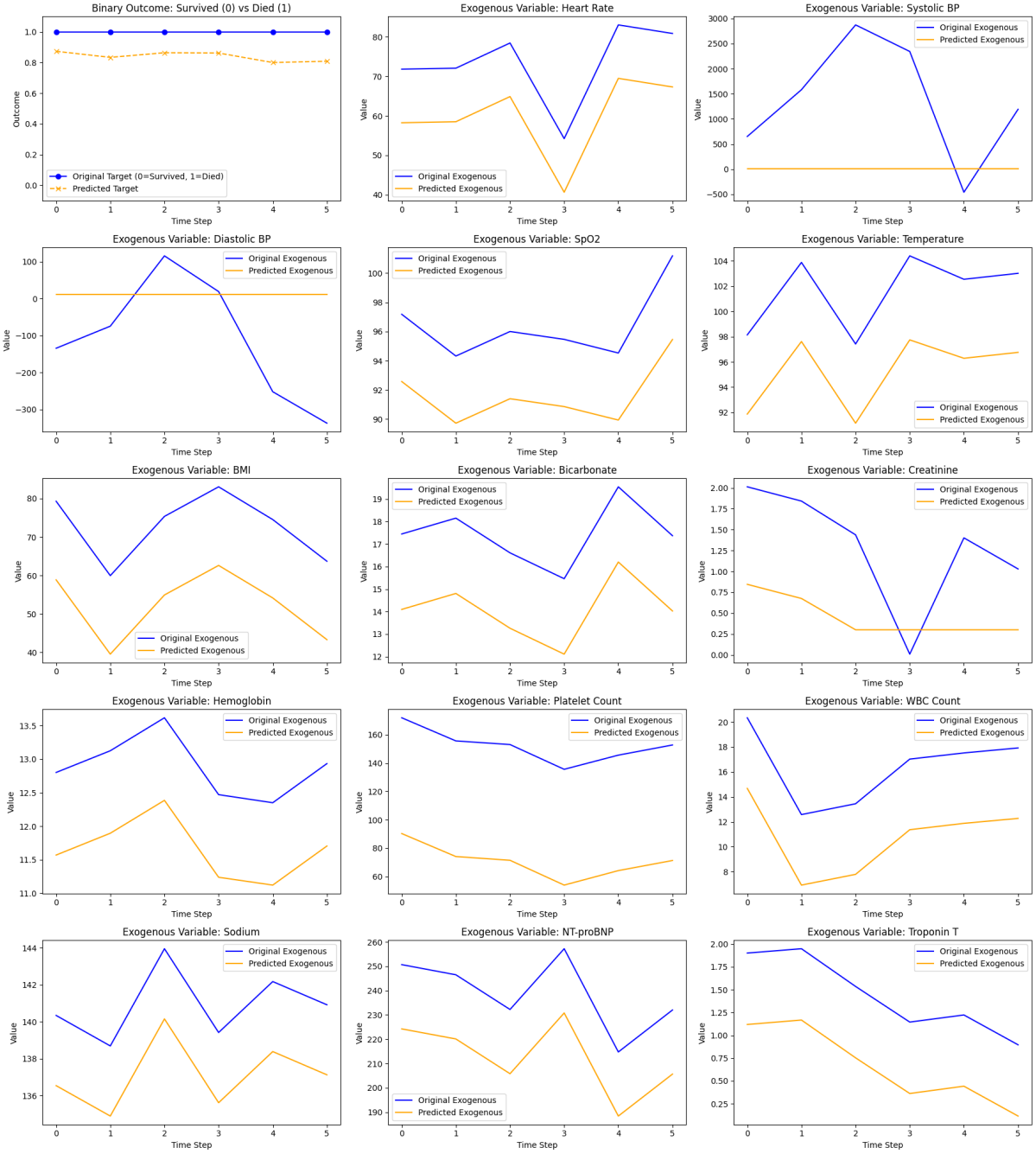


Figure 35: Example of the counterfactuals generated for the MIMIC dataset using SARIMAX, with survival. The different exogenous variables, with the original values in blue and the predicted in yellow.

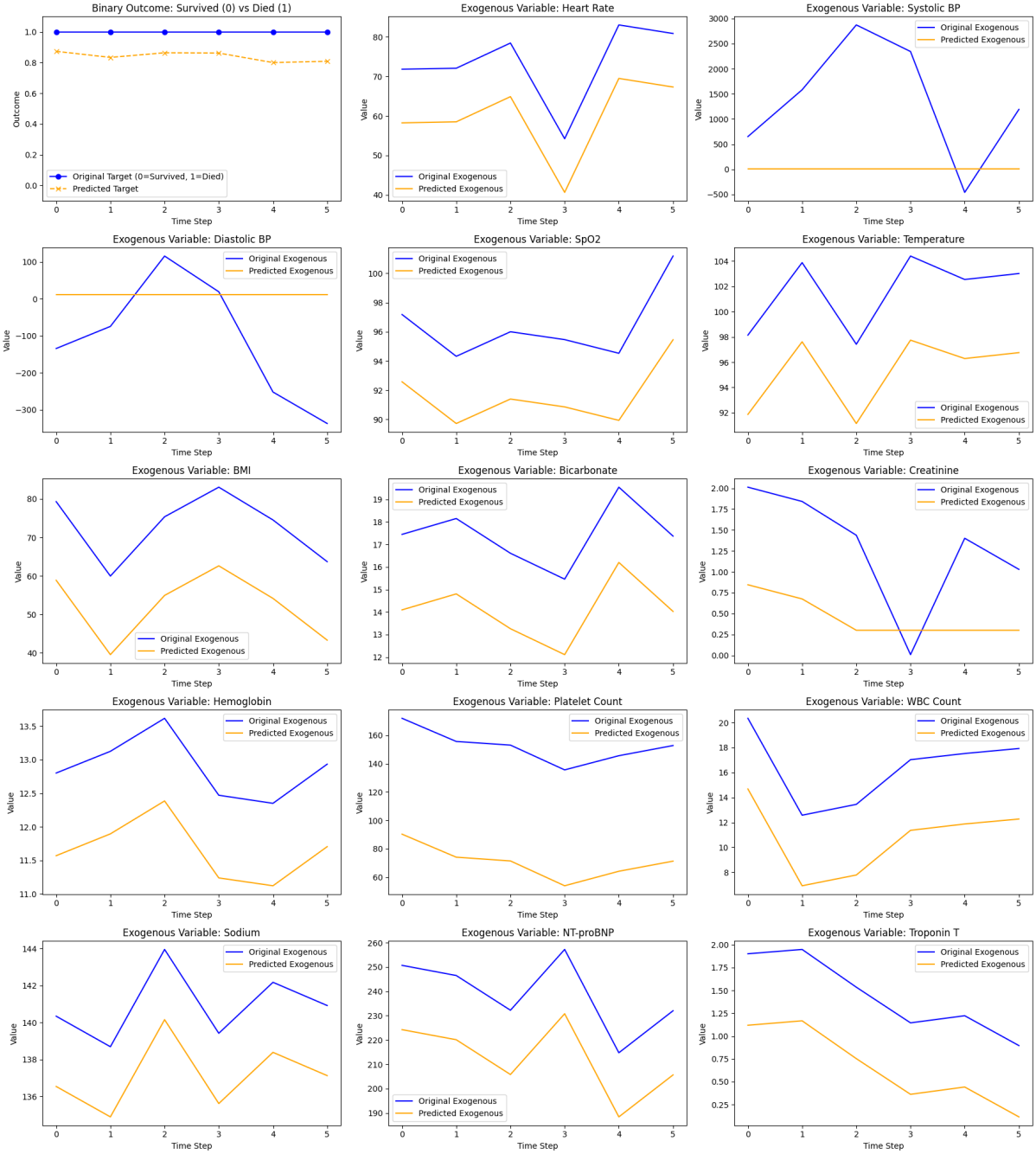


Figure 36: Example of the counterfactuals generated for the MIMIC dataset using OLS, with survival. The different exogenous variables, with the original values in blue and the predicted in yellow.

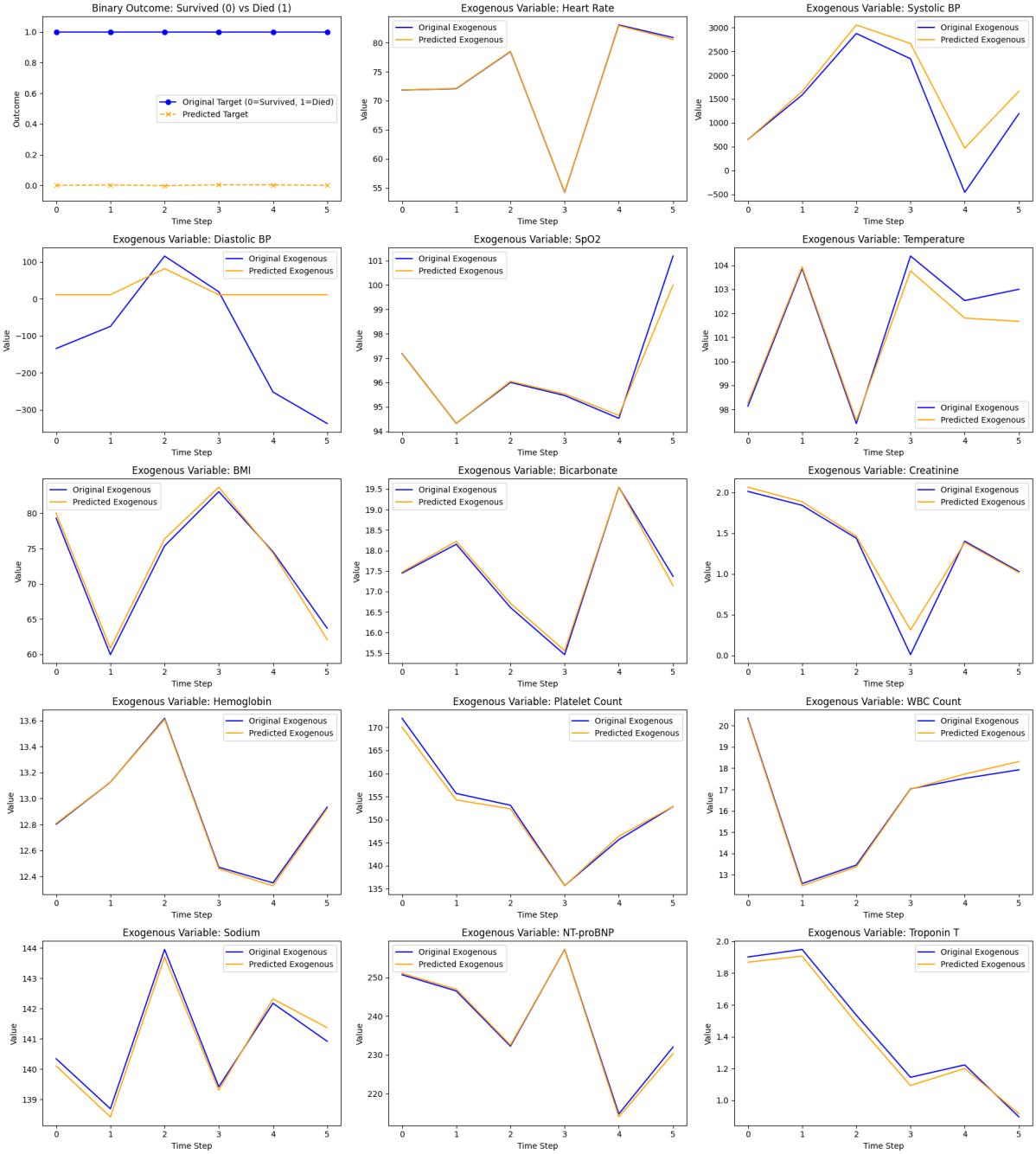


Figure 37: Example of the counterfactuals generated for the MIMIC dataset using GRU, with survival. The different exogenous variables, with the original values in blue and the predicted in yellow.

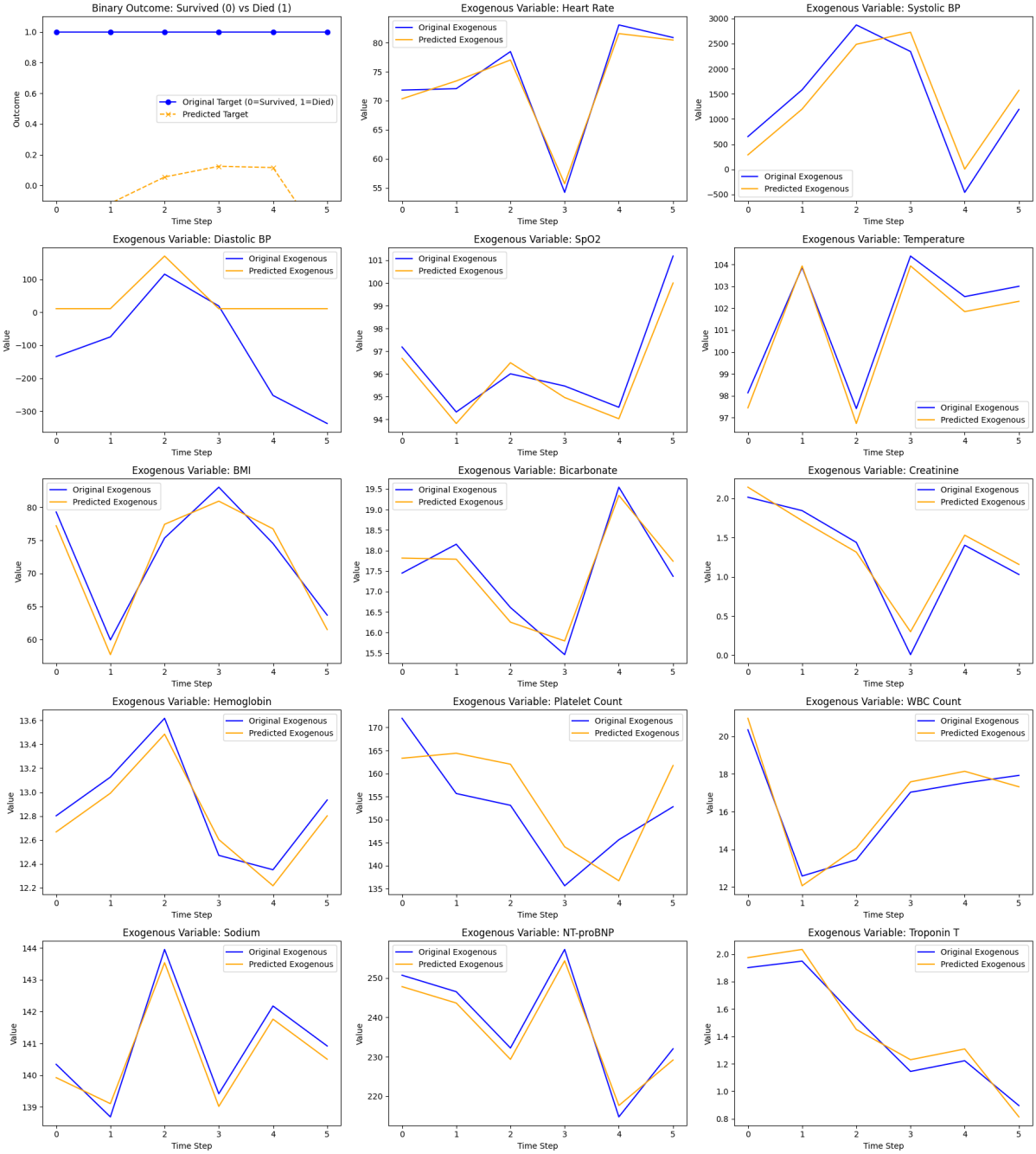


Figure 38: Example of the counterfactuals generated for the MIMIC dataset using N-BEATS, with survival. The different exogenous variables, with the original values in blue and the predicted in yellow.

Metric	GRU	SARIMAX	OLS	N-BEATS
Target Prediction Deviation				
MAE	0.00	0.06	0.06	0.02
RMSE	0.00	0.06	0.06	0.02
Max Deviation	0.04	1.09	1.09	0.32
MAPE (%)	251495.87	6293314.07	6292890.42	1973553.19
Survival Changes				
Changes to Survived ($0 \rightarrow 1$)	0	0	0	0
Changes to Dead ($1 \rightarrow 0$)	0	44	44	0
Mean % Change in Exogenous Features				
Heart Rate	99.6%	6.7%	6.3%	288.2%
Systolic BP	175.7%	8.6%	7.4%	309.4%
Diastolic BP	2.1%	2.1%	2.1%	28.0%
SpO2	3.4%	1.0%	1.0%	14.5%
Temperature	0.1%	0.1%	0.1%	0.9%
BMI	31.8%	2.0%	2.0%	75.9%
Bicarbonate	3.5%	1.2%	1.2%	19.4%
Creatinine	13.3%	3.5%	3.4%	78.2%
Hemoglobin	2.1%	0.9%	0.9%	16.2%
Platelet Count	4.5%	2.3%	2.3%	32.8%
WBC Count	40.0%	21.3%	21.3%	182.3%
Sodium	0.4%	0.3%	0.3%	4.5%
NT-proBNP	7.4%	3.4%	3.4%	55.2%
Troponin T	89.0%	8.4%	7.7%	918.7%

Table 29: Cluster 0: Comparison of counterfactual generation performance across GRU, SARIMAX, OLS, and N-BEATS. Metrics include prediction deviation, survival outcome flips, and average percent change in key exogenous clinical features. Lower deviations and smaller, targeted feature changes are desirable.

Metric	GRU	SARIMAX	OLS	N-BEATS
Target Prediction Deviation				
MAE	0.01	0.08	0.08	0.05
RMSE	0.01	0.09	0.09	0.06
Max Deviation	1.00	1.23	1.23	1.10
MAPE (%)	316022.05	7729621.81	7729251.27	4484177.31
Survival Changes				
Changes to Survived ($0 \rightarrow 1$)	4	4	4	4
Changes to Dead ($1 \rightarrow 0$)	0	54	54	0
Mean % Change in Exogenous Features				
Heart Rate	0.9%	1.4%	1.4%	3.0%
Systolic BP	0.7%	1.6%	1.6%	2.5%
Diastolic BP	251.3%	246.3%	246.3%	376.4%
SpO2	1.5%	0.8%	0.8%	2.0%
Temperature	0.1%	0.1%	0.1%	0.2%
BMI	2.8%	1.6%	1.6%	4.7%
Bicarbonate	1.2%	1.6%	1.6%	2.4%
Creatinine	113.3%	4.2%	4.2%	124.1%
Hemoglobin	1.0%	1.1%	1.1%	1.9%
Platelet Count	14.2%	9.1%	9.1%	17.4%
WBC Count	11.4%	10.3%	10.3%	16.9%
Sodium	0.1%	0.3%	0.3%	0.4%
NT-proBNP	15.8%	3.0%	3.0%	21.7%
Troponin T	67.8%	8.1%	8.1%	102.2%

Table 30: Cluster 1: Comparison of counterfactual generation performance across GRU, SARIMAX, OLS, and N-BEATS. Metrics include prediction deviation, survival outcome flips, and average percent change in key exogenous clinical features. Lower deviations and smaller, targeted feature changes are desirable.

Metric	GRU	SARIMAX	OLS	N-BEATS
Target Prediction Deviation				
MAE	0.00	0.11	0.11	0.06
RMSE	0.00	0.11	0.11	0.07
Max Deviation	0.01	1.02	1.02	0.25
MAPE (%)	181858.38	10527415.13	10527024.89	6225617.23
Survival Changes				
Changes to Survived ($0 \rightarrow 1$)	0	0	0	0
Changes to Dead ($1 \rightarrow 0$)	0	78	78	0
Mean % Change in Exogenous Features				
Heart Rate	0.2%	2.1%	2.1%	25.9%
Systolic BP	0.2%	2.0%	2.0%	25.0%
Diastolic BP	73.4%	12.7%	12.7%	652.7%
SpO2	0.3%	0.5%	0.5%	11.1%
Temperature	1.2%	1.4%	1.4%	25.9%
BMI	0.4%	2.9%	2.9%	37.5%
Bicarbonate	0.3%	1.6%	1.6%	26.1%
Creatinine	30.3%	9.1%	9.1%	231.4%
Hemoglobin	0.3%	1.8%	1.8%	25.8%
Platelet Count	0.7%	4.1%	4.1%	59.1%
WBC Count	48.2%	13.0%	13.0%	474.3%
Sodium	0.1%	0.4%	0.4%	5.9%
NT-proBNP	2.7%	4.5%	4.5%	62.7%
Troponin T	58.3%	12.0%	12.0%	3274.8%

Table 31: Cluster 2: Comparison of counterfactual generation performance across GRU, SARIMAX, OLS, and N-BEATS. Metrics include prediction deviation, survival outcome flips, and average percent change in key exogenous clinical features. Lower deviations and smaller, targeted feature changes are desirable.

Metric	GRU	SARIMAX	OLS	N-BEATS
Target Prediction Deviation				
MAE	0.00	0.04	0.04	0.04
RMSE	0.00	0.05	0.05	0.05
Max Deviation	0.01	0.87	0.87	0.16
MAPE (%)	245085.05	4497574.79	4497170.58	4457559.15
Survival Changes				
Changes to Survived ($0 \rightarrow 1$)	0	0	0	0
Changes to Dead ($1 \rightarrow 0$)	0	36	36	0
Mean % Change in Exogenous Features				
Heart Rate	3.7%	1.2%	1.2%	7.6%
Systolic BP	29.9%	0.8%	0.8%	32.9%
Diastolic BP	3.8%	1.5%	1.5%	8.9%
SpO2	2.9%	0.7%	0.7%	3.8%
Temperature	2.5%	0.7%	0.7%	3.9%
BMI	3.1%	1.0%	1.0%	6.8%
Bicarbonate	0.9%	1.0%	1.0%	4.1%
Creatinine	161.8%	4.0%	4.0%	160.0%
Hemoglobin	0.6%	0.7%	0.7%	2.6%
Platelet Count	3.1%	1.5%	1.5%	10.9%
WBC Count	2.0%	1.9%	1.9%	10.5%
Sodium	0.1%	0.2%	0.2%	0.7%
NT-proBNP	13.2%	5.0%	5.0%	18.1%
Troponin T	48.4%	5.8%	5.8%	93.1%

Table 32: Cluster 3: Comparison of counterfactual generation performance across GRU, SARIMAX, OLS, and N-BEATS models. Includes prediction deviation metrics, survival flips, and mean percentage changes in key exogenous clinical features.