Universidad Rey Juan Carlos

TESIS DOCTORAL

Studies on Genetic Programming Techniques for the Short and Medium Term Predictions of the Interstitial Glucose of Diabetic Patients

Autor:

Sergio Contador Pachón

Directores:

José Ignacio Hidalgo Pérez
José Manuel Velasco Cabo

Programa de Doctorado en Tecnologías de la Información y las Comunicaciones

Escuela Internacional de Doctorado

2022

Acknowledgements

To my family:

- My mother and father, Manuela and Florentino, for their unconditional support throughout my life. Without them this work would simply not exist.
- My aunt and uncle, Beatriz and Richard, for their interest in my work and for enlightening me with the English language.
- My grandmother and grandfather, Maria Paz and Florentino, for their wisdom and everything they have taught me.

To my directors, José Ignacio and José Manuel Velasco, and my advisor, José Manuel Colmenar, for their dedication to their work and mine; for their patience and knowledge; for all the hours that they have spent with me preparing this thesis.

To the Rey Juan Carlos University and the Complutense University of Madrid for giving me the opportunity to do this thesis, and providing it with the human and material resources necessary to carry it out.

To the institutions that have partially funded this work:

- Operational Programmes of the European Social Fund and the European Development Regional Fund of the Community of Madrid - projects B2017/BMD3773 (GenObIA-CM) and Y2018/NMT-4668 (Micro-Stress-MAP-CM).
- Spanish Ministerio de Innovación, Ciencia y Universidad - grant RTI2018-095180-B-I00.
- Eugenio Rodríguez Pascual Foundation - development of adaptive and bioinspired systems for glycemic control with continuous subcutaneous insulin infusers and continuous glucose monitoring systems.

And finally to the readers of this thesis, my thanks.

Agradecimientos

A mi familia:

- Mi madre y mi padre, Manuela y Florentino, por su apoyo incondicional a lo largo de mi vida. Simplemente sin ellos este trabajo no existiría.
- Mi tía y mi tío, Beatriz y Richard, por su interés en mi trabajo y por iluminarme con el idioma inglés.
- Mi abuela y mi abuelo, María Paz y Florentino, por su sabiduría y todo lo que me han enseñado.

A mis directores, José Ignacio y José Manuel Velasco, y a mi tutor, José Manuel Colmenar, por su dedicación a su trabajo y al mío; por su paciencia y conocimiento; por todas las horas que han pasado conmigo elaborando esta tesis.

A la Universidad Rey Juan Carlos y a la Universidad Complutense de Madrid por darme la oportunidad de hacer esta tesis, y dotarla de los recursos humanos y materiales necesarios para llevarla a cabo.

A las instituciones que han financiado parcialmente este trabajo:

- Programas Operativos del Fondo Social Europeo y del Fondo Europeo de Desarrollo Regional de la Comunidad de Madrid - proyectos B2017/BMD3773 (GenObIA-CM) y Y2018/NMT-4668 (Micro-Stress-MAP-CM).
- Ministerio de Innovación, Ciencia y Universidad de España - grant RTI2018-095180-B-I00.
- Fundación Eugenio Rodríguez Pascual - development of adaptive and bioinspired systems for glycemic control with continuous subcutaneous insulin infusers and continuous glucose monitoring systems.

Y finalmente a los lectores de esta tesis, gracias.

Abstract

Background

Diabetes Mellitus (DM) is a chronic disease that increases the morbidity and mortality, and causes a significant deterioration in the quality of life. There are mainly two types of diabetes:

- Type 1 Diabetes Mellitus (T1DM): due to an autoimmune process, the pancreas is not able to generate enough insulin to process the sugar produced after carbohydrate intake.
- Type 2 Diabetes Mellitus (T2DM): the insulin produced by the pancreas does not work properly, in a phenomenon known as insulin resistance.

T1DM can only be treated with synthetic insulin injected into the bloodstream. Depending on the amount of insulin in the body two scenarios can happen. On the one hand, an excess of insulin can cause hypoglycemia, defined as a Blood Glucose (BG) value of less than 70 mg/dl. If this situation continues over time, it can cause short-term complications. On the other hand, if the insulin dose is insufficient, it can lead to hyperglycemia, defined as a BG value greater than 180 mg/dl, which can lead to long-term complications. The goal is to keep BG levels within the target range most of the time (defined as a BG value between [70, 180] mg/dl).

BG control in insulin-dependent patients requires predicting future glucose values to determine the amount of insulin to inject. This amount depends on many factors that People with Diabetes (PwD) have to estimate manually, following two different primary therapies: Continuous Subcutaneous Insulin Infusion (CSII) and Multiple Doses of Insulin (MDI). One of the most important factors is Glucose Variability (GV), defined as the fluctuation of glucose levels during a day. GV makes the prediction process more complicated.

There are different types of BG control strategies (*Manual*, *Semi-automated* and *Automated* solutions), but in all of them, it is important to develop mathematical models or Artificial Intelligence (AI) systems that describe the interaction between the glucose system and insulin using the measurements and stored data.

Different ways to produce models for BG prediction are used depending on the information available at the time of the forecast. One option is the *What-if* scenario, where the model predicts future glucose values by taking into account not only past values of the three variables (glucose, carbohydrates, and insulin) but also future carbohydrate intakes and insulin injections from the present until the prediction horizon. For instance, the model can predict the BG level, supposing that the patient eats a certain amount of carbohydrates m minutes from the prediction time. Another option is the *Agnostic* scenario, that produces model without information about future events in the prediction phase. This type of model needs to implicitly predict those events. For example, the model must identify the fasting periods or the physical exercise. Both scenarios are used in this thesis.

There are three main methods commonly used to solve the glucose prediction problem. The first, called *Physiological* models, are linear mathematical models that simulate the physiology of the glucose-insulin regulatory system, and require specific *Physiological* knowledge. The second method is *Data-driven* models, which can predict glucose concentration based only on existing input and output data. The last method combines both solutions in a *Hybrid* way, where the models take the simplest parts of glucose-insulin physiology and include data to determine the parameters of the models. All the models created in this thesis are based on *Data-driven* method.

There are some research based on Data Mining and Machine Learning (ML) techniques using decision trees studying the problem of glucose prediction and insulin recommendation in PwD. One of the most

decision tree algorithms used is CHi-square Automatic Interaction Detection (CHAID). This algorithm recursively divides the data by a response/objective variable using multiple divisions between the different input/predictor variables. The algorithm computes the Chi-test through the predictor. A Chi-square test yields a probability value as a result lying anywhere between 0 and 1. A value closer to 0 indicates a significant difference between the two classes being compared, and a value more relative to 1 means no significant difference among them. If the test is not statistically significant, the algorithm determines the predictor which is least significantly different corcerning the criterion variable and merge the predictor categories. Else, it computes the Bonferroni adjusted p-value for the set of categories of the predictor. Bonferroni correction is used to counteract multiple classes to find the class which did the next best split. If the test is not statistically significant, the predictor variable with the smallest p-value will be considered for the next split. Else, the respective node will become a terminal node. The algorithm repeats this process until no further splits can be performanced.

Another works applied solutions based on evolutionary techniques such as Genetic Programming (GP) or Grammatical Evolution (GE). GP is an evolutionary computation technique that automatically solves problems without requiring the user to know the structure of the solution in advance. In GP we evolve a population of computer programs. The algorithm randomly creates an initial population of programs (individuals or solutions). Next, it executes each program and ascertain its fitness to select one or two programs from the population which become the parent programs of the process. A child program is created by combining randomly chosen parts from two selected parent programs (crossover). Another new child program is created by randomly altering a randomly chosen part of a parent program (mutation). The algorithm repeats this process until an acceptable solution is found or some other stopping condition is met. Finally, it returns the best-so-far individual (solution). GE is a GP technique that offers a solution using a specific grammar. The mathematical expression of the solution is achieved by decoding it through the grammar. Therefore the search space can be restricted, and domain knowledge of the problem can be incorporated. Usually the grammar is implemented in Backus-Naur Form (BNF) format designed to find a predictive model for future BG levels.

Clarke Error Grid (CEG) method was published in 1987 as a technique to evaluate clinical accuracy of Continuous Glucose Monitoring (CGM) systems. Parkes Error Grid (PEG) was published in 2000 as an alternative to CEG. Both methods can be used to represent the differences between the values estimated in a prediction and the current or reference values in a graph with Cartesian coordinates, where the X-axis represents the reference values and the Y-axis the values of the prediction, and Y = X is the ideal prediction. The unique feature of this representation is that the graph is divided into five zones depending on the degree of accuracy of the glucose estimates. The difference between the two methods lies in the definition of the zones. In CEG the zones are defined as follows: zone A represents glucose values that deviate from the reference by no more than 20%; zone B are points that are outside of 20% but would not lead to inappropriate treatment; zone C are points that would result in over-correcting acceptable BG levels leading to unnecessary treatment; zone D are points indicating a potentially dangerous failure to detect hypoglycemia or hyperglycemia; zone E are points that would confuse treatment between hypoglycemia and hyperglycemia. In PEG, the zones are redefined based in the zones of CEG and in the limits established by 100 medical experts in diabetes in a survey carried out in the *American Diabetes Association Meetings* in June 1994. In both alternatives, the goal is to maximize the predictions included in zones A and B and minimize those in zones C, D, and E.

Objectives

The main objectives of this thesis are to improve the state-of-the-art models, in terms of accuracy and effectiveness, applied in the prediction of BG levels in T1DM patients and increase the time horizon of reliable prediction. To carry out these primary objectives, we will take on the following sub-objectives:

- To demonstrate that using statistical techniques, it is possible to extract a small number of glucose profiles that characterize the majority of glucose patterns in patients with T1DM.
- To create a new methodology to obtain accurate predictors of glucose using a three-step process: (i)

data is collected using CGM, preprocessed and divided using division criteria; (ii) data is clustered to obtained customized models for different glucose profiles; (iii) a training step is used to create ensemble prediction models based on evolutionary computation.

- To design a *Data-driven* modeling approach that generates prediction models based on evolutionary algorithms that take into account both classified glucose profiles and several latent GV measures as new input features of the modeling engine.
- To investigate the benefits of applying a multi-objective approach to solve a Symbolic Regression (SR) problem based on Non-dominated Sorting Genetic Algorithm (NSGA-II) via GE, testing two different scenarios: *What-if* and *Agnostic* (the most common in daily clinical practice).

Methodology

Below we describe the general methodology used in this thesis.

First, glucose, insulin and carbohydrates data are collected from different T1DM patients. The collection is made by CGM systems, and includes information on Interstitial Glucose (IG), injected units of insulin, and carbohydrate intake.

Next step is preprocessing. Glucose data is imputated by performing a correction of the values using segmented spline interpolation of degree 3, where the maximum number of consecutive imputated values is one hour. In the case of insulin doses, a function to simulate the plasma dynamics after subcutaneous injection of insulin based on Berger model is used.

Once the data is curated, organized, and synchronized by timestamps, models based on evolutionary algorithms are created. We tackle the problem as a SR problem. Models are generated using time series of glucose, insulin, carbohydrates, and a set of historical and future features generated by averaging and aggregating the values in different periods of time. 10-fold cross-validation is used in the training step. The models are evaluated using two different metrics that calculate the accuracy of the predictions for each time horizon.

Finally, to select the suitable test for a statistical assessment of the results, we analyse the distribution of the results using a Kernel Density Estimation (KDE) of the distribution of the samples. To make a deeper statistical study of the results, we follow the Bayesian model based on the Plackett-Luce distribution over rankings to analyze multiple methods in multiple time horizons.

Next, we describe the variants of the general methodology in each of the works presented in this thesis.

In the work presented in chapter 3, after the preprocessing step, data is divided according to a criterion that creates different glucose profiles. The division criteria are based on the behavior of glucose time series under study. In this research, we split time series in groups obtained by the variable day of the week into seven categories (Monday; Tuesday; Wednesday; Thursday; Friday; Saturday; Sunday) and by the variable time slot into six categories defined as the division of glucose values into sections of four hours each (00:00h-04:00h; 04:00h-08:00h; 08:00h-12:00h; 12:00h-16:00h; 16:00h-20:00h; 20:00h-24:00h). A clustering algorithm is applied to explore the data and get hidden information. Here CHAID is applied to recursively divide the data concerning a target variable (glucose) using multiple divisions between the different input variables (day of the week and time slot). The objective is to obtained glucose profiles to train specific models on each profile.

We generate models based on GP using two different methods in chapter 4. The first, called *CHAID-GP*, are models created with GP and classified glucose values obtained in the work presented in chapter 3. The second, called *GP*, are models created with GP and no classified glucose values. The best model of each method is selected using the Akaike Information Criterion (AIC). This method is used to measure the relative quality of models based on the entropy of information. One model for each time horizon and fold is selected. Models for predicting glucose values at intervals of 30 minutes for a maximum of four hours are obtained (time horizons at 30, 60, 90, 120, 150, 180, 210 and 240 minutes). PEG method is used to evaluate the accuracy of the predictions for each time horizon. The idea is to compare these two methods to find out if using a clustering step, the models are able to improve the accuracy in the prediction of glucose values for different time horizons.

After the preprocessing step, a set of additional features are generated based on different measures of GV in the study presented in chapter 5. Again, after all the Latent Variables (LV) have been generated, data is divided and clustering according to a criteria that creates different glucose profiles. The division criteria and the clustering algorithm used are the same as applied in chapter 3. A data augmentation algorithm that generates synthetic glucose time series is used to train datasets for developing meaningful information as well as significantly enhance data quality. We generate models based on GP using two new methods. First, called *LV-GP*, are models created with GP and Latent Glucose Variability (LGV) features. Second, called *LV-CHAID-GP*, are models created with GP, LGV features, and classified glucose values. Once more, we use AIC to select the best model for each method, and the same different time horizons as in chapter 4. An ensemble method is used to create the predictor. Finally, the predictor is selected to forecast future glucose value. The main objective is to investigate whether or not the use of LGV in combination with classified glucose, improve the accuracy in the prediction of glucose values.

Finally, in the research presented in chapter 6, we analyze two different scenarios: *What-if* and *Agnostic*. For both scenarios, a single-objective (GE) model and a multi-objective (MO-GE) approach based on the NSGA-II are used, considering the Root Mean Square Error (RMSE) and an ad-hoc fitness function called F_{CLARKE} as objectives. This ad-hoc function is based on the CEG analysis, which helps show the potential danger of mispredictions in PwD. In addition to the GE models, we have also fitted Auto Regressive Integrated Moving Average (ARIMA) models to estimate the glucose values for the four prediction horizons (30, 60, 90, and 120 minutes). For the *Agnostic* scenario, the dataset includes, in addition to the information of the dataset used in previous works and in the *What-if* scenario, data collected from an activity band. Particularly, Galvanic Skin Response (GSR), skin temperature, and acceleration. We compare two options regarding the information available at the time of prediction (60 and 120 minutes).

Results

We present in chapter 3 the results of glucose patterns using clustering techniques. A retrospective study of ten patients with T1DM is performed. Significant differences were found in glucose profiles classified by the variables day of the week and time slot in each patient. There are some patients with differences among the glucose profiles obtained for each day. In other patients the opposite happens, and the glucose profiles obtained are similar.

The results for glucose prediction using clustering and GP are shown in chapter 4. In this work we used the same dataset as in chapter 3. A comparison between *CHAID-GP* and *GP* is made. We studied the ratio of predictions in zone A+B for the different clusters of data and time horizons for models created with *CHAID-GP* and *GP*. Results show that models created with *CHAID-GP* are more accurate or at least of equal accuracy than models created with *GP*. Finally, we performanced an analysis of the predictions in the evaluation phase obtained for the zone A+B with PEG for all time horizons and patients. In general, the accuracy of predictions is better for shorter time horizons and gradually gets worse as the time horizon increases.

The contribution of LGV features are exposed in chapter 5. The same dataset as in chapters 3 and 4 is used. We introduced two new methods called *LV-GP* and *LV-CHAID-GP*. We studied the ratio among the values obtained in zone A+B for the different clusters of data and time horizons between *GP* and *LV-CHAID-GP*, and between *GP* and *CHAID-GP*, respectively. Results show that models created with *CHAID-GP* and *LV-CHAID-GP* are more accurate or at least of equal accuracy than models created with *GP*. We analysed the different solutions obtained with PEG for all patients with all methods for all time horizons. It should be noted that, as expected, when time horizon increases, the points are more scattered, obtaining a greater number of points in undesirable zones (C, D and E). We can also see that *LV-CHAID-GP* is not giving predictions in the worst zone E and only a few points in zone D. We also compared the predictions in the evaluation phase obtained for the zone A+B with PEG for all time horizons. Models created with *LV-GP* and *LV-CHAID-GP* are more accurate. We also performanced an analysis of the 25 most significant relative importance variables, regarding their inclusion in the models, out of a total of 83. The MEAN, Percentage Spent in Target Range (PSTR), and J index (JI) measures appear in the top ten positions. For the non-LV variables, glucose, insulin, and carbohydrates appear in the top five positions.

We analysed the distributions of the results for all methods and time horizons. The data is not distributed according to a Gaussian distribution and is multi-modal, so a non-parametric test is necessary. Next, we selected the Nemenyi test to apply an all pairwise comparison to the results. The test determines the Critical Difference (CD). Results show significant differences for all time horizons (except for 150 minutes). Finally, we studied the Bayesian test for all methods and time horizons. Results show that *LV-GP* has the highest probability of being the best for 30, 60, 90, and 120 minutes. For 150 and 180 minutes, *CHAID-GP* is the method with the highest probability of being the best.

In chapter 6 we present the results for glucose prediction using multi-objective GE. Ten T1DM patients have been selected for the *What-if* scenario, based on conditions of reasonable glucose control. We studied the solutions obtained with both the single-objective (GE) evolution approach and the solutions obtained with the multi-objective (MO-GE) method, and each point represents a solution referenced by its coordinates (CDE, RMSE). We can observed that solutions developed by MO-GE dominate all the solutions generated by GE. We also studied the distribution of the solutions through the different horizons with both approaches GE and MO-GE. As expected, the error and the number of solutions in dangerous prediction zones increase with the prediction horizon. An additional analysis of the solutions taking into account the different folds of the 10-fold cross-validation is made. It should be noted that some of the folds are more difficult to solve than others, and a deeper study of the data should be done to improve the algorithm. We analyzed the solutions quantitatively by comparing the 40 different instances (ten patients by four different time horizons) for both GE and MO-GE approaches. Solutions obtained with the MO-GE method dominate solutions obtained with GE. We performanced an analysis on the aggregated results for all the different methods: GE, MO-GE and ARIMA. Results show that MO-GE algorithm reaches the best performance, reducing predictions in the most dangerous zones D and E for 30 and 60 minutes. ARIMA gets the worst results for all horizons except 30 minutes, where it gets better results than the GE approach.

The six patients of the dataset used in the *Agnostic* scenario were selected from the OhioT1DM dataset for BG Level Prediction: update 2020. We examined the differences in the MO-GE approach when compared to GE taking into account the historical information WS=$\{60, 120\}$ minutes. There are some cases where solutions obtained with GE are dominated by solutions generated with MO-GE. There are other cases where solutions obtained with MO-GE are dominated by solutions generated with GE. We studied the results comparing both historical values. We found solutions in the Pareto front for the different historical values and time horizons, so there is no dominance between methods. We also analyzed the solutions quantitatively by comparing the 24 different instances (six patients by four different time horizons) for both GE and MO-GE and the two historical values. The solutions obtained with MO-GE dominate the solutions obtained with GE. We also examined the aggregated results for ws=120 minutes for GE, MO-GE and ARIMA methods. As well as in the *What-if* scenario, the MO-GE algorithm reaches the best performance in all horizons except 30 minutes, where ARIMA gets the best results.

To analyze the complexity of the solutions in terms of the number of parameters and length of the solutions, we represent the average RMSE for each number of parameters and the length of the models found in the solutions for both GE and MO-GE methods and for both the *What-if* and *Agnostic* scenarios. We have observed that solutions obtained with GE have a greater length, a greater number of parameters, and a higher RMSE than solutions obtained with MO-GE in both scenarios. To study how different historical values WS={60, 120} minutes contributed to the models, a deeper analysis is done to assess the statistical significance of the results. We first created density plots using a KDE for the distribution of the samples. Data distribution is non-unimodal, and a non-parametric test is necessary. Then, we follow the Bayesian model over historical values and time horizons with both methods and objective functions. It can be seen that the prediction horizon of 30 minutes is the best since both configurations with this horizon reach the highest probabilities. Also, the historical value of 60 minutes has the highest probability. Even so, as the confidence interval overlaps with the results obtained with the historical values of 120 minutes, there is no statistical evidence that one method is better than the other.

Conclusions

The conclusions obtained from chapter 3 are that significant differences and dependencies have been observed among glucose profiles classified according to the variables day of the week and time slot, and the groups found have been different for each patient, demonstrating the need for an individualized study.

The main conclusions of the work presented in chapter 4 can be summarized as follows. Glucose predictions with models created with GP are better for shorter time horizons and gradually worsen as the time horizon increases from 30 to 240 minutes. Models created with glucose values classified in categories with fewer elements obtained the best results. In general, when using classified glucose values (*CHAID-GP*), the accuracy of the predictions of glucose values improves compared to those of models with the original dataset (*GP*).

In the research presented in chapter 5, we concluded that GV can be incorporated by generating models for different patterns of glucose profiles or by including LV. Models created with LV improve the quality predictions and do not produce forecasts in the worst zone E and only a few points in the second-worst zone D. *CHAID-GP* and *LV-CHAID-GP* are the best for all time horizons and patients. In general, for all patients and short-term horizons, *LV-GP*, *CHAID-GP*, and *LV-CHAID-GP* have more accurate models than *GP*. The analysis of the relative importance of the variables reveals that MEAN, PSTR, and JI measures are in the top-ranking positions. The non-LV features (glucose, carbohydrates and insulin) appear in the top five positions of influence. The statistical analysis was performance with a novel approach based on a Bayesian model and the Plackett-Luce distribution over rankings. It reveals that *LV-GP* has the highest probability of being the best for 30, 60, 90, and 120 minutes. For 150 and 180 minutes, *CHAID-GP* is the method with the highest probability of being the best.

Results from chapter 6 show that the multi-objective approach produces better models, reducing the number of predictions in the most dangerous zones of the CEG metric for both scenarios. These results are achieved because the multi-objective approach has a better ability than the single-objective approach to traverse the different areas of the search space defined by the F_{CLARKE} objective function. This is an essential conclusion since medical criteria are included in this function, which penalizes the models' most dangerous mispredictions. In addition, we have also found that GE can obtained good results in terms of CEG for both scenarios, despite it not being considered an objective function. However, the multi-objective approach can be regarded as safer since the CEG metric is explicitly included, and a decision-maker could examine non-dominated models and decide which one best fits a patient.

Now, we are developing a framework to generate models created with the multi-objective approach for both the *What-if* and *Agnostic* scenarios. The models created in this thesis could be, after clinical testing, directly applicable to daily clinical practice.

Our future work will explore other combinations of evolutionary computation techniques with fuzzy logic and neural network approaches, and different clustering algorithms will be applied (e.g., K-means, K-shape, Recurrent Neural Network (RNN)).

Resumen

Antecedentes

Diabetes Mellitus (DM) es una enfermedad crónica que aumenta la morbilidad y mortalidad, y provoca un deterioro significativo en la calidad de vida. Existen principalmente dos tipos de diabetes:

- Diabetes Mellitus Tipo 1 (DMT1): debido a un proceso autoinmune, el páncreas no puede generar suficiente insulina para procesar el azúcar producido después de la ingesta de carbohidratos.
- Diabetes Mellitus Tipo 2 (DMT2): la insulina producida por el páncreas no funciona correctamente, en un fenómeno conocido como resistencia a la insulina.

La DMT1 solo se puede tratar con insulina sintética inyectada en el torrente sanguíneo. Dependiendo de la cantidad de insulina en el cuerpo se pueden dar dos escenarios. Por un lado, un exceso de insulina puede provocar hipoglucemia, definida como un valor de Glucosa en Sangre (GS) inferior a 70 mg/dl. Si esta situación continúa en el tiempo, puede causar complicaciones a corto plazo. Por otro lado, si la dosis de insulina es insuficiente, puede provocar hiperglucemia, definida como un valor de GS superior a 180 mg/dl, lo que puede derivar en complicaciones a largo plazo. El objetivo es mantener los niveles de GS dentro del rango objetivo la mayor parte del tiempo (definido como un valor de GS entre [70, 180] mg/dl).

El control de GS en pacientes insulino-dependientes requiere predecir los valores futuros de glucosa para determinar la cantidad de insulina que se necesita inyectar. Esta cantidad depende de muchos factores que las Personas con Diabetes (PcD) tienen que estimar manualmente, siguiendo fundamentalmente dos terapias distintas: Infusión Continua de Insulina Subcutánea y Múltiples Dosis de Insulina. Uno de los factores más importantes es la Variabilidad Glucémica (VG), definida como la fluctuación de los niveles de glucosa durante un día. La VG hace que el proceso de predicción sea más complicado.

Existen diferentes tipos de estrategias de control de GS (*Manual*, *Semiautomática* y *Automática*), pero en todas ellas es importante desarrollar modelos matemáticos o sistemas basados en Inteligencia Artificial que describan la interacción entre el sistema de glucosa y la insulina utilizando las mediciones y los datos almacenados.

Hay diferentes formas de producir modelos de predicción de GS dependiendo de la información disponible en el momento del pronóstico. Una opción es el escenario *Que-pasaría-si*, donde el modelo predice los valores futuros de glucosa teniendo en cuenta no solo los valores pasados de las tres variables (glucosa, carbohidratos e insulina), sino también la ingesta futura de carbohidratos y las inyecciones de insulina presentes hasta el horizonte de predicción. Por ejemplo, el modelo puede predecir el nivel de GS, suponiendo que el paciente ingiere una cierta cantidad de carbohidratos m minutos desde el tiempo de la predicción. Otra opción es el escenario *Agnóstico*, que genera el modelo sin tener en cuenta información sobre eventos futuros en la fase de predicción. Este tipo de modelo necesita predecir implícitamente esos eventos. Por ejemplo, el modelo debe identificar los periodos de ayuno o el ejercicio físico. Ambos escenarios se utilizan en esta tesis.

Existen tres métodos principales comúnmente utilizados para resolver el problema de predicción de glucosa. El primer método está compuesto por el modelo *Fisiológico*, que utiliza ecuaciones lineales que simulan la fisiología del sistema regulador de glucosa-insulina, y requiere conocimientos específicos de fisiología. El segundo método, el modelo *Basado-en-datos*, puede predecir la concentración de glucosa basándose únicamente en los datos de entrada y salida existentes. El último método combina ambas

soluciones de forma *Híbrida*, donde el modelo toma las partes más simples de la fisiología de la glucosa-insulina e incluye datos para determinar los parámetros del modelo. Todos los modelos creados en esta tesis se basan en el método *Basado-en-datos.*

Hay algunas investigaciones que utilizan técnicas de Minería de Datos y Aprendizaje Automático para estudiar el problema de la predicción de glucosa y recomendación de insulina en PcD. Una de esas técnicas son los árboles de decisión. Uno de los más utilizados es el algoritmo CHi-square Automatic Interaction Detection (CHAID). Este algoritmo divide recursivamente los datos mediante una variable respuesta/objetivo utilizando múltiples divisiones entre las diferentes variables de entrada/predictores. El algoritmo calcula la prueba Chi a través del predictor. Una prueba Chi-cuadrado da como resultado una probabilidad entre 0 y 1. Un valor cercano a 0 indica una diferencia significativa entre las dos clases que se comparan, y un valor cercano a 1 significa que no hay diferencia significativa entre ellas. Si la prueba no es estadísticamente significativa, el algoritmo determina el predictor que es menos diferente significativamente con respecto a la variable criterio y junta las categorías del predictor. De lo contrario, calcula el p-valor de Bonferroni para el conjunto de categorías del predictor. La corrección de Bonferroni se usa para comparar múltiples clases y encontrar la clase que obtuvo la mejor división. Si la prueba no es estadísticamente significativa, la variable predictora con el p-valor más bajo se utiliza para la próxima división. De lo contrario, el nodo respectivo se convierte en un nodo terminal. El algoritmo repite este proceso hasta que no se pueden realizar más divisiones.

Otras soluciones trabajan con técnicas evolutivas como Programación Genética (PG) o Gramáticas Evolutivas (GEs). La PG es una técnica de computación evolutiva que resuelve problemas de forma autónoma sin requerir que el usuario conozca la estructura de la solución de antemano. En PG evolucionamos una población de programas de ordenador. El algoritmo crea aleatoriamente una población inicial de programas (individuos o soluciones). A continuación, ejecuta cada programa y determina su aptitud para seleccionar uno o dos programas de la población que pasan a ser los programas padre del proceso. Un programa hijo se crea combinando partes elegidas al azar de los dos programas seleccionados (cruce). Se crea otro nuevo programa hijo alterando aleatoriamente una parte elegida al azar de un programa padre (mutación). El algoritmo repite este proceso hasta que encuentra una solución aceptable o se cumple alguna otra condición de parada. Finalmente, devuelve como solución al mejor individuo que ha encontrado. Las GEs son una variante de la PG, donde la solución se calcula mediante el uso de una gramática específica. La expresión matemática de la solución se logra decodificándola a través de la gramática. Por lo tanto, se puede restringir el espacio de búsqueda y se puede incorporar el conocimiento del dominio del problema. Por lo general, la gramática se implementa en formato Backus-Naur Form (BNF) y se diseña para encontrar un modelo predictivo de niveles de GS.

El método Clarke Error Grid (CEG) se publicó en 1987 como técnica para evaluar la precisión clínica de los Monitores Continuos de Glucosa (MCG). Parkes Error Grid (PEG) se publicó en 2000 como una alternativa a CEG. Ambos métodos se pueden utilizar para representar las diferencias entre los valores estimados en una predicción y los valores actuales o de referencia en un gráfico con coordenadas cartesianas donde el eje X representa los valores de referencia y el eje Y los valores de la predicción, con Y = X siendo la predicción ideal. La característica única de esta representación es que el gráfico se divide en cinco zonas según el grado de precisión de las estimaciones de glucosa. La diferencia entre los dos métodos radica en la definición de las zonas. En CEG las zonas se definen de la siguiente manera: la zona A representa los valores de glucosa que se desvían de la referencia no más del 20%; la zona B son puntos que están fuera del 20% pero que darían lugar a un tratamiento adecuado; la zona C son puntos que darían como resultado una corrección excesiva de los niveles aceptables de GS, lo que conduciría a un tratamiento innecesario; la zona D son puntos que indican un error potencialmente peligroso en la detección de hipoglucemias o hiperglucemias; la zona E son puntos que confundirían el tratamiento entre hipoglucemias e hiperglucemias. En PEG, las zonas se redefinen en base a las zonas de CEG y a los límites establecidos por 100 médicos expertos en diabetes en una encuesta realizada en la *American Diabetes Association Meetings* en junio de 1994. En ambas alternativas, el objetivo es maximizar las predicciones incluidas en las zonas A y B y minimizarlas en las zonas C, D y E.

Objetivos

Los principales objetivos de esta tesis son mejorar los modelos de última generación, en términos de precisión y efectividad, aplicados en la predicción de los niveles de GS en pacientes con DMT1 y aumentar el horizonte temporal de predicción. Para llevar a cabo estos objetivos primarios, se definen los siguientes sub-objetivos:

- Demostrar usando técnicas estadísticas, que es posible extraer una pequeña cantidad de perfiles de glucosa que caractericen la mayoría de los patrones de glucosa en pacientes con DMT1.
- Crear una nueva metodología para obtener predictores de glucosa precisos mediante un proceso compuesto por tres pasos: (i) los datos se recopilan usando MCG, se preprocesan y se dividen usando criterios de división; (ii) los datos se agrupan para obtener modelos personalizados de diferentes perfiles de glucosa; (iii) se utiliza un paso de entrenamiento para crear modelos de predicción ensamblados basados en computación evolutiva.
- Diseñar mediante un enfoque *Basado-en-datos* modelos de predicción utilizando algoritmos evolutivos que tengan en cuenta la clasificación de los perfiles de glucosa y varias medidas de VG latentes como nuevas características de entrada del motor de modelado.
- Investigar los beneficios de aplicar un enfoque multi-objetivo para resolver un problema de Regresión Simbólica (RS) basado en el algoritmo Non-dominated Sorting Genetic Algorithm (NSGA-II) a través de las GEs, probando dos escenarios diferentes: *Que-pasaría-si* y *Agnóstico* (el más común en la práctica clínica diaria).

Metodología

A continuación describimos la metodología general utilizada en esta tesis.

En primer lugar, se recopilan datos de glucosa, insulina y carbohidratos de diferentes pacientes con DMT1. La recopilación se realiza mediante MCG e incluye información sobre Glucosa Intersticial, unidades inyectadas de insulina y cantidades de carbohidratos consumidas.

El siguiente paso es el preprocesado. Los datos de glucosa que no se han podido medir se completan realizando una corrección de los valores utilizando interpolación segmentada de grado 3, donde el número máximo de valores consecutivos que se pueden completar es una hora. En el caso de las dosis de insulina, se utiliza una función basada en el modelo de Berger para simular la dinámica del plasma tras la inyección subcutánea de insulina.

Después de que los datos se seleccionan, organizan y sincronizan mediante marcas de tiempo, se crean modelos basados en algoritmos evolutivos. Abordamos el problema como un problema de RS. Los modelos se generan utilizando series temporales de glucosa, insulina, carbohidratos y un conjunto de características con valores históricos y futuros promediados y agregados en diferentes períodos de tiempo. La validación cruzada de 10 iteraciones se utiliza en la fase de entrenamiento. Los modelos se evalúan utilizando dos métricas diferentes que calculan la precisión de las predicciones para cada horizonte temporal.

Finalmente, para seleccionar la prueba adecuada para evaluar de forma estadística los resultados, se analiza la distribución de los mismos usando la técnica de Estimación de la Densidad del Kernel (EDK) de la distribución de las muestras. Para hacer un estudio estadístico más profundo de los resultados, se utiliza el modelo Bayesiano basado en la distribución de Plackett-Luce sobre rankings para analizar múltiples métodos en múltiples horizontes de tiempo.

A continuación, describimos las variantes de la metodología general en cada uno de los trabajos presentados en esta tesis.

En el trabajo presentado en el capítulo 3, después del paso de preprocesamiento, los datos se dividen según un criterio que crea diferentes perfiles de glucosa. Los criterios de división se basan en el comportamiento de la serie temporal de glucosa en estudio. En esta investigación dividimos las series de tiempo en grupos obtenidos mediante la variable día de la semana en siete categorías (Monday; Tuesday; Wednesday; Thursday; Friday; Saturday; Sunday) y la variable franja horaria en seis categorías dividiendo los valores de glucosa en secciones de cuatro horas cada una (00:00h-04:00h; 04:00h-08:00h; 08:00h-12:00h; 12:00h-16:00h; 16:00h-20:00h; 20:00h-24:00h). Se aplica un algoritmo de clustering para explorar los datos y obtener la información oculta en ellos. En concreto, se utiliza el algoritmo CHAID para dividir recursivamente los datos relacionados con una variable objetivo (glucosa) usando múltiples divisiones entre las diferentes variables de entrada (día de la semana y franja horaria). El objetivo es obtener perfiles de glucosa para entrenar los modelos de forma específica en cada perfil.

En el capítulo 4 generamos modelos basados en PG utilizando dos métodos diferentes. El primero, llamado *CHAID-GP*, genera modelos con PG y valores de glucosa clasificados obtenidos en el trabajo presentado en el capítulo 3. El segundo, llamado *GP*, genera modelos creados con PG y sin valores de glucosa clasificados. El mejor modelo de cada método se selecciona mediante el Criterio de Información de Akaike (CIA). Este método se utiliza para medir la calidad relativa de los modelos en función de la entropía de información. Se selecciona un modelo por cada horizonte temporal e iteración de la validación cruzada. Se obtienen modelos de predicción de valores de glucosa a intervalos de 30 minutos para un máximo de cuatro horas (horizontes temporales a 30, 60, 90, 120, 150, 180, 210 y 240 minutos). El método PEG se utiliza para evaluar la precisión de las predicciones para cada horizonte temporal. La idea es comparar estos dos métodos para averiguar si utilizando los valores de glucosa clasificados, los modelos pueden mejorar la precisión en la predicción de los valores de glucosa para diferentes horizontes temporales.

En el estudio presentado en el capítulo 5, después del paso de preprocesamiento, se genera un conjunto de características adicionales basadas en diferentes medidas de VG. Una vez más, después de que se hayan generado todas las Variables Latentes (VL), los datos se dividen y agrupan de acuerdo con un criterio que crea diferentes perfiles de glucosa. Los criterios de división y el algoritmo de clustering utilizados son los mismos que se aplicaron en el capítulo 3. Se utiliza un algoritmo de aumento de datos que genera de forma sintética series temporales de glucosa para entrenar conjuntos de datos. Con esta técnica se consigue incluir información significativa y mejorar la calidad de los datos. Generamos modelos basados en PG usando dos nuevos métodos. El Primero, llamado *LV-GP*, crea modelos basados en PG utilizando características de Variabilidad de Glucosa Latente (VGL). El segundo método, llamado *LV-CHAID-GP*, genera modelos basados en PG, con características de VGL y valores de glucosa clasificados. De nuevo se utiliza CIA para seleccionar el mejor modelo para cada método y los mismos horizontes de tiempo que en el capítulo 4. Se utiliza un método de ensamblado para crear el predictor. Finalmente, el predictor se selecciona para pronosticar el valor futuro de glucosa. El objetivo principal es investigar si el uso de las características de VGL en combinación con la glucosa clasificada mejora o no la precisión en la predicción.

Finalmente, en la investigación presentada en el capítulo 6, analizamos dos escenarios diferentes: *Que-pasaría-si* y *Agnóstico*. Para ambos escenarios, se utiliza un modelo mono-objetivo (GE) y un enfoque multi-objetivo (MO-GE) basado en el algoritmo NSGA-II, considerando como funciones objetivo el Error Cuadrático Medio (ECM) y una función de aptitud ad-hoc llamada F_{CLARKE}. Esta función ad-hoc se basa en el análisis CEG, que ayuda a mostrar el potencial peligro de predicciones erróneas en PcD. Además de los modelos basados en GEs, también hemos ajustado modelos basados en el algoritmo Auto Regressive Integrated Moving Average (ARIMA) para estimar los valores de glucosa en los cuatro horizontes de predicción (30, 60, 90 y 120 minutos). Para el escenario *Agnóstico*, el conjunto de datos incluye, además de la información del conjunto de datos utilizado en trabajos anteriores y en el escenario *Que-pasaría-si*, datos recopilados de un brazalete de actividad. Particularmente se han tenido en cuenta la respuesta galvánica de la piel, la temperatura de la piel y la aceleración. Comparamos dos opciones teniendo en cuenta la información disponible en el momento de la predicción (60 y 120 minutos).

Resultados

Presentamos en el capítulo 3 los patrones de glucosa obtenidos utilizando técnicas de clustering. Se realiza un estudio retrospectivo de diez pacientes con DMT1. Se encontraron diferencias significativas en los perfiles de glucosa clasificados por las variables día de la semana y franja horaria en cada paciente. Hay algunos pacientes con diferencias entre los perfiles de glucosa obtenidos para cada día. En otros pacientes sucede lo contrario, y los perfiles de glucosa obtenidos son similares.

Los resultados de la predicción de glucosa mediante clustering y PG se muestran en el capítulo 4. En este trabajo usamos el mismo conjunto de datos que en el capítulo 3. Se hace una comparación entre *CHAID-GP* y *GP* estudiando la proporción de predicciones en la zona A+B obtenida con PEG para los diferentes grupos de datos y horizontes temporales. Los resultados muestran que los modelos creados con *CHAID-GP* son más precisos o al menos tienen la misma precisión que los modelos creados con *GP*. Finalmente, realizamos un análisis de las predicciones en la fase de evaluación obtenidas para la zona A+B con PEG para todos los horizontes temporales y pacientes. En general, la precisión de las predicciones es mejor para horizontes de tiempo más cortos y empeora gradualmente a medida que aumenta el horizonte temporal.

La contribución de las características de VGL se expone en el capítulo 5. Se utiliza el mismo conjunto de datos que en los capítulos 3 y 4. Utilizamos dos nuevos métodos llamados *LV-GP* y *LV-CHAID-GP*. Estudiamos la proporción entre los valores obtenidos en la zona A+B para los diferentes clusters de datos y horizontes temporales entre *GP* y *LV-CHAID-GP*, y entre *GP* y *CHAID-GP*, respectivamente. Los resultados muestran que los modelos creados con *CHAID-GP* y *LV-CHAID-GP* son más precisos o al menos tienen la misma precisión que los modelos creados con *GP*. Analizamos las diferentes soluciones obtenidas con PEG para todos los pacientes con todos los métodos y horizontes temporales. Cabe señalar que, como era de esperar, cuando aumenta el horizonte temporal, los puntos están más dispersos, obteniéndose un número mayor de puntos en las zonas no deseadas (C, D y E). También podemos ver que *LV-CHAID-GP* no está dando predicciones en la peor zona E y solo algunos puntos en la zona D. También comparamos las predicciones en la fase de evaluación obtenidas para la zona A+B con PEG para todos los horizontes temporales. Los modelos creados con *LV-GP* y *LV-CHAID-GP* son más precisos. También realizamos un análisis de las 25 variables más significativas, en cuanto a su importancia relativa de aparición en los modelos, de un total de 83. Las medidas MEAN, Percentage Spent in Target Range (PSTR) y J Index (JI) aparecen en las diez primeras posiciones. Para las variables que no son VL, la glucosa, la insulina y los carbohidratos aparecen en las cinco primeras posiciones.

Analizamos las distribuciones de los resultados para todos los métodos y horizontes temporales. Los datos no se distribuyen según una distribución Gaussiana y son multi-modales, por lo que es necesario una prueba no paramétrica. A continuación, seleccionamos la prueba de Nemenyi para realizar una comparación de los resultados por pares. La prueba determina las Diferencias Críticas. Los resultados muestran diferencias significativas para todos los horizontes temporales (excepto para 150 minutos). Finalmente, utilizamos la prueba Bayesiana para todos los métodos y horizontes temporales. Los resultados muestran que *LV-GP* tiene la mayor probabilidad de ser el mejor método para los horizontes de tiempo a 30, 60, 90 y 120 minutos. Para 150 y 180 minutos, *CHAID-GP* es el método con mayor probabilidad de ser el mejor.

En el capítulo 6 presentamos los resultados obtenidos en la predicción de glucosa utilizando las GEs multi-objetivo. Se seleccionaron diez pacientes con DMT1 para el escenario *Qué-pasaría-si*, en base a un control adecuado de los niveles de glucosa. Estudiamos las soluciones obtenidas con los enfoques mono-objetivo (GE) y multi-objetivo (MO-GE), donde cada punto representa una solución referenciada por sus coordenadas (CDE, ECM). Podemos observar que las soluciones obtenidas con MO-GE dominan a todas las soluciones generadas por GE. También estudiamos la distribución de las soluciones a través de los diferentes horizontes con ambos enfoques. Como era de esperar, el error y el número de soluciones en zonas de predicción peligrosas aumentan con el horizonte de predicción. Se realiza un análisis adicional de las soluciones teniendo en cuenta las diferentes iteraciones de la validación cruzada. Cabe señalar que algunas de las soluciones obtenidas con las distintas iteraciones son más difíciles de resolver que otras, y se debe hacer un estudio más profundo de los datos para mejorar el algoritmo. Analizamos las soluciones cuantitativamente comparando las 40 instancias diferentes (diez pacientes por cuatro horizontes de tiempo diferentes) para ambos métodos. Las soluciones obtenidas con el método MO-GE dominan a las soluciones obtenidas con GE. Realizamos un análisis de los resultados agregados para los diferentes métodos: GE, MO-GE y ARIMA. Los resultados muestran que el algoritmo MO-GE obtiene las mejores soluciones, reduciendo las predicciones en las zonas más peligrosas D y E para los horizontes de tiempo a 30 y 60 minutos. ARIMA obtiene los peores resultados para todos los horizontes excepto a 30 minutos, donde obtiene mejores resultados que el enfoque GE.

Los seis pacientes del conjunto de datos utilizados en el escenario *Agnóstico* se seleccionaron del conjunto de datos OhioT1DM para la predicción del nivel de glucosa en sangre: actualización 2020. Examinamos las diferencias entre los enfoques MO-GE y GE teniendo en cuenta la información histórica WS={60, 120} minutos. Hay algunos casos en los que las soluciones obtenidas con GE están dominadas por soluciones generadas con MO-GE. Hay otros casos donde las soluciones obtenidas con MO-GE están dominadas por soluciones generadas con GE. Estudiamos los resultados comparando ambos valores históricos. Encontramos soluciones en el frente de Pareto para los diferentes valores históricos y horizontes temporales, por lo que no existe dominancia entre métodos. También analizamos las soluciones cuantitativamente al comparar las 24 instancias diferentes (seis pacientes por cuatro horizontes de tiempo diferentes) para ambos métodos y los dos valores históricos. Las soluciones obtenidas con MO-GE dominan a las soluciones obtenidas con GE. También examinamos los resultados agregados para ws=120 minutos para los métodos GE, MO-GE y ARIMA. Así como en el escenario *Que-pasaría-si*, el algoritmo MO-GE obtiene los mejores resultados en todos los horizontes excepto a 30 minutos, donde ARIMA es el mejor método.

Para analizar la complejidad de las soluciones en términos del número de parámetros que contienen y de su longitud, representamos el ECM en función del número de parámetros y de la longitud de los modelos encontrados en las soluciones para los métodos GE y MO-GE y para ambos escenarios *Qué-pasaría-si* y *Agnóstico*. Hemos observado que las soluciones obtenidas con GE tienen una mayor longitud, un mayor número de parámetros y un ECM más alto que las soluciones obtenidas con MO-GE en ambos escenarios. Para estudiar cómo los diferentes valores históricos WS={60, 120} minutos contribuyeron a los modelos, se realiza un análisis estadístico para evaluar la importancia de los resultados. Primero creamos gráficas de densidad usando la EDK para la distribución de las muestras. La distribución de datos no es unimodal y es necesaria una prueba no paramétrica. Luego, utilizamos el modelo Bayesiano sobre los valores históricos y los horizontes temporales teniendo en cuenta los distintos métodos y funciones objetivo. Se puede observar que las mejores soluciones se obtienen con el horizonte de predicción a 30 minutos ya que ambas configuraciones alcanzan las mayores probabilidades. Además, el valor histórico a 60 minutos tiene la probabilidad más alta. Aun así, como el intervalo de confianza de las soluciones obtenidas con los diferentes valores históricos se superponen, no existe evidencia estadística de que un método sea mejor que el otro.

Conclusiones

Las conclusiones obtenidas en el capítulo 3 son que se han observado diferencias y dependencias significativas entre los perfiles de glucosa clasificados según las variables día de la semana y franja horaria, y los grupos encontrados han sido diferentes para cada paciente, demostrando la necesidad de un estudio individualizado.

Las principales conclusiones del trabajo presentado en el capítulo 4 se pueden resumir de la siguiente manera. Las predicciones de glucosa con modelos creados con PG son mejores para horizontes de tiempo más cortos y empeoran gradualmente a medida que el horizonte de tiempo aumenta de 30 a 240 minutos. Los modelos creados con valores de glucosa clasificados en categorías con menos elementos obtuvieron los mejores resultados. En general, cuando se utilizan valores de glucosa clasificados (*CHAID-GP*), la precisión en la predicción de los valores de glucosa mejora en comparación con los modelos obtenidos con el conjunto de datos original (*GP*).

En la investigación presentada en el capítulo 5, concluimos que la VG se puede incorporar para generar modelos utilizando diferentes patrones de perfiles de glucosa o incluyendo las VL. Los modelos creados con las VL mejoran la calidad de las predicciones y no producen pronósticos en la peor zona E y solo unos pocos puntos en la segunda peor zona D. *CHAID-GP* y *LV-CHAID- GP* son los mejores métodos para todos los horizontes temporales y pacientes. En general, teniendo en cuenta todos los pacientes y horizontes a corto plazo, *LV-GP*, *CHAID-GP* y *LV-CHAID-GP* obtienen modelos más precisos que *GP*. El análisis de la importancia relativa de las variables revela que las medidas MEAN, PSTR y JI están en las posiciones más altas, siendo por tanto las medidas más relevantes. Las características que no son VL (glucosa, carbohidratos e insulina) aparecen en las cinco primeras posiciones de influencia. El análisis estadístico se realizó utilizando un enfoque novedoso basado en el modelo Bayesiano y la distribución de Plackett-Luce sobre rankings. Los resultados obtenidos muestran que *LV-GP* tiene la mayor probabilidad de ser el mejor método para los horizontes de tiempo a 30, 60, 90 y 120 minutos. Para 150 y 180 minutos, *CHAID-GP* es el método con mayor probabilidad de ser el mejor.

Los resultados del capítulo 6 muestran que el enfoque multi-objetivo genera los mejores modelos, reduciendo la cantidad de predicciones en las zonas más peligrosas de la métrica CEG para ambos escenarios. Estos resultados se logran porque el enfoque multi-objetivo tiene una mejor capacidad que el enfoque mono-objetivo para atravesar las diferentes áreas del espacio de búsqueda definidas por la función objetivo F_{CLARKE}. Esta es una conclusión esencial ya que con esta función se obtienen modelos que incluyen criterios médicos que penalizan los errores de predicción más peligrosos. Además, hemos encontrado que las GEs pueden obtener buenos resultados en términos de CEG para ambos escenarios, a pesar de no ser considerada como función objetivo. Sin embargo, el enfoque multi-objetivo puede verse como un enfoque más seguro ya que la métrica CEG se incluye explícitamente, y el responsable de la toma de decisiones va a poder examinar los modelos no dominados y decidir cuál se adapta mejor a cada paciente.

Ahora, estamos desarrollando un marco de trabajo para generar modelos creados con el enfoque multi-objetivo para los escenarios *Que-pasaría-si* y *Agnóstico*. Los modelos creados en esta tesis van a poder ser, después de las pruebas clínicas, directamente aplicables a la práctica clínica diaria.

Como trabajo futuro exploraremos otras combinaciones de técnicas de computación evolutiva con lógica difusa y enfoques basados en redes neuronales, utilizando diferentes algoritmos de clustering (por ejemplo, K-means, K-shape, y Redes Neuronales Recurrentes).

Contents

List of Figures **23**

List of Tables **29**

List of Acronyms **31**

1 Introduction **35**

2 Glucose Prediction and Insulin Recommendation **39**

2.1 Prediction Scenarios . . . 41

2.2 State of the Art . . . 42

2.2.1 Solutions based on Classical Techniques . . . 43

2.2.2 Solutions based on Evolutionary Techniques: Genetic Programming and Grammatical Evolution . . . 44

2.2.3 Solutions based on Data Mining and Machine Learning Techniques: Decision Trees and Neural Networks . . . 46

2.3 Conclusions and our Contributions . . . 48

3 Glucose Patterns using Clustering Techniques **51**

3.1 Methodology . . . 51

3.1.1 Decision Trees . . . 51

3.1.2 Chi-Square Automatic Interaction Detection . . . 52

3.2 Experimental Results . . . 53

3.2.1 Dataset . . . 53

3.2.2 Configuration . . . 54

3.2.3 Results 55
3.2.4 Discussion 55
3.3 Conclusions 63

4 Glucose Prediction using Clustering and GP 65

4.1 Methodology 65
4.1.1 Data Collection, Preprocessing and Division 65
4.1.2 Data Clustering 66
4.1.3 Models Training 67
4.1.4 Selection of the Best Models 68
4.1.5 Parkes Error Grid 69
4.2 Experimental Results 69
4.2.1 Dataset 69
4.2.2 Configuration 69
4.2.3 Results 70
4.3 Conclusions 73

5 Contribution of Latent Glucose Variability Features 75

5.1 Methodology 75
5.1.1 Data Collection, Preprocessing and Division 76
5.1.2 Data Clustering 77
5.1.3 Models Training 77
5.2 Experimental Results 79
5.2.1 Dataset 79
5.2.2 Configuration 79
5.2.3 Results 80
5.3 Conclusions 87

6 Glucose Prediction using Multi-objective GE 89

6.1 Methodology 89

6.1.1 What-if Scenario or Insulin-Carbohydrate Recommendation 90

6.1.2 Agnostic Scenario or Glucose Prediction in Absence of Event Information 91

6.1.3 Fitness Functions 92

6.1.4 Multi-Objective Grammatical Evolution 94

6.2 Experimental Results 95

6.2.1 Datasets 95

6.2.2 Configuration 97

6.2.3 Results 98

6.3 Conclusions 109

7 Conclusions and Future Works **111**

8 Publications **115**

Appendices **117**

A Additional Material for Chapter 4 **119**

B Additional Material for Chapter 5 **121**

C Additional Material for Chapter 6 **131**

List of Figures

2.1 The graph on the left represents the time series of glucose for six consecutive days from a Type 1 Diabetes Mellitus patient. Each day is plotted with a different color. The graph on the right represents the different ranges of glucose values using a bar chart. Each range is plotted with a different color. 40

2.2 Artificial pancreas. 42

3.1 A decision tree that detects patterns of hypoglycemia [1]. Nodes that correspond to risk situations are drawn in blue. 52

3.2 Chart plots for patients 1, 2, 3, 4, 5 and 6. 56

3.3 Chart plots for patients 7, 8, 9, and 10. 57

4.1 Flow diagram describing the data clustering and models training. 66

4.2 Flow diagram describing the data collection, preprocessing, and division. 66

4.3 The ratio of models with better predictions (higher values in zone A+B) for different data clusters. Labels in the Y-axis identify clusters accordingly to table 3.4, columns day of the week + time slot. Dark-gray segments indicate that the best prediction is made by *CHAID-GP* while light-gray means a better model from *GP*. 71

4.4 The ratio of models with better predictions (higher values in zone A+B) for the different time horizons of data. Dark-gray segments indicate that the best prediction is made by *CHAID-GP* while light-gray means a better model from *GP*. 72

5.1 Flow diagram describing the data collection, preprocessing, and division. 76

5.2 Flow diagram describing the data clustering and models training. For N clusters found in the classification process, N predictors are generated. 78

5.3 The ratio of models with better predictions (higher values in zone A+B) for different data clusters. Labels in the X-axis represent the number of cases. Each case represents a time horizon, and the label Y-axis identifies clusters found with the clustering algorithm according to the criteria day of the week and time slot. Dark-gray segments indicate that the best prediction is made by *LV-CHAID-GP* while light-gray means a better model from *GP*. 81

5.4 Parkes Error Grid from test data for all different models of all patients for fold-0 and time horizon of 30 minutes. The X-axis represents the reference values of glucose, and the Y-axis the values of the prediction. Each zone of the Parkes Error Grid is identified with a different color. 82

5.5 Parkes Error Grid from test data for all different models of all patients for fold-0 and time horizon of 90 minutes. The X-axis represents the reference values of glucose and the Y-axis the values of the prediction. Each zone of the Parkes Error Grid is identified with a different color. 82

5.6 Parkes Error Grid from test data for all different models of all patients for fold-0 and time horizon of 150 minutes. The X-axis represents the reference values of glucose, and the Y-axis the values of the prediction. Each zone of the Parkes Error Grid is identified with a different color. 83

5.7 The relative importance of the first 25 variables that make up the models based on *LV-GP* and *LV-CHAID-GP*. The importance is between 0 and 1, where 0 is a variable that never appears in the model and 1 is a variable that always appears in the model. Latent Variables are represented in green, and historical and future variables in purple. 85

5.8 Density plots of the distribution of the results for all methods and time horizons (except 210 and 240 minutes). Each algorithm is plotted with a different color. The distributions are multi-modal. All the plots were obtained with [2]. 86

5.9 Nemenyi test for all methods and time horizons (except 210 and 240 minutes) using the graphical representation of [3]. Each algorithm is placed on an axis according to its average ranking. Those algorithms that show no significant difference are grouped using a horizontal line. The plot also shows the size of the Critical Difference required for considering two algorithms as significantly different. 86

5.10 Bayesian model of [4] to analyze all methods and time horizons (except 210 and 240 minutes). The X-axis represents the probability of winning, with values between 0 and 1, where 0 is a solution with 0% of the probability of winning and 1 is a solution with 100% of the probability of winning. Y-axis represents each of the algorithms. 87

6.1 Workflow illustrating the generation of models for the glucose prediction problem. 90

6.2 Grammar for Blood Glucose prediction. 94

6.3 Workflow illustrating the prediction problem setup. 95

6.4 Histograms and boxplots describe the glucose level of the patients from the dataset used in the *What-if* scenario. The upper figure shows the percentage of time the patient has a very low glucose level (<54 mg/dl in dark-red), low ([54,70) mg/dl in red), in range ([70,180] mg/dl in green), high ((180,250] mg/dl in yellow), and very high (>250 mg/dl in dark-yellow). The numbers (from top to bottom) represent the percentage of time the patient has a glucose level >250 mg/dl, in range [70,180] mg/dl and <70 mg/dl. The lower figure shows the Inter-Quartile Ranges of glucose, where the mean values are represented with red dots. 96

6.5 Histograms and boxplots describe the glucose level of the patients from the dataset used in the *Agnostic* scenario. The upper figure shows the percentage of time the patient has a very low glucose level (<54 mg/dl in dark-red), low ([54,70) mg/dl in red), in range ([70,180] mg/dl in green), high ((180,250] mg/dl in yellow), and very high (>250 mg/dl in dark-yellow). The numbers (from top to bottom) represent the percentage of time the patient has a glucose level >250 mg/dl, in range [70,180] mg/dl and <70 mg/dl. The lower figure shows the Inter-Quartile Ranges of glucose with mean values as red dots. . . . 97

6.6 Comparison between GE and MO-GE (red and green points) for patients 4 and 7 in the *What-if* scenario for all prediction horizons (30 and 90 min in left column, and 60 and 120 min in right column) and both historical values. 99

6.7 Root Mean Square Error and CDE values for all prediction horizons (30 min in red points, 60 min in green points, 90 min in blue points, and 120 min in purple points) for patients 1, 5, and 10 (rows one, two and three) in the *What-if* scenario for GE and MO-GE (left and the right column) and both historical values. 100

6.8 All folds (each color represents a different fold) for patient 6 in the *What-if* scenario for GE and MO-GE (left and right column) and all prediction horizons (row one for 30 min, row two for 60 min, row three for 90 min and row four for 120 min) and both historical values. 101

6.9 GE vs. MO-GE (red and green dots) for patients 567 and 584 (rows 1,2 and 3,4) in the *Agnostic* scenario for all prediction horizons (30 and 90 min in left column, and 60 and 120 min in right column) and historical values of 60 min. 104

6.10 GE vs. MO-GE (red and green dots) for patients 567 and 584 (rows 1,2 and 3,4) in the *Agnostic* scenario for all prediction horizons (30 and 90 min in left column, and 60 and 120 min in right column) and historical values of 120 min. 105

6.11 Solutions coming from historical values (60 min in dot shape and 120 min in star shape) for patient 552 in the *Agnostic* scenario for all prediction horizons (30 and 90 min in left column, and 60 and 120 min in right column) and both methods GE and MO-GE (red and green dots). 106

6.12 Analysis of the complexity of the solutions based on two criteria: the number of parameters and the length of the solutions. First row shows Root Mean Square Error vs. the number of parameters (left) and Root Mean Square Error vs. the length of the solutions (right) in the *What-if* scenario. The second row shows the results in the *Agnostic* scenario. 107

6.13 Density plots of the Root Mean Square Error distribution for GE and MO-GE result for all the time horizons in the *Agnostic* scenario. The distributions are non-unimodal, and a non-parametric test is recommended. 108

6.14 Bayesian model of [4] to analyse the GE results with Root Mean Square Error as objective function for the *Agnostic* scenario. 108

B.1 The ratio of models with better predictions (higher values in zone A+B) for the different time horizons of data. Dark-gray segments indicate that the best prediction is made by *LV-CHAID-GP* while light-gray means a better model from *GP*. 123

B.2 The ratio of models with better predictions (higher values in zone A+B) for different data clusters. Labels in the X-axis represent the number of cases. Each case represents a time horizon, and the label Y-axis identifies clusters found with the clustering algorithm according to the criteria day of the week and time slot. Dark-gray segments indicate that the best prediction is made by *LV-CHAID-GP* while light-gray means a better model from *CHAID-GP*. . . . 124

B.3 The ratio of models with better predictions (higher values in zone A+B) for the different time horizons of data. Dark-gray segments indicate that the best prediction is made by *LV-CHAID-GP* while light-gray means a better model from *CHAID-GP*. . . . 125

B.4 The relative importance of all variables that make up the models is based on *LV-GP* and *LV-CHAID-GP*. The importance is between 0 and 1, where 0 is a variable that never appears in the model and 1 is a variable that always appears in the model. Latent Variables are represented in green, and historical and future variables in purple. . . . 126

B.5 Parkes Error Grid from test data for all different models of all patients for fold-0 and time horizon of 60 minutes. The X-axis represents the reference values of glucose, and the Y-axis the values of the prediction. Each zone of the Parkes Error Grid is identified with a different color. . . . 127

B.6 Parkes Error Grid from test data for all different models of all patients for fold-0 and time horizon of 120 minutes. The X-axis represents the reference values of glucose, and the Y-axis the values of the prediction. Each zone of the Parkes Error Grid is identified with a different color. . . . 127

B.7 Parkes Error Grid from test data for all different models of all patients for fold-0 and time horizon of 180 minutes. The X-axis represents the reference values of glucose, and the Y-axis the values of the prediction. Each zone of the Parkes Error Grid is identified with a different color. . . . 128

B.8 Parkes Error Grid from test data for all different models of all patients for fold-0 and time horizon of 210 minutes. The X-axis represents the reference values of glucose, and the Y-axis the values of the prediction. Each zone of the Parkes Error Grid is identified with a different color. . . . 128

B.9 Parkes Error Grid from test data for all different models of all patients for fold-0 and time horizon of 240 minutes. The X-axis represents the reference values of glucose, and the Y-axis the values of the prediction. Each zone of the Parkes Error Grid is identified with a different color. . . . 129

C.1 Comparison between GE and MO-GE (red and green points) for patients 1 and 2 in the *What-if* scenario for all prediction horizons (30 and 90 min in left column, and 60 and 120 min in right column) and both historical values. . . . 132

C.2 Comparison between GE and MO-GE (red and green points) for patients 3 and 5 in the *What-if* scenario for all prediction horizons (30 and 90 min in left column, and 60 and 120 min in right column) and both historical values. . . . 133

C.3 Comparison between GE and MO-GE (red and green points) for patients 6 and 8 in the *What-if* scenario for all prediction horizons (30 and 90 min in left column, and 60 and 120 min in right column) and both historical values. . . . 134

C.4 Comparison between GE and MO-GE (red and green points) for patients 9 and 10 in the *What-if* scenario for all prediction horizons (30 and 90 min in left column, and 60 and 120 min in right column) and both historical values. 135

C.5 Root Mean Square Error and CDE values for all prediction horizons (30 min in red points, 60 min in green points, 90 min in blue points and 120 min in purple points) for patients 2, 3, 4, and 6 (rows one, two and three) in the *What-if* scenario for GE and MO-GE (left and right column) and both historical values. 136

C.6 Root Mean Square Error and CDE values for all prediction horizons (30 min in red points, 60 min in green points, 90 min in blue points and 120 min in purple points) for patients 7, 8 and 9 (rows one, two and three) in the *What-if* scenario for GE and MO-GE (left and right column) and both historical values. 137

C.7 All folds (each color represents a different fold) for patient 1 in the *What-if* scenario for GE and MO-GE (left and right column) and all prediction horizons (row one for 30 min, row two for 60 min, row three for 90 min and row four for 120 min) and both historical values. 138

C.8 All folds (each color represents a different fold) for patient 2 in the *What-if* scenario for GE and MO-GE (left and right column) and all prediction horizons (row one for 30 min, row two for 60 min, row three for 90 min and row four for 120 min) and both historical values. 139

C.9 All folds (each color represents a different fold) for patient 3 in the *What-if* scenario for GE and MO-GE (left and right column) and all prediction horizons (row one for 30 min, row two for 60 min, row three for 90 min and row four for 120 min) and both historical values. 140

C.10 All folds (each color represents a different fold) for patient 4 in the *What-if* scenario for GE and MO-GE (left and right column) and all prediction horizons (row one for 30 min, row two for 60 min, row three for 90 min and row four for 120 min) and both historical values. 141

C.11 All folds (each color represents a different fold) for patient 5 in the *What-if* scenario for GE and MO-GE (left and right column) and all prediction horizons (row one for 30 min, row two for 60 min, row three for 90 min and row four for 120 min) and both historical values. 142

C.12 All folds (each color represents a different fold) for patient 7 in the *What-if* scenario for GE and MO-GE (left and right column) and all prediction horizons (row one for 30 min, row two for 60 min, row three for 90 min and row four for 120 min) and both historical values. 143

C.13 All folds (each color represents a different fold) for patient 8 in the *What-if* scenario for GE and MO-GE (left and right column) and all prediction horizons (row one for 30 min, row two for 60 min, row three for 90 min and row four for 120 min) and both historical values. 144

C.14 All folds (each color represents a different fold) for patient 9 in the *What-if* scenario for GE and MO-GE (left and right column) and all prediction horizons (row one for 30 min, row two for 60 min, row three for 90 min and row four for 120 min) and both historical values. 145

C.15 All folds (each color represents a different fold) for patient 10 in the *What-if* scenario for GE and MO-GE (left and right column) and all prediction horizons (row one for 30 min, row two for 60 min, row three for 90 min and row four for 120 min) and both historical values. . . . 146

C.16 GE vs. MO-GE (red and green dots) for patients 540 and 544 (rows 1,2 and 3,4) in the *Agnostic* scenario for all prediction horizons (30 and 90 min in left column, and 60 and 120 min in right column) and historical values of 60 min. . . . 147

C.17 GE vs. MO-GE (red and green dots) for patients 552 and 596 (rows 1,2 and 3,4) in the *Agnostic* scenario for all prediction horizons (30 and 90 min in left column, and 60 and 120 min in right column) and historical values of 60 min. . . . 148

C.18 GE vs. MO-GE (red and green dots) for patients 540 and 544 (rows 1,2 and 3,4) in the *Agnostic* scenario for all prediction horizons (30 and 90 min in left column, and 60 and 120 min in right column) and historical values of 120 min. . . . 149

C.19 GE vs. MO-GE (red and green dots) for patients 552 and 596 (rows 1,2 and 3,4) in the *Agnostic* scenario for all prediction horizons (30 and 90 min in left column, and 60 and 120 min in right column) and historical values of 120 min. . . . 150

C.20 Solutions coming from historical values (60 min in dot shape and 120 min in star shape) for patients 540 and 544 in the *Agnostic* scenario for all prediction horizons (30 and 90 min in left column, and 60 and 120 min in right column) and both methods GE and MO-GE (red and green dots). . . . 151

C.21 Solutions coming from historical values (60 min in dot shape and 120 min in star shape) for patients 567 and 584 in the *Agnostic* scenario for all prediction horizons (30 and 90 min in left column, and 60 and 120 min in right column) and both methods GE and MO-GE (red and green dots). . . . 152

C.22 Solutions coming from historical values (60 min in dot shape and 120 min in star shape) for patient 596 in the *Agnostic* scenario for all prediction horizons (30 and 90 min in left column, and 60 and 120 min in right column) and both methods GE and MO-GE (red and green dots). . . . 153

List of Tables

2.1 Classification of types of insulin and their characteristics. Ultra long-acting insulin does not peak. . . . 39

3.1 Characterization of the patients. The table is divided into three parts. The first part shows the clinical characteristics of the patients. The second part shows the glycemic characterization of patients, where $\overline{G}$ is the average glucose, *Std* the Standard Deviation, and T_G is the percentage of time that glucose values spent in each range (< 70 mg/dl, > 250 mg/dl and [70, 180] mg/dl or time in range). The last part shows the number of days and glucose values available. . . . 54

3.2 Tree depth, number of nodes, and number of end nodes obtained for each patient. . . . 54

3.3 Coding used for the variables day of the week and time slot. . . . 54

3.4 Results obtained with the CHi-square Automatic Interaction Detection algorithm for the groupings obtained, taking into account the different days of the week and the time slots for all patients. $\overline{G}$ is the average glucose, *Std* is the Standard Deviation, and T_G is the percentage of time glucose values spent in each range (< 70 mg/dl, > 250 mg/dl and [70, 180] mg/dl or time in range). The colors in the time zone represent the number of elements that make up a group. The color blue represents clusters of one element, green with two, yellow with three, and red with four. . . . 58

3.5 Results obtained with the CHi-square Automatic Interaction Detection algorithm for the groupings obtained, taking into account the different days of the week and the time slots for all patients. The time slots are included in brackets and bold (upper left of each record) along with the average values of glucose (upper right) in mg/dl for each grouping. . . . 59

3.6 Results obtained with the CHi-square Automatic Interaction Detection algorithm for the groups obtained, taking into account the different days of the week and the time slots for all patients. The time slots are included in brackets and bold (upper left of each record) along with the average values of glucose (upper right) in mg/dl for each grouping. In addition, the average insulin values (lower left) are shown in $U/4h$ and carbohydrates (lower right) in gr. . . . 60

3.7 Results obtained with the CHi-square Automatic Interaction Detection algorithm for the groups obtained, taking into account the different days of the week and the time slots for patients 1, 2, 3, 4, and 5. In addition, the average values of insulin $\overline{I}$ and carbohydrates $\overline{C}$ are shown. . . . 61

3.8 Results obtained with the CHi-square Automatic Interaction Detection algorithm for the groups obtained, taking into account the different days of the week and the time slots for patients 6, 7, 8, 9, and 10. In addition, the average values of insulin $\bar{I}$ and carbohydrates $\bar{C}$ are shown. 62

4.1 Predictions (in percentage) obtained for the zone A+B with Parkes Error Grid for *GP* models, and the average percentage for *CHAID-GP* models. *Green* values indicate better solutions for *CHAID-GP*. The best result is highlighted in bold. 73

5.1 Glucose variability measures used. Measures of glycemic risk can also be considered as measures of glucose variability. 77

5.2 Four different variants of the Genetic Programming approach were applied in this study. The differences among algorithms are the use or not of classified glucose values and the service or not of Latent Variables. 80

5.3 Predictions (in percentage) obtained for the zone A+B with Parkes Error Grid for *GP* and *LV-GP* models, and the average percentage for *CHAID-GP* and *LV-CHAID-GP* models. *Green* values indicate better solutions compared to *GP*. The best result is highlighted in bold. 84

6.1 Configuration of the evolutionary engine. 98

6.2 Percentage of predictions for the *What-if* scenario and each zone of the Clarke Error Grid metric. Aggregated data of all the patients in the four-time horizons. The figures reported in the bold point out the best method for each horizon. 103

6.3 Percentage of predictions for *Agnostic* scenario with ws=120 min and each zone of the Clarke Error Grid metric. Aggregated data of all the patients in the four-time horizons. The figures reported in the bold point out the best method for each horizon. 106

A.1 Predictions (in percentage) obtained for the zone A+B with Parkes Error Grid for Genetic Programming models, and the average percentage for *CHAID-GP* models. *Green* values indicate better solutions for *CHAID-GP*. The best result is highlighted in bold. 120

B.1 Predictions (in percentage) obtained for the zone A+B with Parkes Error Grid for *GP* and *LV-GP* models, and the average percentage for *CHAID-GP* and *LV-CHAID-GP* models. *Green* values indicate better solutions compared to *GP*. The best result is highlighted in bold. 122

List of Acronyms

ABBA Adaptive Basal-Bolus Algorithm. 47

ABSys Adaptive and Bioinspired Systems research group. 37

ADA American Diabetes Association. 39

ADRR Average Daily Risk Range. 41, 49

AI Artificial Intelligence. 36, 37, 41, 63, 74, 111, 112

AIC Akaike Information Criterion. 68, 79

AP Artificial Pancreas. 36, 40, 41, 43, 45, 49

AR Auto Regressive. 43, 48

ARIMA Auto Regressive Integrated Moving Average. 44, 47, 98, 103, 106, 107

AUC Area Under Curve. 49, 76

BG Blood Glucose. 35–37, 39–45, 47–49, 52, 70, 90–97, 109, 111, 113

BGI Blood Glucose Index. 41

Bi-LSTM Bidirectional Long-Short Term Memory. 47

BNF Backus-Naur Form. 94

CC Correlation Coefficient. 47

CD Critical Difference. 85

CDT Customer Decision Tree. 46

CEG Clarke Error Grid. 45, 47, 49, 69, 89, 93, 102, 103, 109, 112

CGM Continuous Glucose Monitoring. 36, 37, 41–48, 53, 65, 66, 69, 75, 76, 90–92, 97

CHAID CHi-square Automatic Interaction Detection. 37, 52, 53, 66, 70, 77, 79

CNN Convolutional Neural Network. 48

COD Coefficient of Determination. 43, 47, 70

CONGA Continuous Overall Net Glycemic Action. 41, 49, 76

CRNN Convolutional Recurrent Neural Network. 48

CSII Continuous Subcutaneous Insulin Infusion. 39, 90

CV Coefficient of Variation. 41, 49

DAB Diabetic Association of Bangladesh. 46

DCCT Diabetes Control Complications Trial. 41

DEA Differential Evolution Algorithm. 44, 45

DM Diabetes Mellitus. 35, 41, 44, 46

DSS Decision Support System. 46

FDA Food and Drug Administration. 43

FIT Index of Fiting. 43, 47

FN False Negative. 46

FP False Positive. 46

FPR False Positive Rate. 46

GA Genetic Algorithm. 44, 94

GE Grammatical Evolution. 36, 37, 43–45, 47, 49, 67, 89, 90, 92, 94, 109, 112

GP Genetic Programming. 37, 43–45, 47, 67, 70, 73, 75–79, 83, 87, 111

GRADE Glycemic Risk Assessment Diabetes Equation. 41

GSR Galvanic Skin Response. 48, 92, 97

GV Glucose Variability. 36, 37, 40, 41, 49, 53, 75, 76, 78, 87, 88, 111

HBGI High Blood Glucose Index. 43, 47

IG Interstitial Glucose. 41, 44, 45, 75, 91

IGC Index of Glycemic Control. 41

IQR Inter-Quartile Range. 41, 96

JI J Index. 41, 49, 83, 88, 111

KDE Kernel Density Estimation. 79, 85, 107

KNN K-Nearest Neighbors. 45

LBGI Low Blood Glucose Index. 43, 47

LGV Latent Glucose Variability. 49, 75, 83, 87, 111, 113

LI Lability Index. 41, 49, 76

LMT Logistic Model Tree. 46

LSTM Long-Short Term Memory. 47, 48

LV Latent Variables. 37, 76–79, 83, 88, 111, 112

LVM Latent Variable Model. 48

MAD Mean Absolute Difference. 41, 44

MAG Mean Absolute Glucose. 41

MAGE Mean Amplitude of Glycemic Excursions. 41, 47, 49, 76

MARD Mean Absolute Relative Difference. 48

MCC Matthews Correlation Coefficient. 48

MDI Multiple Doses of Insulin. 40, 46

ML Machine Learning. 36, 37, 43, 45–47, 49, 63, 112

MODD Mean Of Daily Differences. 41

MPC Model Predictive Control. 43

MSE Mean Square Error. 44, 45, 47, 48

MV M Value. 41, 49, 76

NBT Naive Bayes Tree. 46

NLL Negative Log Likelihood. 47, 48

NME Negative Max Error. 43

NNs Neural Networks. 43, 47

NSGA-II Non-dominated Sorting Genetic Algorithm. 36, 37, 46, 49, 89, 94, 97, 109, 112

PCC Padova Clinical Center. 47

PEG Parkes Error Grid. 69, 70, 73, 79, 80, 83, 87, 111

PID Proportional Integral Derivative. 43

PME Positive Max Error. 43

PSO Particle Swarm Optimization. 94

PSTR Percentage Spent in Target Range. 41, 43, 47, 49, 76, 83, 88, 111

PwD People with Diabetes. 35, 37, 39, 40, 42, 45, 49, 65, 69, 76, 79, 87, 89, 90, 92, 96, 109

REPT Reduce Error Pruning Tree. 46

RF Random Forest. 45, 46

RKHS Reproducing Kernel Hilbert Space. 43

RL Reinforcement Learning. 46, 47

RMSE Root Mean Square Error. 37, 43, 45–48, 89, 92, 93, 98, 102, 107–109

RNN Recurrent Neural Network. 47, 48, 113

ROC Receiver Operating Characteristic. 46

RT Random Tree. 46

SD Standard Deviation. 41, 49, 54, 69, 79, 108

SE Square Error. 68

SGE Structured Grammatical Evolution. 45

SMBG Self Monitoring of Blood Glucose. 47

SOD Second Order Differences. 44

SR Symbolic Regression. 36, 37, 45, 67, 75, 78, 89, 92, 94, 112

SVC Support Vector Classifier. 46

SVM Support Vector Machine. 46

SVR Support Vector Regression. 46–48

T1DM Type 1 Diabetes Mellitus. 35–37, 39–41, 43–48, 52, 53, 69, 79, 95, 96, 111, 113

T2DM Type 2 Diabetes Mellitus. 35

TL Time Lag. 47

TN True Negative. 46

TP True Positive. 46

TPR True Positive Rate. 46

UCI University of California Irvine. 46

WHO World Health Organization. 35

Chapter 1

Introduction

Diabetes Mellitus (DM) is a chronic disease of high prevalence in the world population, which increases the morbidity and mortality of People with Diabetes (PwD), and causes a significant deterioration in their quality of life [5]. According to estimates from the World Health Organization (WHO) [6], 422 million adults worldwide suffered from DM in 2016, compared to 108 million in 1980, and this number is expected to grow to 692 million by the year 2045. The global economic costs of diabetes were estimated to be US$ 727 billion.

The human body cells absorb the glucose present in the bloodstream through a hormone called insulin, and the pancreas produces this hormone. Glucose is a sugar extracted from carbohydrates during the digestion of food. When the level of glucose in the blood is elevated, insulin production is triggered, and, as a result, the sugar enters the cells, and glucose level is reduced. In PwD, this process does not work correctly, which results in patients often having very high Blood Glucose (BG) levels (hyperglycemia). Depending on the cause of this imbalance, two main types of diabetes are defined:

- Type 1 Diabetes Mellitus (T1DM): due to an autoimmune process, the pancreas in patients with T1DM is not able to generate enough insulin to process the sugar produced after the food ingestion. At present, the only treatment is the injection of artificial insulin by the patient, or by an insulin pump, with each meal and sometimes among meals to maintain healthy glucose levels.
- Type 2 Diabetes Mellitus (T2DM): the insulin produced by the pancreas is not working correctly, in a phenomenon known as insulin resistance. In advanced stages of the disease, it appears combined with a deficiency of insulin which requires that patients also need to inject some insulin.

T2DM is the most common type of diabetes, while only 5% of PwD have T1DM. However, all kinds of DM cause complications in many organs of the body and increase the overall risk of premature death. Possible complications include, but are not limited to, heart disease, stroke, kidney failure, leg amputation, loss of vision, neurological damage, or foot ulcers [7, 8]. During pregnancy, if DM is not controlled correctly, it increases the risk of stillbirth and other complications.

Some patients with severe complications have undergone pancreas transplantation, but this option can lead to several new problems. The transplant patient must take immunosuppressive medications to prevent their immune system from attacking their new pancreas, and this treatment can contribute to developing bacterial or viral infections and cancer.

As we have said, T1DM can only be treated with synthetic insulin injected into the bloodstream. However, this is not an easy task. On the one hand, an excessive dose of insulin can lead to hypoglycemia, defined as a BG value of less than 70 mg/dl. If the hypoglycemia is very severe (BG less than 50 mg/dl), it can lead to short-term complications, and in the worst case, to unconsciousness and a diabetic coma. On the other hand, if the dose of insulin is insufficient, it can lead to hyperglycemia, defined as a BG value higher than 180 mg/dl. If the hyperglycemia is very severe (higher than 250 mg/dl), it can lead to long-term complications. The goal is to maintain the BG levels within the target range most of the time (defined as a BG value between [70, 180] mg/dl) [9]. It has been shown that both short-term and long-term complications can emerge when these values are not maintained in a healthy range, or there is high variability.

The main objectives of this thesis are to improve the state-of-the-art models, in terms of accuracy and effectiveness, applied in the prediction of BG levels in T1DM patients and increase the time horizon of reliable prediction. The models could be embedded in an automatic system for solving the problem of insulin recommendation and maintaining the BG levels within the target range most of the time. To carry out these primary objectives, we will take on the following sub-objectives:

- To demonstrate that using statistical techniques, it is possible to extract a small number of glucose profiles that characterize the majority of glucose patterns in patients with T1DM. In this way, we aim to facilitate individualized mathematical modeling for each patient, which could allow clinicians to find significant differences and may eventually lead to more accurate models using Machine Learning (ML) and Artificial Intelligence (AI) techniques.

- To create a new methodology to obtain accurate predictors of glucose using a three-step process. First, data is collected using Continuous Glucose Monitoring (CGM), preprocessed and divided using division criteria based on the behavior of the glucose time series. Second, data is clustered to obtained customized models for different glucose profiles. Third, a training step is used to create ensemble prediction models based on evolutionary computation.

- To design a *Data-driven* modeling approach that generates prediction models based on evolutionary algorithms that take into account both classified glucose profiles and several Glucose Variability (GV) measures (calculated through different measures of glycemic average, GV and glycemic risk) as new input features of the modeling engine. In this way, we intend to demonstrate that the inclusion within the input parameters of GV can lead to better and more accurate predictions.

- To investigate the benefits of applying a multi-objective approach to solve a Symbolic Regression (SR) problem based on Non-dominated Sorting Genetic Algorithm (NSGA-II) via Grammatical Evolution (GE), testing two different scenarios: *What-if* and *Agnostic* (the most common in daily clinical practice). The reason for using these two scenarios is to generate insulin or carbohydrate recommendation models that are useful in practice and allow testing possible modifications of treatments without putting patient's health at risk. We also think that these models could be useful when more smart devices are available and integrated into the patients' daily life, allowing alarm signals to be generated when the patient is in unhealthy situations.

The models created in this thesis would, after clinical testing, be directly applicable to daily clinical practice and could be integrated into mobile and web applications. The recommendations generated could help the user to decide the insulin dose and the actions to take. The joint action of the models and digital technologies will make it possible to automate patient therapies, thus improving quality of life and patient autonomy.

The remainder of this thesis is organized as follows. In chapter 2, we explain the problems of glucose prediction and insulin recommendation. First, we classify the types of insulin and the different alternatives in the insulin injection task. Next, we explain the problem of glucose prediction and control of BG in insulin-dependent patients. Then, we describe the Artificial Pancreas (AP), which is an ideal solution for T1DM patients. Subsequently, we explain the main ways to develop BG prediction models and the main works in the literature that applied different techniques to solve them. In particular, we show works based on traditional, evolutionary, and Data Mining and ML techniques. Finally, we offer the conclusions of these works, and the contributions of this thesis to solve the problem of glucose prediction and insulin recommendation.

In chapter 3, we show that by using decision trees, it is possible to obtain a small number of glucose profiles that allow to identify most of the patterns of glucose behavior in patients with T1DM. A retrospective study of ten patients with T1DM, with data acquired through CGM, is performed. A decision tree-based Data Mining technique called CHi-square Automatic Interaction Detection (CHAID) is used to classify glucose profiles using two decision criteria. On the one hand, the different days of the week (Monday; Tuesday; Wednesday; Thursday; Friday; Saturday; Sunday). On the other hand, the different time slots, dividing the day into six sections of four hours each (00:00h-04:00h; 04:00h-08:00h; 08:00h-12:00h; 12:00h-16:00h; 16:00h-20:00h; 20:00h-24:00h). The grouping is made according to the glucose levels recorded, using the statistically significant differences found.

Continuing the research from chapter 3, a method to obtain accurate predictions of subcutaneous glucose values of PwD is proposed in chapter 4. Statistical techniques are applied to identify glucose patterns and profiles and this knowledge is used to create prediction models using Genetic Programming (GP). The time series of glucose values, measured with CGM, are classified based on previous work to customize models for different profiles. The GP models created from the original dataset are compared to the models made with the classified glucose values.

Chapter 5 is an extension of the research in chapter 4. In this work, a set of GP methods are investigated, and predictions are improved by exploring the utility of different features that identify latent GV. New features, including mean glucose, GV and glycemic risk, are generated as input variables of the GP algorithm to improve the accuracy of the models in the prediction phase. The performance of traditional GP, and models created with classified glucose values, are compared with that of using latent GV features. We experimented with 18 different features and studied the significance of the variables in the models. We performed a Bayesian statistical analysis to study whether the use of GV as Latent Variables (LV) improved the predictions of the models.

In chapter 6 we explored the advantages of applying a multi-objective approach to solve a SR problem using GE. In particular, we continue previous research on finding expressions to model BG levels of PwD. We used a multi-objective GE approach based on the NSGA-II algorithm, considering the Root Mean Square Error (RMSE) and an ad-hoc fitness function as objectives. This ad-hoc function is based on the fact that two predictions with the same absolute error can produce correct or incorrect treatments depending on the range in which the glucose level is placed.

Finally, in chapter 7, we present the conclusions and future work.

This thesis is included in the PhD program Information and Communication Technologies in AI of the Rey Juan Carlos University, in the department of Computer Science, Computer Architecture, Computer Languages and Systems, and Statistics and Operations Research. All the work has been done with the research group Adaptive and Bioinspired Systems research group (ABSys) of the department of Computer Architecture and Automation of the Complutense University of Madrid [1]. One of the main research topics of the group is the development of an embedded system that improves glycemic control in PwD by maintaining glucose levels in an adequate range of values. This thesis is framed within this topic.

[1] Esta tesis está incluida en el programa de Doctorado en Tecnologías de la Información y las Comunicaciones en Inteligencia Artificial de la Universidad Rey Juan Carlos, del departamento de Ciencias de la Computación, Arquitectura de Computadores, Lenguajes y Sistemas Informáticos, y Estadística e Investigación Operativa. Todo el trabajo se ha realizado con el grupo de investigación ABSys del departamento de Arquitectura de Computadores y Automática de la Universidad Complutense de Madrid.

Chapter 2

Glucose Prediction and Insulin Recommendation

PwD must control BG levels throughout their lives, trying to keep them at adequate levels. This control is quite complicated, especially in patients with T1DM. When patients have a meal, they need to decide the insulin units to maintain healthy glucose levels. The control of BG in insulin-dependent patients requires predicting the future glucose values to determine the amount of insulin to inject. This amount depends on many factors, but, above all, the patient should account for four of them: (i) the glucose value at the time of injection; (ii) the estimate of the amount of food ingested, usually measured in carbohydrate rations; (iii) the insulin previously injected; and (iv) the estimate of the ratio of how much is still active in the body. Estimating these factors is a complicated process and has to be done manually and several times every day. From a scientific point of view, this process is complex and is not clearly defined by the variables involved.

Correct techniques in insulin delivery, particularly selecting the right type and amount of insulin, are critical for optimal control of diabetes. Accuracy recommendations should guide to more effective therapies and improve outcomes for patients with diabetes.

Fortunately, recent advances in both devices and algorithms, allow automating some parts of this control process and facilitating the task for PwD.

American Diabetes Association (ADA) classified types of insulin into five groups (rapid-acting, regular/short-acting, intermediate-acting, long-acting, and ultra long-acting insulin) depending on how quickly they work, when they peak, and how long they last [10]. Table 2.1 shows the timing of the action in the body of the different types of insulin.

The patient can perform the task of insulin injection following two different primary therapies. The first one is through a Continuous Subcutaneous Insulin Infusion (CSII) device, also known as an insulin

Table 2.1: Classification of types of insulin and their characteristics. Ultra long-acting insulin does not peak.

Type of insulin	Start	Peak	Last
Rapid-acting	15 minutes	1-2 hours	2-4 hours
Regular/short-acting	30 minutes	2-3 hours	3-6 hours
Intermediate-acting	2-4 hours	4-12 hours	12-18 hours
Long-acting	1-2 hours	6-8 hours	8-24 hours
Ultra long-acting	6 hours		36 hours

pump. This device can be programmable and adjusted to administrate the desired amount of insulin at different instant times. The other option is Multiple Doses of Insulin (MDI), which consists of injecting long-acting insulin once or twice daily as a background dose and rapid-acting insulin injections at each mealtime. In both alternatives, the decisions about the amount of insulin to be injected are challenging and have to consider many factors, including the prediction of BG levels.

One of the factors that has a higher importance in the process of the prediction of BG levels is the GV [11]. GV is the fluctuation of glucose levels during a day [12]. It also appears in non-diabetes people as a consequence of the circadian rhythm of the hormones [13]. GV is not necessarily problematic. However, in PwD, an excessive GV could indicate an inadequate control of the disease and lead to short-term and long-term complications [14]. GV also predicts glucose levels even more complicated because the glucose values become more unpredictable [15].

Figure 2.1 shows the evolution of BG levels in a T1DM patient during six consecutive days. The graph on the left shows the time series of glucose values where a different color represents each day. The graph on the right shows the different ranges of values in a bar chart where each range is represented with a different color. It is observed how glucose levels have a high GV during the six consecutive days. Glucose values are in dangerous zones more than 36% of the time (a long period): 23.04% of the time in hyperglycemia, 7.21% of the time in severe hyperglycemia, 5.44% of the time in hypoglycemia, and 1.05% of the time in severe hypoglycemia. Notice that the BG values are almost five more times in hyperglycemia zones than in hypoglycemia zones. This behavior is usual due to the control of the disease in PwD. The goal is to maximize the time in which BG levels are within the green zone in the bar chart, i.e., the target range (in this case, only 63.25% of the time).

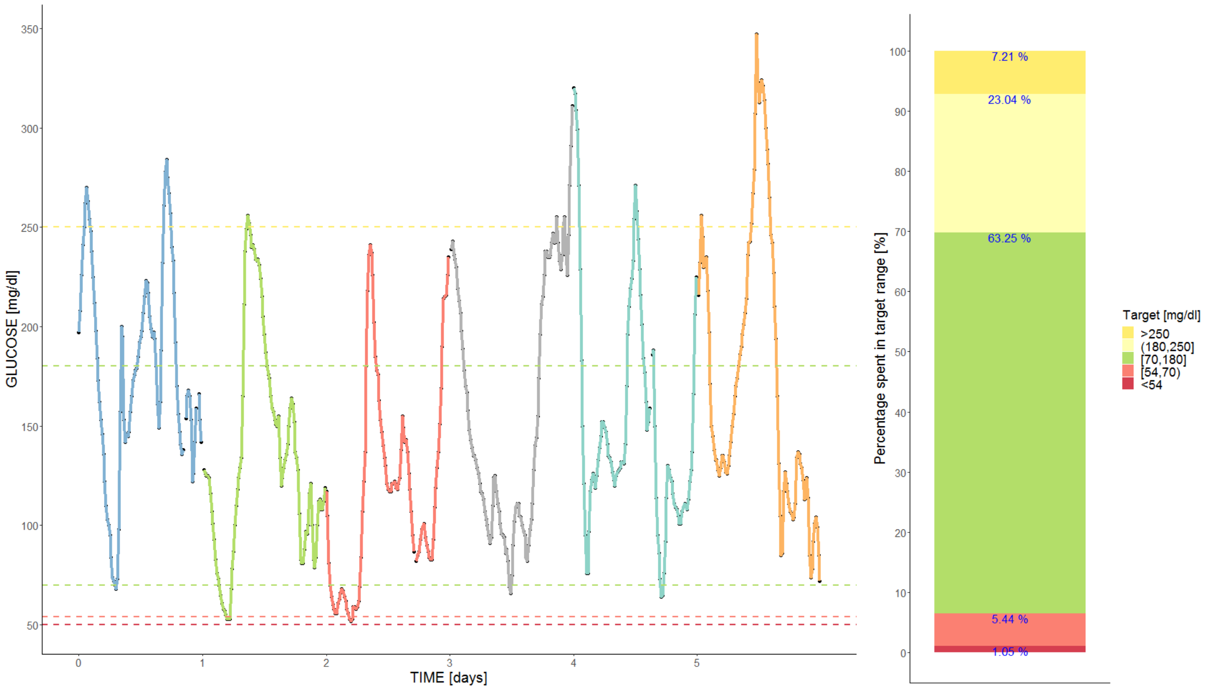


Figure 2.1: The graph on the left represents the time series of glucose for six consecutive days from a Type 1 Diabetes Mellitus patient. Each day is plotted with a different color. The graph on the right represents the different ranges of glucose values using a bar chart. Each range is plotted with a different color.

There are different kinds of BG control strategies [16]: traditional therapies with *Manual* calculation and administration of the insulin protocol [17], *Semi-automated* insulin pump therapies [18], and *Automated* solutions based on the AP [19]. For all of them, it is extremely important to develop mathematical models or AI systems to describe the interaction between the glucose system and the insulin using the measurements and stored data.

An ideal solution for T1DM would be an AP [20] capable of maintaining control of blood sugar levels and allowing the patient to have a healthy life while avoiding, or at least delaying, the appearance of complications. Figure 2.2 shows the basic structure of an AP. One of the components is a CGM consisting of a biosensor [21, 22] (usually a sensor that measures Interstitial Glucose (IG) level [23]) and a wireless transmitter. The other component is an embedded device with various functionalities. On the one hand, it receives the glucose samples and stores them. On the other hand, it controls an insulin pump, which injects insulin boluses into the bloodstream. The device considers several factors to determine the amount of insulin, including the type and characteristics of the insulin.

The continuous recording of IG has revealed the instability of glycemic control in T1DM. The current state of therapies is far from approaching the ideal, resulting in glycemic fluctuations that may harm the pathogenesis of acute and chronic complications in DM. Although the HbA1c level has been used as the reference criterion to evaluate the average glycemic control in patients with T1DM [24], the Diabetes Control Complications Trial (DCCT) [25, 26] confirmed that the HbA1c level is not the most complete of the measures of the degree of glycemia. This measure is insufficient to evaluate the GV where patients with more significant fluctuations have greater hypoglycemia and hyperglycemia [27]. Currently, it is assumed that GV in combination with the HbA1c level may be a more reliable indicator of glycemic control than only the HbA1c level.

In addition to the HbA1c level, there are different measures of glycemic averages such as Arithmetic Mean (MEAN), or Percentage Spent in Target Range (PSTR) [28], among others, and numerous measures that have been developed to evaluate the GV, including Standard Deviation (SD) and its subtypes [29] SD_T (all days and all data points), SD_w (between points within days), SD_{ws} (within series), $SD_{hh:mm}$ (any time of day for all days), $SD_{bhh:mm}$ (between days within time points), etc ..., Coefficient of Variation (CV), Continuous Overall Net Glycemic Action (CONGA) [30], J Index (JI) [28], Lability Index (LI) [31], Mean Amplitude of Glycemic Excursions (MAGE) [12], Mean Of Daily Differences (MODD) [32] , M Value (MV) [33], Average Daily Risk Range (ADRR) [34], Blood Glucose Index (BGI) [34], Inter-Quartile Range (IQR) [29], Glycemic Risk Assessment Diabetes Equation (GRADE) [35], Index of Glycemic Control (IGC) [36], Mean Absolute Difference (MAD) [28], Mean Absolute Glucose (MAG) [28], HYPO score [31], GOLD score [37], Clarke score [37], and Pedersen-Bjergaard score [37].

Due to the number of measures to calculate GV and the high degree of correlation among some of them, it is difficult to determine which one is the most indicated [38]. Some authors distinguish measures of variability between measures of variability in the strictest sense and measures of glycemic control quality [29]. Other authors classify the measures among measures that capture variance, measures of average variability, measures of frequency of crossing of physiologically relevant thresholds, and measures of risk of extreme values, among others [39].

2.1 Prediction Scenarios

There are three different ways of producing models for BG prediction depending on the information available at the time of prognosis: *What-if*, *Agnostic* and *Inertial* scenarios.

The first option is the *What-if* scenario where the model can consider *What-if* events, i.e., future information for some of the variables. For instance, the model can predict the BG level, supposing that the patient eats a certain amount of carbohydrates m minutes from the prediction time. These models are handy for designing insulin or carbohydrates recommendation systems.

Recently, with the appearance of new smart devices, the possibility of incorporating more predictive variables into the models has been unveiled to improve the accuracy of the models. For example, the

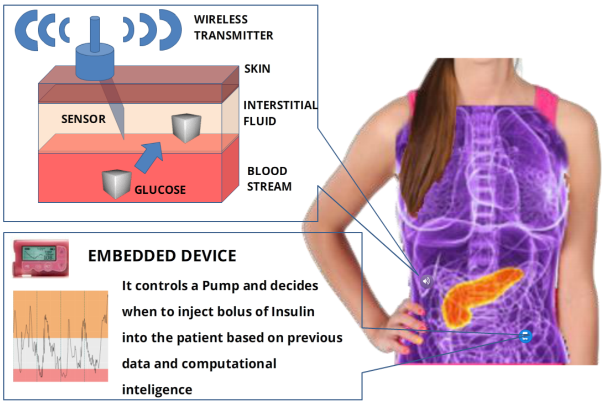

WIRELESS TRANSMITTER
SKIN
SENSOR
INTERSTITIAL FLUID
BLOOD STREAM
GLUCOSE
EMBEDDED DEVICE
It controls a Pump and decides when to inject bolus of Insulin into the patient based on previous data and computational inteligence

Figure 2.2: Artificial pancreas.

activity bracelets on the market provide accurate information on many variables, including, but not limited to, exercise, sleep, heart rate, body temperature, and caloric consumption. When all this information is incorporated into the dataset, it becomes much more complicated to make models of the type *What-if* since the number of possible combinations of the variables and assumptions involved in an event is enormous, and its usefulness would be minimal. However, it is possible to produce models with no information on future events in the prediction phase. This type of model needs to predict, in an implicit way, those events. For example, the model must identify the fasting periods or the physical exercise. This model is framed under the second option called *Agnostic* models, where we cannot access *What-if* events. However, models are trained on all samples, even if they include entries on *What-if* variables. The test phase is performed in the same way for fairness. This approach could theoretically be more difficult since a model trained under *Agnostic* conditions need to predict events in the future time horizon implicitly. *Agnostic* models are more useful when the number of variables, i.e., device measurements, is high.

Finally, under the *Inertial* view, the models are trained using only samples for which the prediction time window only has measurements from the CGM. We cannot use the data where other events occur, which is an unrealistic situation since the selection of the samples needs to be carefully made. Data from other events closely reported could affect a variable, even if they are not precisely located in the dataset. In addition, as [40] pointed out, for large prediction horizons, the datasets are expected to be much smaller and less diverse because that *Inertial* situations are probably going to take place during the night. However, *Inertial* models can be helpful for the study of specific conditions, such as the study of *The Dawn Phenomenon* [41], an early morning BG level rise, which is a particular situation that appears in some PwD.

In this work, we focus only on the two first scenarios, excluding the *Inertial* situation.

2.2 State of the Art

Many works studying the problem of glucose prediction and insulin recommendation in PwD have been presented in the literature. There are three principal methods commonly used [42]. The first one includes mathematical models that simulate the physiology of the glucose-insulin regulatory system. This method uses compartmental models, which are a class of linear dynamic models, named as *Physiological* models. These models are very accurate, but they are typically complex and specific *Physiological* knowledge about the glucose-insulin dynamics is required. The second method is *Data-driven* models, which can predict the glucose concentration based only on existing input-output data. These models allow the incorporation of additional inputs and accurately capture the relationship between the information and the outcome but at the cost of losing part of the *Physiological* meaning of the model. They rely entirely on BG values, carbohydrate intake, and insulin administration data, and possibly other inputs. The last method combines both solutions in a *Hybrid* way, where the models take the simpler parts of the physiology of glucose-insulin and include data to determine the parameters of the models.

In what follows, we review essential literature about those models with an emphasis on *Data-driven* models since these are the models used in our work. In section 2.2.1, we briefly introduce solutions based on traditional techniques, particularly linear and Auto Regressive (AR) models. In section 2.2.2, models based on evolutionary techniques are presented, specifically GP and GE methods. In section 2.2.3, some approaches based on Data Mining and ML techniques are discussed, focusing on decision trees and Neural Networks (NNs) methods.

2.2.1 Solutions based on Classical Techniques

Several papers apply traditional modeling techniques resulting in models or profiles defined by linear equations with a limited set of inputs based on Model Predictive Control (MPC). In [43] the authors proposed two approaches for linear glucose-insulin models to a specific patient. The first one is a non-parametric approach used to identify the glucose-insulin dynamics of the patient. The predictor estimation of the model is calculated through an optimization problem by a Reproducing Kernel Hilbert Space (RKHS). The predictor is converted into an individualized linear model. The second approach required identifying the parametric structure of the model, starting from the linearization of the UVA/Padova metabolic model. The UVA/Padova T1DM simulator [44] is a software program approved by the U.S. Food and Drug Administration (FDA) as a substitute for preclinical trials for certain insulin treatments. The model solutions are evaluated following the next criteria: Coefficient of Determination (COD), Index of Fiting (FIT), Positive Max Error (PME), Negative Max Error (NME) and RMSE. Both approaches are tested in 100 virtual patients of the UVA/Padova T1DM simulator for four days. Results show that the non-parametric models provide the best performance and could improve the closed-loop control concerning those created with non-individualized models.

Leading research groups on AP have presented other personalized control approaches that follow the clinical practice. In [45] a new in-silico model is exploited for both design and validation of a linear MPC based on glucose-insulin model. The new approach provided some modifications to simulate the metabolic specifics of T1DM. The linear model is compared with the classical Proportional Integral Derivative (PID) Control. Dataset consisted of 100 virtual subjects with T1DM with data generated during four days. The performance of the models is evaluated based on four indices: Low Blood Glucose Index (LBGI) and High Blood Glucose Index (HBGI) with coefficients modified concerning literature values to suit control performance result better, and minimum and maximum of BG concentration. MPC results in better regulation than PID, limiting the oscillation of glucose levels significantly.

In [46] the authors compared closed-loop control to open-loop therapy in T1DM patients. The control algorithm uses a simulation model of the glucose-insulin system in the postprandial state. The algorithm is tested with 300 virtual subjects from the UVA/Padova T1DM simulator [44]. Experiments are conducted with 20 T1DM using CGM and insulin pumps, and taking into account two scenarios. During the first scenario, open-loop control is used with their insulin pump. In the second scenario, an insulin management system is inserted and used for the closed-loop control. This system does not automatically control the insulin pump, instead, suggested insulin boluses every 15 minutes. A non-parametric Wilcoxon matched-pairs test is applied to compare open-loop vs. closed-loop control through the number of hypoglycemic events ($< 70\ mg/dl$) and PSTR. Results show that compared to open-loop treatment under identical conditions, closed-loop control improves the overnight regulation of diabetes.

Several works use AR techniques for modeling glucose time series. In [9] the authors applied two prediction strategies based on the description of past glucose data. One is the first-order polynomial, and the other is the first-order AR model. In both methods, parameters are identified using weighted least squares techniques. Then, models are used to forecast glucose levels for two-time horizons: 30 and

45 minutes. The performance of the methods is evaluated based on two classical metrics: Mean Square Error (MSE) and Second Order Differences (SOD). Dataset applied in the study consists of 28 T1DM patients with measurements of glucose values collected with CGM for two days. Results demonstrate that glucose can be accurately predicted within a time horizon of 30 minutes.

Auto Regressive Integrated Moving Average (ARIMA) model to predict future glucose values is used in [47]. ARIMA is a statistical technique that uses variations and regressions of the data, in particular time series, to find the most suitable model parameters to predict future values. It is usually expressed as ARIMA(p, d, q), where the parameters p, d, and q, indicate the order of the different components of the model (p is the autoregressive, d the integrated, and q the moving average component). These parameters are estimated using the Differential Evolution Algorithm (DEA), a method to manage non-differentiable, non-linear, and multi-modal objective functions. The suggested algorithm can trace the *Physiological* changes efficiently. The predictions are calculated for different time horizons: 30, 45, and 60 minutes. A total of 30 different datasets from T1DM patients are simulated using Glucosim [48], a web-based DM educator developed by the Illinois Institute of Technology. MAD between the predicted and actual glucose values is used as the performance metric. Results can be improved using this approach by extending the number of iterations and changing the gain value.

2.2.2 Solutions based on Evolutionary Techniques: Genetic Programming and Grammatical Evolution

A solution that has proven to be suitable for predicting glucose levels is the use of algorithms based on GE [49]. However, one of the main obstacles to training GE models is the lack of essential data. Data collection from real T1DM patients is very complex. GE models trained with a small data set usually suffer from over-fitting. In [50] a method has been proposed to increase the data records of glucose with synthetic data that have good results. Three evolutionary algorithms that generate artificial glucose time series using actual data from a person with diabetes are used to train GE models. Together with the original model, these models are combined into an ensemble to produce a final result. Experimental results show that GE models can get more accurate and robust predictions using data augmentation. The ensemble approach offers excellent performance when compared to not only classical techniques such as ARIMA but also with other GE approaches.

Other comparisons have also been made among techniques related to GE, such as GP with offspring selection. However, the latter has a high execution time. There have been some improvements from the original work [51] to solving this problem. For example, in [52] the offspring selection is conducted by early detection of unsuccessful individuals that estimates if a solution candidate will not be accepted based on partial solution evaluation. For the empirical studies, four artificial benchmark functions to create synthetic data (Friedman-1, Friedman-2, Poly-10, and Pagie-1) and two datasets from actual patients (Tower and Chemical-1) have been used. Results show that the algorithm's efficiency improves as the datasets become larger utilizing this technique, which can be used for Big Data analysis.

The authors in [53] proposed an improved method for predicting the patients' BG trend based on the minimal model used in [54], where the parameters are obtained using a Genetic Algorithm (GA) and identified for every couple (patient, meal) of interest. The study is carried out with data from 15 actual patients for at least 15 days. Glucose levels are registered through a CGM and their estimated intakes of carbohydrates and insulin administration. Optimal control is applied to generate individualized meal-specific insulin profiles. Glucose predictions are made for a multiple-day horizon. Results provide a relationship between variations in insulin sensitivity and variability of BG during the day, as clinical practice suggests.

A work on GP-based induction of a glucose-dynamics model for telemedicine is presented in [55]. The work aims to create a regression model that allows the determination of BG values from IG in five actual patients with T1DM, using it in a telemedicine portal. One of the main problems working with IG values is that they usually differ considerably from BG due to *Physiological* reasons. So, using a large amount of IG measures is remarkably endorse to get as much of the BG-IG dynamics as possible. CGM is used in this work, allowing a large amount of IG values. Another problem is the readily available number of IG values in contrast with the low number of corresponding BG values. To solve this problem, the authors proposed a new approach to enrich the dataset using the Steil-Rebrin model [56] to estimate these missing BG values. The estimation is carried out employing DEA. After this process, the extraction of an explicit SR model using a GP algorithm is applied. The model parameters are selected using RMSE as the fitness function to make the most accurate estimates. The experiments are divided into two scenarios. In the first scenario, the global dataset creates one general model that works for all patients. In the second scenario, the best model for each of the five patients is supposed to find the best model. Results obtained seem that a simple model can fit all of the subjects, and the AP for all of them could be based on this unique model.

In [16] the authors proposed the application of several techniques for modeling and predicting glucose values to obtain customized models of patients. In particular, four methods based on evolutionary algorithms and ML techniques are applied: GE, GP, Random Forest (RF) and K-Nearest Neighbors (KNN). The authors proposed two new enhanced modeling evolutionary algorithms. The first one is a variant of GE which uses an optimal grammar that includes some knowledge about the glucose prediction problem. The second one is a variant of tree-based GP which uses a three-compartment model for carbohydrate and insulin dynamics. Both GP algorithms use strict offspring selection and SR. Both grammars are implemented using compilable phenotypes to speed up the evaluation process and the extraction of models of glycemia. All algorithms use the original datasets plus a set of features that consider historical values of glucose and historical and future values of carbohydrates and insulin, calculated as average values overtime periods to enrich the datasets for the predictions. The predictions are made for four different time horizons: 30, 60, 90, and 120 minutes. The experiments are conducted using ten real PwD with T1DM selected based on conditions of good glucose control with at least ten full days of data not necessarily consecutive neither the same days for each patient. Data is collected using CGM. Results are compared with two baseline predictors: one that considers the average glucose in the past two hours and the other that considers the last known value of the glucose. Clarke Error Grid (CEG) [57] is used to test the accuracy of the glucose predictions. The work concludes that the performance of GE and GP methods are better than or equal to the performance of RF and KNN in terms of CEG. GP and GE variants are performing better than their generic counterparts.

The work has been extended recently in [58] where a novel GE approach called Structured Grammatical Evolution (SGE) is applied. This new variant of GE represents the individuals as chromosomes, where a combination of each non-terminal symbol forms each gene. This way, the locality of the representation is increased, obtaining models more robust than GE. The authors compare this new method with traditional GE. The evolutionary process of both GE and SGE used the RMSE as fitness function. The datasets are the same used in [16], and again solutions are compared in terms of CEG. The authors consider a time horizon of 120 minutes using historical values up to 240 minutes. Experimental results show that SGE achieves better results than GE in terms of both RMSE and CEG.

In [59] the authors presented a *Hybrid* model for predicting glucose in the medium-term (120 minutes) for T1DM patients. A GE-based algorithm is the output of a *Physiological* model in conjunction with a glucose-specific fitness function and a customized grammar. Hence, the final solution can capture the patient's dynamics. The system used 100 virtual patients with synthetic data over 14 days generated by the UVA/Padova T1DM simulator [44]. The datasets incorporate BG readings collected by CGM, the carbohydrates from three meals per day (breakfast, lunch, and dinner), and insulin administration as the sum of basal insulin and the bolus insulin. Both the fitness function of the evolutionary grammar, which is the MSE and the performance metrics, used a penalty factor based on CEG to take into account the physical damage caused by deviations in BG prediction. Four models are tested matching different phases of the day: night, breakfast, lunch, and dinner. Results suggest that by dividing the day into these phases, customized models can be generated, improving the accuracy of the glucose predictions. The data obtained for the night phase are quite good. However, results with data of real patients have not been reported.

The authors in [60] suggest a method based on the Support Vector Regression (SVR) classification model. To find out and tune the SVR parameters optimally, a kind of multi-objective evolutionary algorithm called NSGA-II [61] is used. For multi-objective problems, there is usually no single optimal solution. The classifier is built based on pairwise coupling, a popular multi-class classification method. This work uses four different metrics as fitness functions: recall, precision, accuracy, and F-Measure. Datasets (liver disorder, breast cancer, hepatitis, and Pima Indian diabetes) are obtained from the University of California Irvine (UCI) Machine Learning Repository [62]. The prediction method is compared with other competing methods and Support Vector Machine (SVM) without using regression. Results demonstrate that the proposed method achieves the best accuracy compared to the other methods for all the datasets applied. SVR based on NSGA-II can be very helpful to physicians in the treatment of diabetes and could assist in the diagnosis of other diseases.

2.2.3 Solutions based on Data Mining and Machine Learning Techniques: Decision Trees and Neural Networks

Although more centered in the classification, i.e., prediction of a class instead of a glucose value, there are other interesting works based on Data Mining and ML techniques. They offer technical and methodological solutions to solve problems of medical data analysis and the creation of prediction models. Data Mining finds and discovers unknown patterns or relationships that provide a clear and helpful result [63]. It is a field of science that has developed rapidly in recent years, which helps explain the data and gain knowledge about them [64].

Quite a few works use decision trees in predicting DM. For example, in [65] five traditional ML tree classifiers are analyzed. Particularly, Random Tree (RT) [66], decision tree C4.5 [67], RF [68], Logistic Model Tree (LMT) [69] and Reduce Error Pruning Tree (REPT) [70] are used in this study. The accuracy of the classifications are calculated based on the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) of the classifiers. From the confusion matrix, True Positive Rate (TPR), False Positive Rate (FPR), precision, recall and Receiver Operating Characteristic (ROC) values of the classifiers are calculated. Diabetes dataset from the UCI Machine Learning Repository is used. Results demonstrate that LMT and RF classifiers achieved the best accuracy predicting DM.

Other works applied decision trees to mind significant features of DM. In [71] the authors used four decision tree-based classifiers: Customer Decision Tree (CDT) [72], Naive Bayes Tree (NBT) [73], J48 [74] and REPT. Different metrics evaluate these classifiers: accuracy, Kappa statistics, precision, recall, F-Measure, ROC and RMSE. The study used 220 patient's data with 13 attributes from Diabetic Association of Bangladesh (DAB). Results show that CDT gets the best performance. Several significant features that can help to build a Decision Support System (DSS) for physicians to predict the severity of diabetes are also found. Particularly, three attributes represent the severity of diabetes: plasma glucose, plasma glucose two hours after glucose, and HDL-Cholesterol.

A method for predicting postprandial hypoglycemia using a classification approach with ML techniques personalized to each patient is developed in [75]. The process to generate a hypoglycemic prediction model is made by Support Vector Classifier (SVC) machines for binary classification, trained and tested using Scikit-Learn [76]. SVC selects a small number of critical boundary instances called support vectors from each class and builds a linear discriminant function that separates them as widely as possible to solve a given classification problem. Hypoglycemia risk as a feature and as a class-labeling factor is used in the study. Data from ten patients with T1DM who have used insulin pumps and CGM for several months are collected. The study is conducted under free-living conditions for the patients, and the prediction window of interest is four hours after every meal. Results demonstrate excellent performance in terms of specificity, sensitivity, and accuracy predictions of postprandial hypoglycemic events. The methodology created will support decision systems for patients using MDI or insulin pumps.

Dual mode adaptive basal-bolus advisor based on Reinforcement Learning (RL) is presented in [77]. RL is one of three basic ML paradigms, alongside supervised and unsupervised learning, concerned with how intelligence agents ought to take actions in an environment in order to maximize the notion of cumulative reward. The authors proposed an Adaptive Basal-Bolus Algorithm (ABBA) which provides personalised recommendations for the daily insulin doses using the information of the previous day. ABBA supports inputs from either Self Monitoring of Blood Glucose (SMBG) or CGM. The two versions of ABBA are evaluated using UVA/Padova T1DM simulator [44] with 100 virtual subjects. The following widely used metrics are implemented to evaluate the results: PSTR, percentage time in hypoglycemia ([50, 70) mg/dl) and hyperglycemia ((180, 300] mg/dl), percentage time in severe hypoglycemia (< 50 mg/dl) and hyperglymecia (> 300 mg/dl), LBGI, HBGI, MAGE and the total daily insulin intakes. The proposed RL approach can learn the patient's features and contribute to personalized suggestions on insulin treatment.

Some approaches have solved the prediction problem using artificial NNs with similar results in terms of the quality of the solutions based on GP and GE. NNs are based on a collection of connected units/nodes. Each connection/edge can transmit a signal/actual number, and a non-linear function of the inputs computes the output. Nodes and edges typically have a weight that adjusts as learning proceeds. Commonly, nodes are aggregate in layers. The signal travels from the first layer to the last layer, and the final solution is achieved after multiple layers traversing. There are many kinds of NNs and, due to the availability of new high-performance computing architectures, their development and applicability have increased exponentially during the last five years. Long-Short Term Memory (LSTM) networks have achieved outstanding performance in modeling several time-dependent phenomena. This is the reason why most of the approximations found in the literature for the prediction of BG levels are LSTM since BG data are usually obtained as a time series from CGM.

In [78] the authors make a glucose prediction system using a sequential model consisting of a LSTM, a Bidirectional Long-Short Term Memory (Bi-LSTM), each with four units, and three fully connected layers with 8, 64, and 8 units, respectively. LSTM introduces the memory cell improving prediction by combining their memories, and the inputs [79]. LSTM can learn speedily than others networks and solve complex tasks like time series prediction. In the Bi-LSTM approach, each cell is enabled to access information from both past and future events. Results from LSTM are compared with two baseline methods (ARIMA and SVR) in terms of four criteria: RMSE, Correlation Coefficient (CC), Time Lag (TL) and FIT. Four different prediction horizons are calculated: 15, 30, 45, and 60 minutes. The authors use two different datasets: 11 virtual subjects from the UVA/Padova simulator [44] and 20 real patients with CGM. Results demonstrate that LSTM network outperformed the baseline methods for all prediction horizons and reduced the RMSE and TL.

The authors of [80] developed a predictive model of BG to define future insulin therapy consisting of two stacked LSTM working in parallel. Various prediction horizons from 5 to 60 minutes in steps of five minutes are considering. The predictions of the model are evaluated in terms of COD, FIT and RMSE. The inputs of the model are composed of three variables: the injected insulin recorded by the insulin pump device, the carbohydrate amount inserted manually by the patient, and the glucose levels collected by CGM. One hundred virtual subjects of the UVA/Padova simulator [44] is used for the training step. The models have been tested on both virtual and real data from a patient with T1DM involved in the AP@Home project [81] of the Padova Clinical Center (PCC) with data collected during a month through a fully automatic closed-loop control. Two different scenarios are designed to simulate the timing and meal size variations among subjects. The performance obtained with LSTM has been compared with the average linear model included in the UVA/Padova simulator. Results show superior prediction performance of LSTM against the average linear model in both scenarios.

Another interesting work is [82] where the authors used a Recurrent Neural Network (RNN) composed of LSTM cells. RNN is a class of artificial NNs where connections between nodes form a direct graph along a temporal sequence allowing the expression of temporal dynamic behavior. RNN can use its internal state (memory) to process a variable-length sequence of inputs. The authors used four different fitness functions: RMSE, MSE, Negative Log Likelihood (NLL) loss function based on the Gaussian probability density function, and a *Physiological* loss function based on a smooth version of CEG. Predictions of glucose levels are up to one hour into the future using historical values up to two hours with an LSTM state size of 8, 32, 96, and 128. The training and evaluating steps are carried out with the Ohio T1DM Dataset for Blood Glucose Level Prediction [83] in six patients with T1DM and glucose values collected by CGM. Results indicate that the model performs the same way when trained with NLL and MSE, but with the added benefit of an estimate of the variance of the prediction.

Convolutional Recurrent Neural Network (CRNN) has also been explored to predict the BG level [84]. In Convolutional Neural Network (CNN) the hidden layers include layers that perform convolutions, typically a dot product of the convolution kernel with the layers' input matrix. The architecture proposed in this work is composed of three parts: a multi-layer CNN that gets data features using convolution, a RNN layer with 64 LSTM cells, and a regression output for time series prediction. The LSTM cells stored prior data patterns over arbitrary time intervals. Thus the internal memory can predict the future output according to the previous states. Time horizons of 30 and 60 minutes are calculated. The performance of the proposed method are compared against benchmark algorithms including SVR, Latent Variable Model (LVM) [85] and AR model [86]. Several criteria are used to test the performance of the method: RMSE and Mean Absolute Relative Difference (MARD) to evaluate the accuracy, Matthews Correlation Coefficient (MCC) for detecting either hypoglycemia or hyperglycemia events, and cross-correlation to determine the effective prediction horizon. Data used in this work includes two datasets, in-silico, and in-vivo data. In-silico data consisting of ten adult T1DM subjects with six months of data each case generated using the UVA/Padova simulator [44]. In-vivo data are obtained from a clinical study consisting of multiple phases evaluating the benefits of an advanced insulin bolus calculator for T1DM [87] involving ten patients with data collected from six months. The proposed CRNN method shows superior performance in predicting BG levels in both in-silico and in-vivo datasets. The CRNN is better at capturing the features relative to the baseline algorithms.

2.3 Conclusions and our Contributions

The problem of modeling and predicting glucose levels and glucose-insulin interaction modeling has been an intensive area of research for the last ten years.

The treatment for subjects with T1DM uses rates of basal insulin delivery, insulin to carbohydrate ratios, and individual correction factors, typically from observations by the endocrinologist. However, those models are often inaccurate since clinical data in T1DM are not extensive enough to identify the exact models, and most data does not come from actual data.

Obtaining data from actual patients is complex because this is information that usually requires special permission from both patients and the medical authorities. In addition, patients must be committed to the research for data completion. It is essential to wear the devices continuously, take notes, and register any unexpected event affecting the data. Moreover, it is usual to discard part of the collected data because of mistakes made by patients. Despite these problems, all data used in this thesis have been obtained from real patients.

In most previous works, the main traditional variables that have been considered to create models for the prediction of BG values are carbohydrates intakes, insulin doses, and BG values before the time of prediction. Recently, with the appearance of new smart devices, the possibility of incorporating more predictive variables into the models has been colossal. In this thesis, we use, in addition to the traditional variables, Galvanic Skin Response (GSR), skin temperature and magnitude of acceleration.

Usually, works use historical glucose values to create models to calculate predictions, i.e., future glucose values. Additionally, historical and future values of carbohydrates intakes and insulin doses have been used. But almost none of them compare the effects of using historical values vs. historical and future values. In this thesis, we present a methodology to study the contribution of these variables in two scenarios called *What-if* and *Agnostic*.

GV measures have mainly been used to evaluate the performance of the models. Due to the great importance of these variables in the process of the predictions of BG levels, we incorporate them as new input features of the modeling engine. Particularly we use a set of different measures of glycemic average (arithmetic mean, Area Under Curve (AUC) and PSTR), measures of GV (SD, CV, CONGA, JI, LI, MAGE and MV), and one measure of glycemic risk (ADRR). We call them Latent Glucose Variability (LGV) features since they are obtained from the original glucose values. We also perform a study of the relative importance of these features. This way, we ensure that GV measures appear explicitly in the models and can contribute to improving the predictions.

CEG has been used in several works to test the accuracy of the models in the task of BG prediction. However, the models created do not use the potential that CEG has in predicting glucose values in different degrees of dangerous zones for the patients. To cover this gap, we develop a specific fitness function based on CEG using a GE algorithm combined with NSGA-II. This way, we incorporate knowledge about mispredictions, in terms of CEG, for searching the best solution/s.

There are some models, used in AP systems or closed-loop control models, that try to emulate the action of the pancreas. They are based on the supposition that reasonable glucose control is possible with approximate models. Experimental results suggest that these approaches, due to the lack of accurate, have a significant risk of excessive insulin administration and, therefore, the possibility that BG levels fall into the hypoglycemia zone. Our evolutionary models in this work try to avoid this situation through individualized models for each patient and specific problem.

Several approaches use Data Mining and ML techniques for the prediction of diabetes, but almost none of them identify glucose patterns in their studies. We improve our models by identifying these glucose patterns and incorporating that knowledge into the final BG predictions.

There is still much work towards predicting BG levels exceeding the 60-minute horizon, despite all the works in the literature. The models proposed in this thesis focus on predicting glucose levels for a predicting horizon of up to two hours to support PwD in the daily management of insulin. Two hours is usually needed to decide if the insulin dose after a meal has been accurate and adequate.

Chapter 3

Glucose Patterns using Clustering Techniques

One of the main tasks of this thesis is to identify glucose patterns in glucose levels to create individualized models for each patient and each glucose pattern to make more precise and accurate predictions that will result in a more appropriate insulin recommendation.

The objective of this chapter is to identify, employing the construction of decision trees, glucose profiles classified in groups obtained by the variables day of the week (Monday; Tuesday; Wednesday; Thursday; Friday; Saturday; Sunday) and time slots defined as the division of glucose values into sections of 4-hour each (00:00h-04:00h; 04:00h-08:00h; 08:00h-12:00h; 12:00h-16:00h; 16:00h-20:00h; 20:00h-24:00h).

The rest of the chapter is organized as follows. Section 3.1 describes the decision trees and the technique used for glucose classification. Experimental results, including the data used, experimental work, results, and discussion, are shown in section 3.2. Conclusions are set out in section 3.3.

3.1 Methodology

3.1.1 Decision Trees

Decision trees are statistical classification techniques used in Data Mining [88, 89, 90, 91, 92, 93]. A decision tree is a clear and concise way to examine and decide about possible relationships among the data, identifying groups or segments of interest among them. Nodes that are part of the tree test a particular attribute. Leaf nodes give a classification that applies to all instances that reach the leaf, a set of categories, or a probability distribution over all possible varieties. When a leaf is reached, the instance is classified according to the class assigned to the leaf.

Decision trees are a Data Mining technique that explores data to extract hidden information. The objective of the construction of the decision tree is to create a model to predict the value of a dependent/objective variable from the independent/predictive variables considered. The decision tree has three types of nodes, namely the root node, internal and final nodes, each representing a class characterized by the statistical values of the objective variable and the categories of the predictor variables that each node contains. Every path in the construction of the decision tree is associated with a decision rule established by the algorithm itself. Thus, according to the established rules, the dataset is recursively divided into independent subsets of more minor data (divisive algorithm).

Figure 3.1 shows a decision tree that detects patterns of hypoglycemia in a patient with T1DM [1]. Each day of data is divided into a series of blocks defined by the different meals that the patient has throughout the day. A time window from two hours before to four hours later is defined for each carbohydrate entry, and it only includes automatic measurements of BG values. The color's intensity increases as the node's impurity decrease. It means that the patterns that end in a leaf with an intense color (either blue or orange) are the ones that are capable of separate samples that only belong to one class. Also, each node indicates its impurity value, the number of samples that follow its rule, the number of samples that belong to each class, and the majority class (which determines the node's label). It can be observed that the patient tends to suffer from hypoglycemia at least one time in each block.

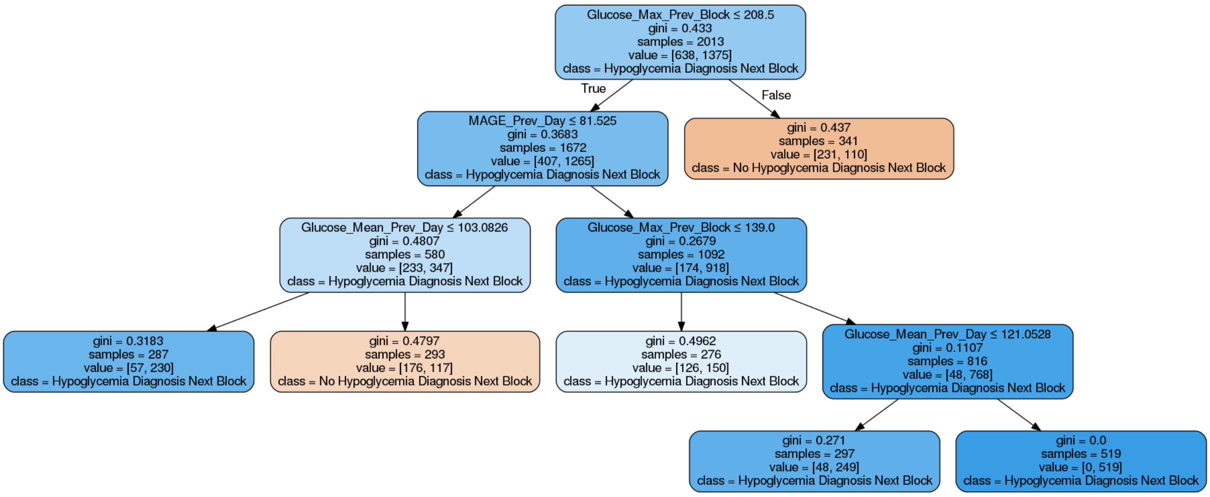

Glucose_Max_Prev_Block ≤ 208.5
gini = 0.433
samples = 2013
value = [638, 1375]
class = Hypoglycemia Diagnosis Next Block
True
False
MAGE_Prev_Day ≤ 81.525
gini = 0.3683
samples = 1672
value = [407, 1265]
class = Hypoglycemia Diagnosis Next Block
gini = 0.437
samples = 341
value = [231, 110]
class = No Hypoglycemia Diagnosis Next Block
Glucose_Mean_Prev_Day ≤ 103.0826
gini = 0.4807
samples = 580
value = [233, 347]
class = Hypoglycemia Diagnosis Next Block
Glucose_Max_Prev_Block ≤ 139.0
gini = 0.2679
samples = 1092
value = [174, 918]
class = Hypoglycemia Diagnosis Next Block
gini = 0.3183
samples = 287
value = [57, 230]
class = Hypoglycemia Diagnosis Next Block
gini = 0.4797
samples = 293
value = [176, 117]
class = No Hypoglycemia Diagnosis Next Block
gini = 0.4962
samples = 276
value = [126, 150]
class = Hypoglycemia Diagnosis Next Block
Glucose_Mean_Prev_Day ≤ 121.0528
gini = 0.1107
samples = 816
value = [48, 768]
class = Hypoglycemia Diagnosis Next Block
gini = 0.271
samples = 297
value = [48, 249]
class = Hypoglycemia Diagnosis Next Block
gini = 0.0
samples = 519
value = [0, 519]
class = Hypoglycemia Diagnosis Next Block

Figure 3.1: A decision tree that detects patterns of hypoglycemia [1]. Nodes that correspond to risk situations are drawn in blue.

3.1.2 Chi-Square Automatic Interaction Detection

One of the most decision tree algorithms used is CHAID [94]. This algorithm recursively divides the data by a response/objective variable using multiple divisions between the different input/predictor variables. A division must reach a threshold level of significance between the nominal values of the objective variable and the branches, or the node is not divided. The search ends when no more branches can be gathered, or there are no significant divisions. The last division is chosen as the solution. The last division does not have to be the most significant examined.

CHAID works with a dependent variable (criterion) and the independent variables (predictors). A Chi-square test yields a probability value as a result lying anywhere between 0 and 1. A Chi-square value closer to 0 indicates a significant difference between the two classes being compared, and a value more relative to 1 means no significant difference between them. Bonferroni correction is used to counteract multiple classes to find the class which did the next best split. In the resulting tree, the most significant independent variable appears in the first node of the classification. The process of node formation ends when there is no significant relationship between the dependent and independent variables. This process is subject to the limitations imposed by the size of the sample. CHAID analysis is restricted by sample size criteria, in particular the sample size required per predictor variable.

The main feature of this method is that no type of distribution of independent variables is assumed a priori because it relies on the use of the Chi-square statistic. There are several advantages of Chi-square. First, it is a non-parametric statistical method of free distribution. Second, the segments can be defined not just by ordinal variables but also by nominal type variables. Any form of variable distribution is accepted in the classification process rather than exclusively a normal one. The range is considerable regarding the kinds of variables that can be included in the tree's construction. Thus, for example, CHAID allows segmentation variables such as age, weight, years of diagnosis of the disease, insulin injection therapy, HbA1c level, and GV measures, among others. Some of these variables are categorical or nominal, and others are ordinal or interval-based. Under such circumstances, a technique that is not subject to the rigidity of the normal distribution and the requirement of ordinal variables will generally be the most appropriate. Hence, Chi-square is the ideal statistical method for these cases. Concerning the dependent variable, CHAID offers, in a natural way, greater flexibility to incorporate continuous criterion variables to the analysis since continuous variables can always be dichotomized.

The basic steps in a CHAID process are shown below:

```
while No further splits can be performed (given the alpha-to-merge and alpha-to-split values) do
    Compute the Chi-test cyclically one by one through all the predictors concerning the criterion
      variable;
    if The test is not statistically significant as defined by an alpha-to-merge value then
        Determine the predictor which is least significantly different concerning the criterion
          variable;
        Merge the respective predictor categories;
    end
    else
        Compute the Bonferroni adjusted p-value for the set of categories of the respective
          predictor;
        if The test is not statistically significant as defined by an alpha-to-split value then
            The predictor variable with the smallest adjusted p-value (the most significant split)
              will be considered for the next split;
        end
        else
            No further splits will be performed, and the respective node will become a terminal
              node;
        end
    end
end
Return to the tree;
```

Algorithm 1: Step-by-step of the CHi-square Automatic Interaction Detection algorithm.

We implement this algorithm with the predictive analysis software IBM SPSS v.21 [95]. The glucose levels are used as an objective variable, and the day of the week and the time slot as predictive variables. A 95% confidence level ($\alpha = 0.05$) is used. The Snedecor F statistic is used as the division criterion and the Bonferroni adjustment for the number of categorical values of the input variable, thus mitigating the bias towards entries with many values.

3.2 Experimental Results

3.2.1 Dataset

A retrospective study of ten patients with T1DM is performed. Measurements were recorded every five minutes using Guardian Real-Time CGM sensors and Minimed insulin pumps from Medtronic. Likewise, carbohydrate estimates made by patients trained in the ration diet process were recorded. The measurements were made on days, not necessarily consecutive nor the same days for each patient. Only those 4-hour sections with at least 46 values are contemplated. Table 3.1 shows the characterization of patients with information on sex, age, weight, glycated hemoglobin HbA1c measured in the three months before the study, and the periods that led to the diagnosis of the disease and treatment with the insulin pump. In addition, for each patient, the average glucose, the SD, and the percentages of time where the patient has glucose levels below 70 mg/dl, above 250 mg/dl, and in time in range [70, 180] mg/dl are shown in the same table.

Table 3.1: Characterization of the patients. The table is divided into three parts. The first part shows the clinical characteristics of the patients. The second part shows the glycemic characterization of patients, where $\overline{G}$ is the average glucose, Std the Standard Deviation, and T_G is the percentage of time that glucose values spent in each range (< 70 mg/dl, > 250 mg/dl and [70, 180] mg/dl or time in range). The last part shows the number of days and glucose values available.

Patient	1	2	3	4	5	6	7	8	9	10
Gender	woman	woman	man	woman	woman	woman	woman	woman	man	woman
Age [$years$]	45	40	70	38	41	31	45	33	34	46
Weight [kg]	74.7	52.0	94.0	55.4	67.3	60.6	50.5	55.5	73.6	64.2
HbA1c [%]	7.2	7.1	7.4	6.9	6.9	8.5	7.6	6.9	7.3	6.9
Diagnosis [$years$]	39	38	40	26	24	18	36	14	14	23
Pump therapy [$years$]	14.0	9.0	15.0	13.0	8.0	12.0	12.0	1.5	1.5	14.0
$\overline{G}$ [mg/dl]	157.67	145.44	143.46	150.47	139.17	142.58	176.33	135.34	146.82	166.13
Std [mg/dl]	62.25	64.02	45.45	56.70	67.93	60.08	68.35	46.11	59.82	86.12
$T\_G_{<70}$ [%]	4.48	8.33	2.14	4.12	14.17	10.08	3.65	4.85	7.54	9.33
$T\_G_{>250}$ [%]	7.31	6.18	2.26	5.14	7.10	4.44	14.28	1.84	5.20	16.16
$T\_G_{[70,180]}$ [%]	37.57	41.83	49.18	41.29	41.89	41.21	27.07	54.06	39.49	35.69
Days	148	181	182	267	190	151	117	82	122	135
Glucose values	37854	49634	53964	65654	43007	40500	27756	18062	28156	32817

3.2.2 Configuration

An individualized study is performed for each patient. The parameters of the decision trees are a maximum tree depth of 3, a minimum number of cases in the parent node of 100, and in the child node of 50. The final depth of the tree, the number of nodes, and the number of end nodes obtained for each patient are shown in table 3.2. The first predictor used in the tree's construction has been the day of the week, and the second predictor has been the time slot. In total, we have seven categories for the day of the week variable and six categories for the time slot variable. The categories of the variables are represented with letters and numbers as shown in table 3.3.

Table 3.2: Tree depth, number of nodes, and number of end nodes obtained for each patient.

Patient	1	2	3	4	5	6	7	8	9	10
Tree depth	2	2	3	2	2	3	2	2	2	2
Number of nodes	33	42	42	36	40	42	32	25	28	46
Number of end nodes	23	30	28	23	27	32	24	17	21	31

Table 3.3: Coding used for the variables day of the week and time slot.

Day of the week	Identifier
Monday	M
Tuesday	T
Wednesday	W
Thursday	R
Friday	F
Saturday	S
Sunday	U

Time slot	Identifier
00:00h-04:00h	0
04:00h-08:00h	1
08:00h-12:00h	2
12:00h-16:00h	3
16:00h-20:00h	4
20:00h-24:00h	5

3.2.3 Results

Table 3.4 represents the groups obtained for each patient. In the first level, the groupings appear per day. For example, we can see that in patient 1, four different glycemic patterns are distinguished; one for Tuesday, another for Thursday, another for Friday, and one last pattern for Monday-Wednesday-Saturday-Sunday, where the lowest average glucose value is for Friday (163.53±58.06 mg/dl) and the highest is for Tuesdays (179.01±55.61 mg/dl). In the second level, the groupings appear by time zone. Colors separate the size of the groups at this level, and it can be seen that in some patients, there are clusters with a large number of bands. In the case of patient 1, we can say that Friday usually has two different behaviors, one for the time slot [20:00h-04:00h] and another for the rest of the day. The same analysis can be performed for the rest of the patients. Table 3.5 includes the glucose average for each time slot. Table 3.6 includes glycemic control information for each day of the week, and tables 3.7 and 3.8 for each time slot.

The relationships found between independent variables and glucose values are also represented as chart plots. The graphics were created using the *circlize* library [96] included in the free statistical analysis software R version 3.5.2 [97]. Figures 3.2 and 3.3 represent the chart plots of the results from table 3.4. A chart plot is created for each patient divided into seven segments, one for each day, and each segment into another six, corresponding to each time slot. Each segment has a letter and a number that identifies the day and the slot (see table 3.3), and the lines represent the relationships found between the days and the slots.

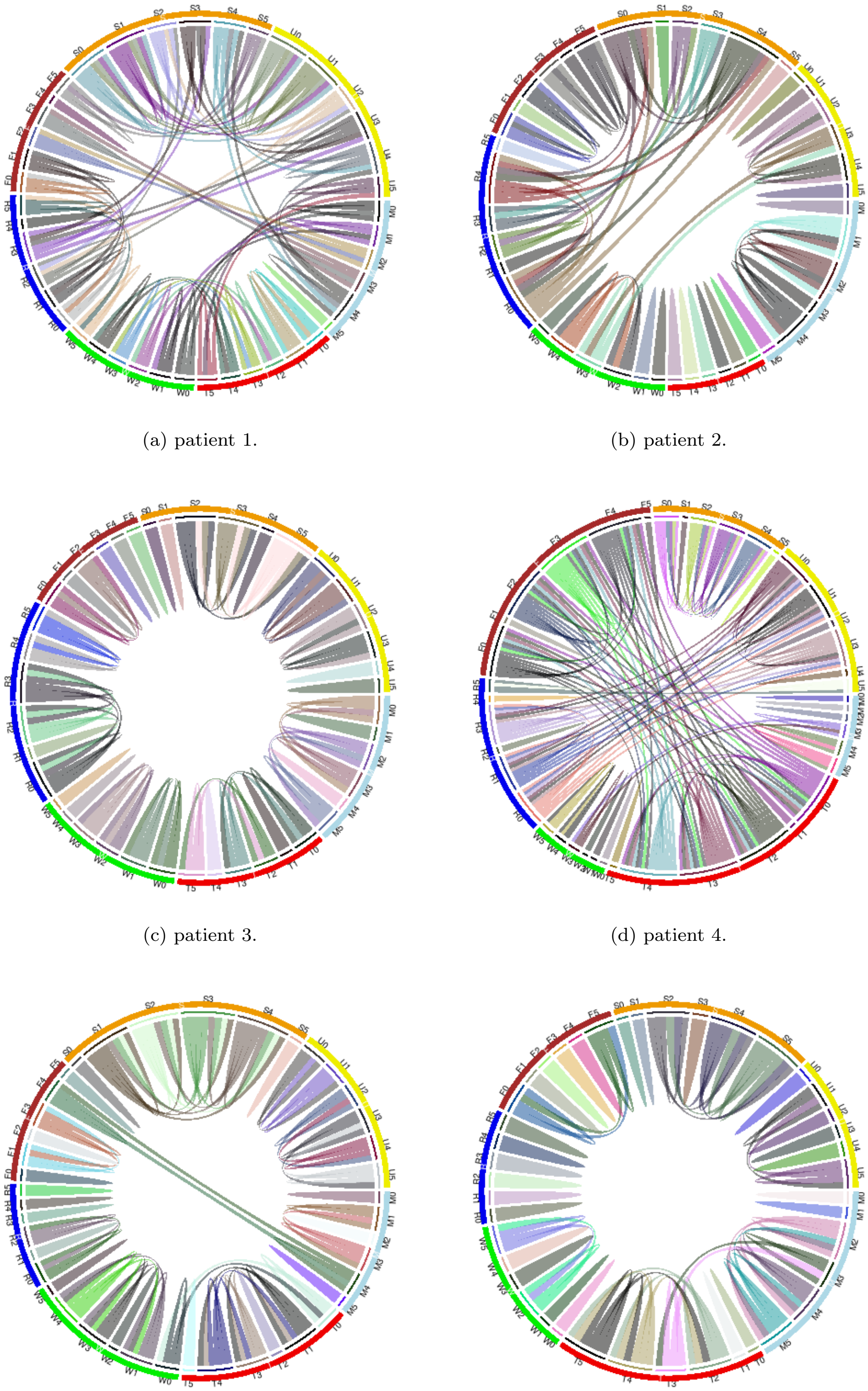
3.2.4 Discussion

Significant differences were found in glucose profiles classified by the variables day of the week and time slot in each patient. The automatic classification has found similarities among glucose profiles of different categories, understanding by category all profiles corresponding to the same time slot of the same day of the week (for example, Mondays from 00:00h to 04:00h). A glucose profile is made up of glucose values measured in that time slot.

The groups obtained with the variable day of the week are heterogeneous. They are made up of one, two, three, and up to four categories. The same applies to the groups concerning the time slot variable. The group size that has been repeated for the day of the week variable is two categories (in all patients) and the least three categories (patients 6 and 7) and four (patients 1 and 9). Regarding the time slot variable, the result is similar. However, a more significant number of similarities have been found among profiles, that is, more clusters than for the day of the week variable and for a more considerable number of the patients analyzed.

From the analysis of tables 3.4 to 3.6, conclusions of clinical applicability can be obtained. In the case of patient 10, it can be observed that on Saturdays, there are more hyperglycemic events than on other days (see table 3.4). With the information in table 3.6 for Saturday, measures can be taken to reduce this behavior, such as modifying its insulin regimen or diet.

If we analyze the chart plots of figures 3.2 and 3.3, it is observed that for patients 3 and 6, the center of the charts is free of lines. This is because the groups in these patients are made up of few elements, with differences among the glucose profiles obtained, both for the days and for the time slots. In patients 4 and 9, the opposite happens. The groupings are made of several elements, and the centers of the graphs are occupied with many lines. In this case, the glucose profiles obtained are similar, with greater glycemic control in these patients (longer in time in range). However, there are graphs (patients 3, 5, and 6) with few connections because the algorithm detects minor differences among glucose values for the different days that were previously grouped in a single set and therefore did not appear.



(a) patient 1.

(b) patient 2.

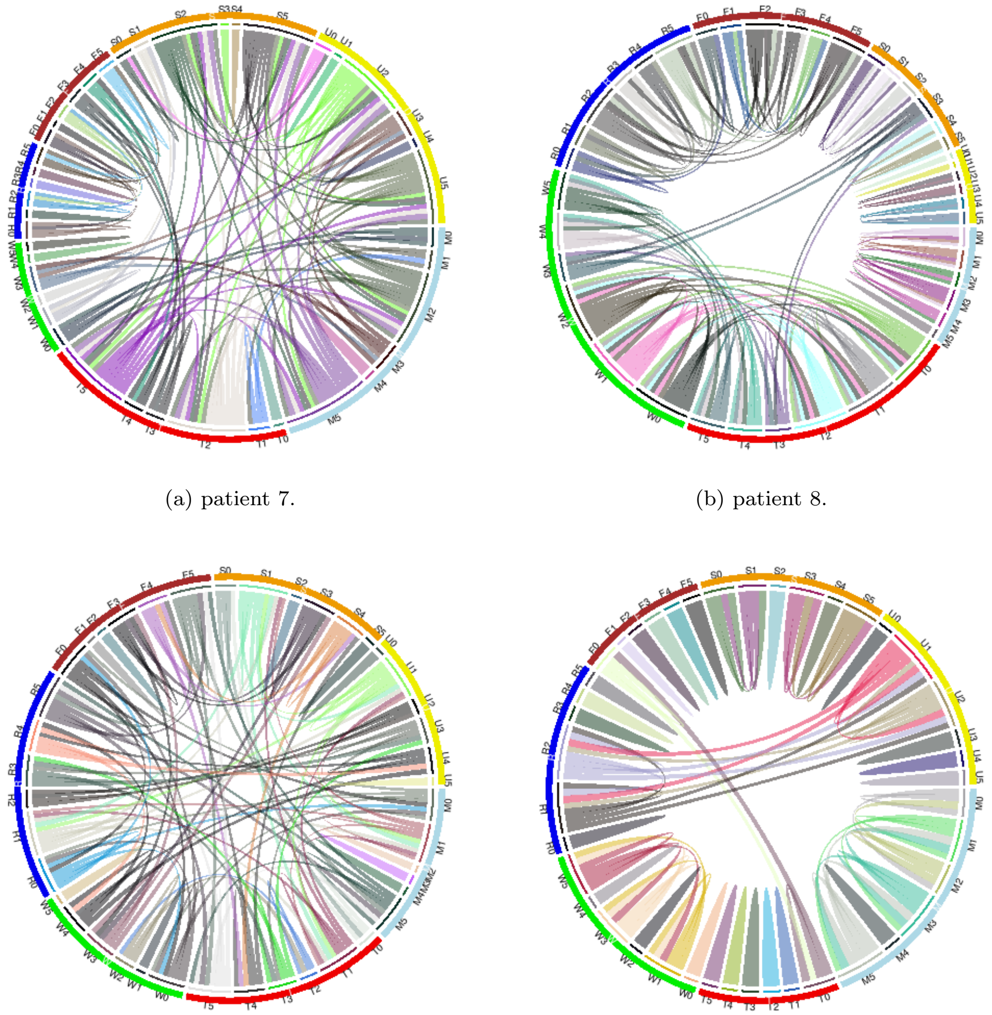
(c) patient 3.

(d) patient 4.

(e) patient 5.

(f) patient 6.

Figure 3.2: Chart plots for patients 1, 2, 3, 4, 5 and 6.



(a) patient 7.

(b) patient 8.

(c) patient 9.

(d) patient 10.

Figure 3.3: Chart plots for patients 7, 8, 9, and 10.

Table 3.4: Results obtained with the CHi-square Automatic Interaction Detection algorithm for the groupings obtained, taking into account the different days of the week and the time slots for all patients. $\overline{G}$ is the average glucose, Std is the Standard Deviation, and T_G is the percentage of time glucose values spent in each range (< 70 mg/dl, > 250 mg/dl and [70, 180] mg/dl or time in range). The colors in the time zone represent the number of elements that make up a group. The color blue represents clusters of one element, green with two, yellow with three, and red with four.

$\overline{G}$ $[mg/dl]$	Std $[mg/dl]$	$T\_G_{<70}$ [%]	$T\_G_{>250}$ [%]	$T\_G_{[70,180]}$ [%]	Day of the week	Time slot					
patient 1											
170.62	61.36	2.22	10.45	57.56	M-W-S-U	0-1		2	3	4	5
179.01	71.15	0.65	13.65	55.61	T	0-1		2-4		3	5
167.22	61.34	1.47	9.81	61.28	R	0	1-2-3			4	5
163.53	58.06	2.40	8.43	62.55	F	0-5		1-2-3-4			
patient 2											
146.28	59.40	6.13	5.08	68.67	M	0	1-2-4			3	5
153.53	58.96	4.42	6.31	67.77	T	0	1	2	3	4	5
158.70	63.25	4.92	9.36	61.93	W-U	0	1	2-4		3	5
151.08	57.70	5.56	5.92	66.63	R-S	0-4		1	2	3	5
148.46	56.29	4.56	5.37	69.56	F	0-2		1	3-5		4
patient 3											
141.65	43.21	2.45	2.02	80.95	M-T	0-3		1	2-5		4
148.25	44.30	2.03	2.56	76.96	W-U	0-1		2-3		4	5
154.49	51.03	0.76	4.92	73.06	R	0-2-3			1	4-5	
139.46	42.90	1.92	0.96	81.83	F	0-2		1	3	4	5
135.85	38.38	1.59	0.53	85.33	S	0	1	2-3-5			4
patient 4											
162.95	58.68	2.96	8.50	61.49	M-W	0	1	2	3	4-5	
151.28	53.76	2.88	4.51	71.18	T-F	0-2-3-4				1	5
157.20	57.62	2.90	7.42	66.88	R-U	0-1-3			2	4	5
154.55	53.01	2.11	5.41	69.49	S	0-2-3-4				1	5
patient 5											
144.50	66.10	10.71	7.19	62.65	M-F	0	1-3		2	4	5
148.19	65.68	8.46	8.11	63.67	T-W	0	1-3-4			2	5
140.77	57.92	8.69	3.91	68.96	R	0-2		1	3	4	5
128.98	55.63	13.65	2.20	67.38	S	0	1-2-3-4				5
132.95	64.48	11.19	6.45	69.40	U	0-1		2-4		3-5	
patient 6											
156.91	61.08	5.81	7.43	61.41	M-T-S	0	1	2-4-5			3
144.54	54.08	8.41	2.76	65.37	W	0	1	2-5		3	4
140.05	55.43	9.73	3.71	67.64	R	0	1	2	3	4	5
152.51	68.46	8.68	7.78	63.12	F	0-5		1	2	3	4
147.10	64.95	10.42	6.67	63.43	U	0	1-5		2	3	4
patient 7											
176.45	70.28	3.28	14.77	55.49	M-S	0	1	2-4		3	5
170.80	64.61	2.06	10.96	61.07	T	0-1-2			3	4	5
179.45	67.80	3.08	15.08	50.18	W	0	1	2-4		3-5	
161.55	60.39	3.40	8.71	62.60	R-F	0-4		1-5		2-3	
173.84	63.19	3.07	11.85	53.03	U	0	1-3-4			2	5
patient 8											
138.60	42.17	2.83	1.03	81.14	M	0-1-3			2-4		5
128.27	44.25	5.40	1.01	82.81	T-W-S	0-1-2			3	4-5	
131.77	48.46	7.15	1.95	77.48	R-F	0	1-4		2-5		3
143.04	45.81	6.04	1.58	74.85	U	0	1	2-3		4-5	
patient 9											
145.90	55.31	7.17	4.83	66.65	M	0-1-2-3				4	5
150.42	61.20	6.23	6.74	66.93	T-W-S-U	0-1		2	3	4-5	
143.12	53.96	7.13	3.09	69.57	R	0-1-4-5				2	3
155.23	62.26	3.25	7.84	66.36	F	0-3-5			1	2	4
patient 10											
177.30	76.22	4.58	14.29	48.78	M	0-2		1-3-5			4
168.56	78.29	7.84	15.96	52.72	T-F	0	1	2	3	4	5
163.74	76.13	10.45	14.86	51.35	W	0	1-3-5			2	4
181.96	85.28	6.77	21.44	46.92	R-U	0	1-2		3	4	5
190.48	90.48	4.89	23.50	45.24	S	0-1		2	3-5		4

Table 3.5: Results obtained with the CHi-square Automatic Interaction Detection algorithm for the groupings obtained, taking into account the different days of the week and the time slots for all patients. The time slots are included in brackets and bold (upper left of each record) along with the average values of glucose (upper right) in mg/dl for each grouping.

$\overline{G}$ [mg/dl]	*Std* [mg/dl]	$T\_G_{<70}$ [%]	$T\_G_{>250}$ [%]	$T\_G_{[70,180]}$ [%]	Day of the week	Time slot					
patient 1											
170.62	61.36	2.22	10.45	57.56	M-W-S-U	**(0-1)** 172		**(2)** 159	**(3)** 164	**(4)** 189	**(5)** 168
179.01	71.15	0.65	13.65	55.61	T	**(0-1)** 198		**(2-4)** 170		**(3)** 158	**(5)** 177
167.22	61.34	1.47	9.81	61.28	R	**(0)** 170	**(1-2-3)** 163			**(4)** 186	**(5)** 156
163.53	58.06	2.40	8.43	62.55	F	**(0-5)** 170		**(1-2-3-4)** 160			
patient 2											
146.28	59.40	6.13	5.08	68.67	M	**(0)** 137	**(1-2-4)** 156			**(3)** 146	**(5)** 127
153.53	58.96	4.42	6.31	67.77	T	**(0)** 170	**(1)** 179	**(2)** 157	**(3)** 139	**(4)** 148	**(5)** 132
158.70	63.25	4.92	9.36	61.93	W-U	**(0)** 169	**(1)** 178	**(2-4)** 163		**(3)** 144	**(5)** 136
151.08	57.70	5.56	5.92	66.63	R-S	**(0-4)** 157		**(1)** 163	**(2)** 152	**(3)** 145	**(5)** 132
148.46	56.29	4.56	5.37	69.56	F	**(0-2)** 148		**(1)** 163	**(3-5)** 139		**(4)** 154
patient 3											
141.65	43.21	2.45	2.02	80.95	M-T	**(0-3)** 142		**(1)** 140	**(2-5)** 147		**(4)** 132
148.25	44.30	2.03	2.56	76.96	W-U	**(0-1)** 149		**(2-3)** 146		**(4)** 136	**(5)** 164
154.49	51.03	0.76	4.92	73.06	R	**(0-2-3)** 160			**(1)** 136	**(4-5)** 155	
139.46	42.90	1.92	0.96	81.83	F	**(0-2)** 147		**(1)** 121	**(3)** 152	**(4)** 132	**(5)** 139
135.85	38.38	1.59	0.53	85.33	S	**(0)** 129	**(1)** 133	**(2-3-5)** 143			**(4)** 122
patient 4											
162.95	58.68	2.96	8.50	61.49	M-W	**(0)** 167	**(1)** 180	**(2)** 171	**(3)** 149	**(4-5)** 156	
151.28	53.76	2.88	4.51	71.18	T-F	**(0-2-3-4)** 150				**(1)** 146	**(5)** 162
157.20	57.62	2.90	7.42	66.88	R-U	**(0-1-3)** 161			**(2)** 168	**(4)** 141	**(5)** 153
154.55	53.01	2.11	5.41	69.49	S	**(0-2-3-4)** 155				**(1)** 146	**(5)** 161
patient 5											
144.50	66.10	10.71	7.19	62.65	M-F	**(0)** 126	**(1-3)** 145		**(2)** 136	**(4)** 149	**(5)** 165
148.19	65.68	8.46	8.11	63.67	T-W	**(0)** 152	**(1-3-4)** 148			**(2)** 128	**(5)** 163
140.77	57.92	8.69	3.91	68.96	R	**(0-2)** 127		**(1)** 137	**(3)** 142	**(4)** 149	**(5)** 161
128.98	55.63	13.65	2.20	67.38	S	**(0)** 119	**(1-2-3-4)** 126				**(5)** 150
132.95	64.48	11.19	6.45	69.40	U	**(0-1)** 147		**(2-4)** 119		**(3-5)** 132	
patient 6											
156.91	61.08	5.81	7.43	61.41	M-T-S	**(0)** 135	**(1)** 167	**(2-4-5)** 161			**(3)** 155
144.54	54.08	8.41	2.76	65.37	W	**(0)** 110	**(1)** 124	**(2-5)** 156		**(3)** 145	**(4)** 168
140.05	55.43	9.73	3.71	67.64	R	**(0)** 110	**(1)** 123	**(2)** 167	**(3)** 139	**(4)** 151	**(5)** 145
152.51	68.46	8.68	7.78	63.12	F	**(0-5)** 122		**(1)** 151	**(2)** 174	**(3)** 182	**(4)** 161
147.10	64.95	10.42	6.67	63.43	U	**(0)** 127	**(1-5)** 144		**(2)** 154	**(3)** 139	**(4)** 176
patient 7											
176.45	70.28	3.28	14.77	55.49	M-S	**(0)** 204	**(1)** 192	**(2-4)** 167		**(3)** 160	**(5)** 174
170.80	64.61	2.06	10.96	61.07	T	**(0-1-2)** 184			**(3)** 156	**(4)** 167	**(5)** 148
179.45	67.80	3.08	15.08	50.18	W	**(0)** 211	**(1)** 197	**(2-4)** 160		**(3-5)** 171	
161.55	60.39	3.40	8.71	62.60	R-F	**(0-4)** 175		**(1-5)** 150		**(2-3)** 160	
173.84	63.19	3.07	11.85	53.03	U	**(0)** 190	**(1-3-4)** 171			**(2)** 157	**(5)** 181
patient 8											
138.60	42.17	2.83	1.03	81.14	M	**(0-1-3)** 128			**(2-4)** 151		**(5)** 142
128.27	44.25	5.40	1.01	82.81	T-W-S	**(0-1-2)** 129			**(3)** 116	**(4-5)** 134	
131.77	48.46	7.15	1.95	77.48	R-F	**(0)** 141	**(1-4)** 131		**(2-5)** 135		**(3)** 118
143.04	45.81	6.04	1.58	74.85	U	**(0)** 154	**(1)** 166	**(2-3)** 120		**(4-5)** 146	
patient 9											
145.90	55.31	7.17	4.83	66.65	M	**(0-1-2-3)** 151				**(4)** 132	**(5)** 143
150.42	61.20	6.23	6.74	66.93	T-W-S-U	**(0-1)** 158		**(2)** 136	**(3)** 142	**(4-5)** 154	
143.12	53.96	7.13	3.09	69.57	R	**(0-1-4-5)** 148				**(2)** 131	**(3)** 136
155.23	62.26	3.25	7.84	66.36	F	**(0-3-5)** 148			**(1)** 188	**(2)** 137	**(4)** 165
patient 10											
177.30	76.22	4.58	14.29	48.78	M	**(0-2)** 159		**(1-3-5)** 179			**(4)** 202
168.56	78.29	7.84	15.96	52.72	T-F	**(0)** 143	**(1)** 164	**(2)** 149	**(3)** 157	**(4)** 194	**(5)** 203
163.74	76.13	10.45	14.86	51.35	W	**(0)** 155	**(1-3-5)** 166			**(2)** 143	**(4)** 190
181.96	85.28	6.77	21.44	46.92	R-U	**(0)** 174	**(1-2)** 161		**(3)** 166	**(4)** 227	**(5)** 194
190.48	90.48	4.89	23.50	45.24	S	**(0-1)** 209		**(2)** 136	**(3-5)** 173		**(4)** 247

Table 3.6: Results obtained with the CHi-square Automatic Interaction Detection algorithm for the groups obtained, taking into account the different days of the week and the time slots for all patients. The time slots are included in brackets and bold (upper left of each record) along with the average values of glucose (upper right) in mg/dl for each grouping. In addition, the average insulin values (lower left) are shown in $U/4h$ and carbohydrates (lower right) in gr.

$\overline{G}$ [mg/dl]	Std [mg/dl]	$T\_G_{<70}$ [%]	$T\_G_{>250}$ [%]	$T\_G_{[70,180]}$ [%]	Day of the week	Time slot					
patient 1											
170.62	61.36	2.22	10.45	57.56	M-W-S-U	**(0-1)** 172 4 3		**(2)** 159 6 25	**(3)** 164 7 136	**(4)** 189 5 12	**(5)** 168 7 92
179.01	71.15	0.65	13.65	55.61	T	**(0-1)** 198 4 3		**(2-4)** 170 6 18		**(3)** 158 8 131	**(5)** 177 7 87
167.22	61.34	1.47	9.81	61.28	R	**(0)** 170 4 0	**(1-2-3)** 163 5 56			**(4)** 186 5 13	**(5)** 156 7 102
163.53	58.06	2.40	8.43	62.55	F	**(0-5)** 170 5 40		**(1-2-3-4)** 160 5 40			
patient 2											
146.28	59.40	6.13	5.08	68.67	M	**(0)** 137 3 1	**(1-2-4)** 156 4 94			**(3)** 146 3 154	**(5)** 127 4 174
153.53	58.96	4.42	6.31	67.77	T	**(0)** 170 3 0	**(1)** 179 4 88	**(2)** 157 4 108	**(3)** 139 3 162	**(4)** 148 4 92	**(5)** 132 4 156
158.70	63.25	4.92	9.36	61.93	W-U	**(0)** 169 3 4	**(1)** 178 4 83	**(2-4)** 163 4 103		**(3)** 144 3 128	**(5)** 136 4 152
151.08	57.70	5.56	5.92	66.63	R-S	**(0-4)** 157 3 45		**(1)** 163 4 68	**(2)** 152 4 114	**(3)** 145 3 131	**(5)** 132 4 141
148.46	56.29	4.56	5.37	69.56	F	**(0-2)** 148 3 46		**(1)** 163 4 68	**(3-5)** 139 3 139		**(4)** 154 3 88
patient 3											
141.65	43.21	2.45	2.02	80.95	M-T	**(0-3)** 142 12 61		**(1)** 140 8 5	**(2-5)** 147 15 115		**(4)** 132 8 2
148.25	44.30	2.03	2.56	76.96	W-U	**(0-1)** 149 8 2		**(2-3)** 146 16 119		**(4)** 136 7 1	**(5)** 164 15 133
154.49	51.03	0.76	4.92	73.06	R	**(0-2-3)** 160 14 77			**(1)** 136 7 0	**(4-5)** 155 11 64	
139.46	42.90	1.92	0.96	81.83	F	**(0-2)** 147 11 43		**(1)** 121 8 10	**(3)** 152 18 145	**(4)** 132 8 2	**(5)** 139 16 143
135.85	38.38	1.59	0.53	85.33	S	**(0)** 129 7 0	**(1)** 133 7 0	**(2-3-5)** 143 16 120			**(4)** 122 7 0
patient 4											
162.95	58.68	2.96	8.50	61.49	M-W	**(0)** 167 5 0	**(1)** 180 10 115	**(2)** 171 6 58	**(3)** 149 11 184	**(4-5)** 156 8 77	
151.28	53.76	2.88	4.51	71.18	T-F	**(0-2-3-4)** 150 6 68				**(1)** 146 9 131	**(5)** 162 10 103
157.20	57.62	2.90	7.42	66.88	R-U	**(0-1-3)** 161 9 98			**(2)** 168 6 44	**(4)** 141 4 11	**(5)** 153 11 115
154.55	53.01	2.11	5.41	69.49	S	**(0-2-3-4)** 155 6 58				**(1)** 146 9 118	**(5)** 161 11 108
patient 5											
144.50	66.10	10.71	7.19	62.65	M-F	**(0)** 126 1 0	**(1-3)** 145 4 158		**(2)** 136 5 75	**(4)** 149 2 13	**(5)** 165 5 190
148.19	65.68	8.46	8.11	63.67	T-W	**(0)** 152 1 0	**(1-3-4)** 148 4 86			**(2)** 128 4 65	**(5)** 163 4 214
140.77	57.92	8.69	3.91	68.96	R	**(0-2)** 127 3 31		**(1)** 137 5 37	**(3)** 142 4 213	**(4)** 149 2 27	**(5)** 161 5 247
128.98	55.63	13.65	2.20	67.38	S	**(0)** 119 1 0	**(1-2-3-4)** 126 4 105				**(5)** 150 4 225
132.95	64.48	11.19	6.45	69.40	U	**(0-1)** 147 3 17		**(2-4)** 119 3 49		**(3-5)** 132 4 254	
patient 6											
156.91	61.08	5.81	7.43	61.41	M-T-S	**(0)** 135 4 1	**(1)** 167 6 15	**(2-4-5)** 161 7 22			**(3)** 155 7 28
144.54	54.08	8.41	2.76	65.37	W	**(0)** 110 3 0	**(1)** 124 7 29	**(2-5)** 156 6 19		**(3)** 145 6 20	**(4)** 168 6 9
140.05	55.43	9.73	3.71	67.64	R	**(0)** 110 3 0	**(1)** 123 8 36	**(2)** 167 8 24	**(3)** 139 6 25	**(4)** 151 9 24	**(5)** 145 8 26
152.51	68.46	8.68	7.78	63.12	F	**(0-5)** 122 6 16		**(1)** 151 9 37	**(2)** 174 6 18	**(3)** 182 7 38	**(4)** 161 7 16
147.10	64.95	10.42	6.67	63.43	U	**(0)** 127 4 0	**(1-5)** 144 6 14		**(2)** 154 9 31	**(3)** 139 5 28	**(4)** 176 7 17
patient 7											
176.45	70.28	3.28	14.77	55.49	M-S	**(0)** 204 2 0	**(1)** 192 2 0	**(2-4)** 167 4 26		**(3)** 160 5 51	**(5)** 174 5 54
170.80	64.61	2.06	10.96	61.07	T	**(0-1-2)** 184 3 17			**(3)** 156 4 49	**(4)** 167 3 0	**(5)** 148 4 46
179.45	67.80	3.08	15.08	50.18	W	**(0)** 211 2 0	**(1)** 197 3 0	**(2-4)** 160 4 25		**(3-5)** 171 4 51	
161.55	60.39	3.40	8.71	62.60	R-F	**(0-4)** 175 2 1		**(1-5)** 150 4 23		**(2-3)** 160 5 47	
173.84	63.19	3.07	11.85	53.03	U	**(0)** 190 2 0	**(1-3-4)** 171 3 19			**(2)** 157 6 49	**(5)** 181 5 47
patient 8											
138.60	42.17	2.83	1.03	81.14	M	**(0-1-3)** 128 6 17			**(2-4)** 151 4 3		**(5)** 142 7 38
128.27	44.25	5.40	1.01	82.81	T-W-S	**(0-1-2)** 129 5 11			**(3)** 116 7 32	**(4-5)** 134 5 19	
131.77	48.46	7.15	1.95	77.48	R-F	**(0)** 141 4 1	**(1-4)** 131 5 21		**(2-5)** 135 5 21		**(3)** 118 7 35
143.04	45.81	6.04	1.58	74.85	U	**(0)** 154 4 0	**(1)** 166 4 0	**(2-3)** 120 6 18		**(4-5)** 146 5 12	
patient 9											
145.90	55.31	7.17	4.83	66.65	M	**(0-1-2-3)** 151 8 27				**(4)** 132 4 1	**(5)** 143 10 55
150.42	61.20	6.23	6.74	66.93	T-W-S-U	**(0-1)** 158 6 7		**(2)** 136 6 27	**(3)** 142 9 54	**(4-5)** 154 8 33	
143.12	53.96	7.13	3.09	69.57	R	**(0-1-4-5)** 148 7 23				**(2)** 131 5 22	**(3)** 136 9 58
155.23	62.26	3.25	7.84	66.36	F	**(0-3-5)** 148 6 26			**(1)** 188 7 11	**(2)** 137 7 54	**(4)** 165 8 37
patient 10											
177.30	76.22	4.58	14.29	48.78	M	**(0-2)** 159 3 14		**(1-3-5)** 179 4 30			**(4)** 202 5 36
168.56	78.29	7.84	15.96	52.72	T-F	**(0)** 143 2 6	**(1)** 164 3 41	**(2)** 149 4 36	**(3)** 157 3 21	**(4)** 194 6 44	**(5)** 203 5 37
163.74	76.13	10.45	14.86	51.35	W	**(0)** 155 3 7	**(1-3-5)** 166 4 30			**(2)** 143 4 39	**(4)** 190 4 15
181.96	85.28	6.77	21.44	46.92	R-U	**(0)** 174 3 9	**(1-2)** 161 3 22		**(3)** 166 3 25	**(4)** 227 6 39	**(5)** 194 5 31
190.48	90.48	4.89	23.50	45.24	S	**(0-1)** 209 3 15		**(2)** 136 4 38	**(3-5)** 173 4 29		**(4)** 247 5 31

Table 3.7: Results obtained with the CHi-square Automatic Interaction Detection algorithm for the groups obtained, taking into account the different days of the week and the time slots for patients 1, 2, 3, 4, and 5. In addition, the average values of insulin $\overline{I}$ and carbohydrates $\overline{C}$ are shown.

Day of the week	Time slot	$\overline{G}$ $[mg/dl]$	Std $[mg/dl]$	$T\_G_{<70}$ [%]	$T\_G_{>250}$ [%]	$T\_G_{[70,180]}$ [%]	$\overline{I}$ $[U/4h]$	$\overline{C}$ $[gr]$
patient 1								
M-W-S-U	**(0-1)**	171.81	56.89	2.35	8.92	55.06	3.69	2.88
M-W-S-U	**(2)**	158.99	57.21	2.80	6.33	64.05	5.89	25.12
M-W-S-U	**(3)**	164.00	60.79	1.60	8.60	67.67	7.44	135.88
M-W-S-U	**(4)**	188.65	67.60	1.36	19.36	44.58	4.94	12.25
M-W-S-U	**(5)**	167.57	63.57	2.92	10.26	59.64	6.93	92.12
T	**(0-1)**	198.45	77.38	0.64	17.89	41.75	4.04	3.50
T	**(2-4)**	169.97	69.57	0.84	10.97	61.73	5.66	18.50
T	**(3)**	158.01	54.07	0.92	7.04	68.17	7.88	131.50
T	**(5)**	176.60	65.35	0.00	16.52	60.27	7.38	87.50
R	**(0)**	170.32	63.92	0.11	11.01	58.68	3.90	0.00
R	**(1-2-3)**	162.89	57.86	1.36	7.40	65.82	5.42	56.00
R	**(4)**	186.35	71.65	0.53	21.65	50.26	5.30	13.50
R	**(5)**	155.94	50.81	4.03	2.84	63.14	7.07	102.00
F	**(0-5)**	169.67	59.49	1.35	9.93	61.50	5.17	40.50
F	**(1-2-3-4)**	160.44	57.08	2.94	7.68	63.07	5.25	40.12
patient 2								
M	**(0)**	136.95	60.66	6.90	4.38	70.78	2.62	1.50
M	**(1-2-4)**	155.93	61.24	4.71	7.24	66.97	3.97	94.00
M	**(3)**	146.11	50.02	4.15	1.16	66.47	3.31	154.00
M	**(5)**	127.01	54.74	11.55	3.19	73.79	4.01	174.00
T	**(0)**	170.44	68.95	5.02	12.37	56.83	2.76	0.00
T	**(1)**	178.86	59.36	0.09	10.61	57.39	4.50	88.00
T	**(2)**	156.51	58.00	2.78	10.04	70.79	4.32	108.00
T	**(3)**	138.60	50.24	7.47	1.64	70.63	3.15	162.00
T	**(4)**	147.69	51.31	6.30	1.69	67.80	3.62	92.00
T	**(5)**	131.75	50.25	4.51	2.46	82.28	4.20	156.00
W-U	**(0)**	168.58	65.71	2.74	12.52	59.71	2.77	4.50
W-U	**(1)**	177.79	63.44	1.26	14.92	55.37	4.30	83.00
W-U	**(2-4)**	162.98	62.46	4.75	9.22	57.83	3.62	103.50
W-U	**(3)**	143.99	58.12	6.30	4.99	69.67	2.69	128.00
W-U	**(5)**	136.01	57.04	9.71	5.28	71.07	3.91	151.75
R-S	**(0-4)**	157.36	59.57	4.58	7.13	64.06	2.99	45.00
R-S	**(1)**	162.96	54.73	2.10	7.66	64.14	3.72	68.00
R-S	**(2)**	151.59	52.55	4.47	4.96	68.66	3.92	114.00
R-S	**(3)**	144.66	54.81	5.47	4.47	69.26	2.89	131.00
R-S	**(5)**	132.43	59.05	12.14	4.15	69.70	3.63	141.00
F	**(0-2)**	147.85	50.37	3.56	3.37	74.05	3.04	46.00
F	**(1)**	163.49	48.66	0.98	5.71	63.21	3.74	68.00
F	**(3-5)**	138.52	56.14	6.64	4.44	71.59	3.33	139.00
F	**(4)**	154.49	68.34	5.74	10.45	63.58	2.98	88.00
patient 3								
M-T	**(0-3)**	142.30	41.21	3.10	1.46	79.75	12.07	60.88
M-T	**(1)**	139.63	35.60	1.13	1.05	88.13	8.16	5.25
M-T	**(2-5)**	147.05	49.21	2.10	3.63	75.30	14.53	115.12
M-T	**(4)**	131.98	39.85	3.18	1.00	86.89	7.60	1.75
W-U	**(0-1)**	149.22	39.11	1.38	1.61	79.54	8.15	2.00
W-U	**(2-3)**	145.84	40.43	1.13	1.70	81.05	15.62	119.38
W-U	**(4)**	135.75	47.15	5.26	2.61	76.51	7.27	0.75
W-U	**(5)**	163.62	53.02	1.82	6.08	64.19	14.58	133.25
R	**(0-2-3)**	160.24	53.78	0.69	6.26	70.39	13.73	76.67
R	**(1)**	135.67	35.08	1.18	0.00	85.73	7.18	0.00
R	**(4-5)**	154.80	51.06	0.67	5.25	71.08	11.40	64.25
F	**(0-2)**	146.63	46.15	1.36	1.04	76.31	10.73	43.00
F	**(1)**	121.16	30.63	3.27	0.00	93.16	7.93	10.50
F	**(3)**	151.87	36.42	0.08	0.15	80.92	17.74	145.00
F	**(4)**	132.00	39.19	2.54	0.07	84.84	8.10	2.00
F	**(5)**	139.13	48.44	2.79	3.36	79.14	15.61	143.00
S	**(0)**	128.66	27.54	1.34	0.00	95.92	7.08	0.00
S	**(1)**	133.11	26.72	0.81	0.00	92.42	7.39	0.00
S	**(2-3-5)**	143.40	42.42	1.73	0.56	77.93	16.20	120.50
S	**(4)**	122.29	38.84	2.14	1.51	90.71	7.50	0.00
patient 4								
M-W	**(0)**	167.37	54.81	2.89	7.96	58.39	4.53	0.00
M-W	**(1)**	179.76	62.77	2.36	16.14	52.72	9.68	114.75
M-W	**(2)**	171.34	63.23	3.52	11.52	53.82	5.58	58.50
M-W	**(3)**	148.66	53.32	3.08	5.06	72.61	10.61	184.50
M-W	**(4-5)**	155.94	55.95	2.96	5.40	65.23	7.88	76.62
T-F	**(0-2-3-4)**	149.77	52.32	3.09	4.07	70.91	6.34	67.94
T-F	**(1)**	146.11	48.23	2.63	3.54	77.50	8.85	130.75
T-F	**(5)**	162.42	62.44	2.28	7.26	66.18	10.49	102.75
R-U	**(0-1-3)**	160.63	56.01	2.44	6.59	65.27	8.62	98.00
R-U	**(2)**	168.24	63.52	3.19	11.98	57.92	5.96	44.50
R-U	**(4)**	140.68	53.90	4.80	5.23	73.54	3.70	11.00
R-U	**(5)**	152.64	55.68	2.05	7.42	73.83	10.55	115.00
S	**(0-2-3-4)**	154.80	53.70	2.37	5.51	68.61	5.93	57.88
S	**(1)**	146.37	52.35	2.59	5.41	77.09	9.50	118.50
S	**(5)**	160.73	49.79	0.65	5.03	66.32	10.95	108.00
patient 5								
M-F	**(0)**	126.18	67.28	21.46	6.77	58.33	1.23	0.00
M-F	**(1-3)**	145.31	59.52	7.78	5.39	66.92	4.06	157.75
M-F	**(2)**	135.62	56.82	7.46	5.45	72.21	4.76	74.75
M-F	**(4)**	149.32	75.22	12.02	9.84	60.01	1.94	13.50
M-F	**(5)**	164.62	69.48	7.63	10.19	52.09	5.20	190.00
T-W	**(0)**	152.30	77.11	11.72	12.71	54.40	1.14	0.00
T-W	**(1-3-4)**	148.35	64.26	7.41	7.99	66.18	3.55	85.83
T-W	**(2)**	128.25	52.86	9.30	2.60	74.32	4.17	65.00
T-W	**(5)**	162.68	63.09	7.23	8.89	55.84	3.95	214.25
R	**(0-2)**	127.30	55.21	12.08	3.75	73.25	2.79	31.25
R	**(1)**	136.74	48.16	10.55	0.65	71.37	5.05	37.00
R	**(3)**	141.51	52.09	7.07	3.09	71.16	3.63	213.00
R	**(4)**	148.71	70.64	4.96	8.60	69.02	2.14	27.00
R	**(5)**	160.54	56.39	5.75	3.70	57.21	5.26	247.50
S	**(0)**	118.57	62.42	25.94	2.06	52.96	0.99	0.00
S	**(1-2-3-4)**	126.37	50.39	12.89	0.81	70.82	3.67	105.12
S	**(5)**	149.79	62.83	4.39	7.88	68.04	4.49	225.50
U	**(0-1)**	146.79	68.16	4.94	8.60	72.08	3.18	17.50
U	**(2-4)**	118.88	60.57	18.27	4.29	68.32	3.29	49.25
U	**(3-5)**	132.48	61.33	10.70	6.35	67.76	4.08	254.00

Table 3.8: Results obtained with the CHi-square Automatic Interaction Detection algorithm for the groups obtained, taking into account the different days of the week and the time slots for patients 6, 7, 8, 9, and 10. In addition, the average values of insulin $\overline{I}$ and carbohydrates $\overline{C}$ are shown.

Day of the week	Time slot	$\overline{G}$ $[mg/dl]$	Std $[mg/dl]$	$T\_G_{<70}$ [%]	$T\_G_{>250}$ [%]	$T\_G_{[70,180]}$ [%]	$\overline{I}$ $[U/4h]$	$\overline{C}$ $[gr]$
patient 6								
M-T-S	**(0)**	135.39	66.19	14.76	7.93	62.39	3.70	1.17
M-T-S	**(1)**	167.41	59.78	1.71	7.39	61.64	6.42	14.83
M-T-S	**(2-4-5)**	160.83	61.35	5.00	8.39	60.17	7.41	21.83
M-T-S	**(3)**	154.81	52.21	4.14	4.31	63.97	6.68	28.17
W	**(0)**	109.69	46.41	21.54	0.35	68.78	3.16	0.00
W	**(1)**	123.62	39.04	11.46	0.00	81.29	6.81	29.00
W	**(2-5)**	155.98	54.79	4.31	5.61	61.28	6.37	19.00
W	**(3)**	145.46	52.34	7.05	0.80	66.07	6.03	20.50
W	**(4)**	168.21	51.00	4.05	3.86	56.46	6.21	9.00
R	**(0)**	110.00	47.18	20.28	1.28	70.12	3.39	0.00
R	**(1)**	123.01	44.52	13.89	0.21	77.47	8.10	36.00
R	**(2)**	166.58	46.84	1.36	4.08	58.89	7.87	24.00
R	**(3)**	138.97	50.09	7.26	1.09	70.51	6.35	25.50
R	**(4)**	151.27	62.10	6.23	8.07	65.99	8.94	24.00
R	**(5)**	145.42	59.30	10.88	6.69	64.05	7.76	26.50
F	**(0-5)**	122.00	60.58	22.43	3.00	60.32	6.01	16.00
F	**(1)**	151.19	49.66	1.61	3.88	75.47	8.83	37.00
F	**(2)**	173.61	66.54	1.94	13.32	55.87	6.29	18.00
F	**(3)**	182.35	83.79	0.28	14.77	63.49	7.13	38.50
F	**(4)**	161.45	58.61	4.25	8.11	63.30	7.19	16.00
U	**(0)**	127.30	75.14	24.69	6.62	54.97	3.55	0.00
U	**(1-5)**	143.76	62.91	10.35	5.78	62.34	5.90	14.50
U	**(2)**	154.46	48.54	4.09	4.53	70.80	8.90	31.50
U	**(3)**	139.02	52.98	7.25	1.97	71.19	5.40	28.00
U	**(4)**	175.57	70.43	4.27	15.04	59.96	7.49	17.50
patient 7								
M-S	**(0)**	204.20	74.85	0.23	25.78	42.92	2.20	0.00
M-S	**(1)**	192.09	67.01	0.63	13.96	46.76	1.98	0.00
M-S	**(2-4)**	166.58	67.12	4.31	11.03	60.34	4.15	26.25
M-S	**(3)**	160.16	67.65	6.52	9.56	60.12	4.53	51.25
M-S	**(5)**	174.31	72.13	3.48	17.51	56.89	4.76	54.00
T	**(0-1-2)**	184.11	70.51	0.91	17.20	55.70	3.20	17.00
T	**(3)**	155.80	60.29	4.29	5.98	65.02	4.38	49.50
T	**(4)**	166.54	49.99	1.38	4.15	63.02	2.74	0.00
T	**(5)**	147.93	51.58	4.18	3.46	72.33	4.05	46.50
W	**(0)**	211.10	71.65	0.94	30.97	34.05	2.27	0.00
W	**(1)**	197.33	49.86	0.00	12.37	35.26	2.81	0.00
W	**(2-4)**	159.56	54.74	2.91	5.82	64.33	4.12	25.25
W	**(3-5)**	171.02	77.44	6.39	17.19	53.72	4.46	51.00
R-F	**(0-4)**	174.57	61.54	1.69	11.34	57.86	2.39	1.38
R-F	**(1-5)**	150.34	50.76	4.27	4.03	68.31	3.53	23.12
R-F	**(2-3)**	159.82	65.78	4.28	10.94	61.52	4.86	47.25
U	**(0)**	190.22	57.89	3.56	10.99	33.13	1.86	0.00
U	**(1-3-4)**	171.11	60.07	3.71	12.05	54.58	3.44	19.50
U	**(2)**	156.57	66.54	0.84	9.20	72.41	5.67	49.00
U	**(5)**	181.15	68.79	2.76	14.37	50.51	4.80	47.50
patient 8								
M	**(0-1-3)**	127.82	35.79	1.97	0.20	89.73	6.20	16.83
M	**(2-4)**	150.83	45.08	3.46	1.86	69.95	3.52	3.00
M	**(5)**	142.33	44.19	3.68	1.52	80.52	6.95	38.00
T-W-S	**(0-1-2)**	128.57	45.38	6.72	0.81	80.48	4.73	11.33
T-W-S	**(3)**	116.48	32.43	3.74	0.00	91.55	7.13	32.17
T-W-S	**(4-5)**	133.97	46.63	4.29	1.83	81.73	4.56	19.17
R-F	**(0)**	141.40	52.86	7.58	2.85	71.66	4.13	1.50
R-F	**(1-4)**	130.77	51.58	10.43	1.86	73.42	4.91	20.62
R-F	**(2-5)**	135.09	48.25	5.05	2.65	77.56	5.16	21.12
R-F	**(3)**	118.26	32.53	4.24	0.00	90.81	7.15	35.50
U	**(0)**	154.03	47.65	5.47	1.29	68.17	4.24	0.00
U	**(1)**	165.53	38.05	2.65	2.65	68.87	3.79	0.50
U	**(2-3)**	119.91	39.30	13.36	0.00	79.76	6.00	18.00
U	**(4-5)**	145.70	45.14	1.71	2.57	77.23	4.53	12.00
patient 9								
M	**(0-1-2-3)**	150.99	58.00	8.02	6.09	59.98	7.97	26.62
M	**(4)**	132.18	46.10	5.49	2.13	81.71	4.13	1.00
M	**(5)**	143.12	52.27	6.07	3.42	73.25	9.67	55.50
T-W-S-U	**(0-1)**	158.43	66.87	7.33	9.43	59.62	5.53	7.06
T-W-S-U	**(2)**	136.18	56.59	8.21	4.80	74.34	5.98	27.50
T-W-S-U	**(3)**	141.94	53.42	4.57	4.06	75.69	8.82	54.12
T-W-S-U	**(4-5)**	153.72	59.49	5.04	6.40	66.10	7.96	32.88
R	**(0-1-4-5)**	147.65	56.56	7.60	3.85	64.89	6.59	22.88
R	**(2)**	131.07	52.35	9.09	1.69	71.34	4.85	22.50
R	**(3)**	136.43	41.73	3.67	1.44	85.58	8.99	58.00
F	**(0-3-5)**	147.75	58.38	2.41	5.34	72.47	6.21	25.67
F	**(1)**	188.24	70.74	1.06	19.16	50.83	6.52	11.50
F	**(2)**	136.80	49.60	3.90	3.76	77.44	7.21	54.00
F	**(4)**	165.06	63.44	7.14	8.71	51.57	8.18	37.00
patient 10								
M	**(0-2)**	159.21	71.65	5.20	11.22	61.01	2.63	13.75
M	**(1-3-5)**	179.09	75.79	4.49	13.69	46.73	3.66	29.67
M	**(4)**	202.40	77.01	3.79	20.98	34.04	5.47	36.50
T-F	**(0)**	142.51	76.45	15.00	8.48	62.62	2.32	6.50
T-F	**(1)**	163.57	69.95	6.32	13.74	58.54	3.21	41.50
T-F	**(2)**	148.95	64.78	9.55	8.35	58.12	4.06	36.25
T-F	**(3)**	156.74	75.08	11.90	11.58	52.61	2.81	21.00
T-F	**(4)**	194.02	79.76	1.91	24.54	45.86	5.62	44.00
T-F	**(5)**	202.58	81.48	2.50	28.11	40.94	4.74	36.75
W	**(0)**	154.82	80.87	11.64	12.70	55.42	2.64	7.50
W	**(1-3-5)**	166.01	80.11	11.83	16.8	47.14	3.53	30.50
W	**(2)**	143.31	60.95	10.98	7.23	67.47	3.73	39.00
W	**(4)**	190.19	63.54	3.99	19.82	41.78	4.25	15.50
R-U	**(0)**	174.50	89.03	5.97	19.30	52.19	2.84	9.25
R-U	**(1-2)**	161.44	75.46	7.48	13.23	57.97	3.03	22.50
R-U	**(3)**	166.24	78.74	9.78	13.98	51.29	3.35	25.50
R-U	**(4)**	226.94	86.61	3.90	42.04	24.95	6.04	38.75
R-U	**(5)**	194.45	83.52	6.06	24.18	41.01	4.76	31.50
S	**(0-1)**	208.59	82.35	3.06	27.98	34.58	2.81	15.25
S	**(2)**	135.70	69.00	11.76	7.26	62.58	3.87	38.50
S	**(3-5)**	173.16	83.28	4.83	16.49	57.30	3.77	28.75
S	**(4)**	247.22	96.84	1.70	45.93	22.36	5.48	31.50

3.3 Conclusions

The conclusions obtained from this work are:

- Significant differences and dependencies have been observed among glucose profiles classified according to the variables day of the week and time slot.
- The groups found have been different for each patient, demonstrating the need for an individualized study.
- Tables 3.4 to 3.8, allow us to find significant differences to correct and improve habits or therapies in patients, and obtain more precise models through ML techniques and AI.
- The results obtained indicate that the techniques applied can facilitate the mathematical modeling of glucose and may be used to create an individualized classifier for each patient that classifies glucose profiles according to the variables day of the week and time slot. The patients' glucose values can be predicted using this classifier by knowing what day the patient is in and in what time slot, obtaining more accurate models.

Chapter 4

Glucose Prediction using Clustering and GP

This chapter is a continuation of the storyline of chapter 3. Here a methodology to obtain accurate predictors of glucose using a three-step process is proposed. First, data is collected from CGM, preprocessed, and divided using division criteria based on the behavior of the glucose time series under study. Second, data is clustered to obtain customized models for different glucose profiles. Third, a training step is used to create ensemble prediction models based on evolutionary computation. Those models will be applied to forecast future glucose values. The main objective of this chapter is to improve the predictions of glucose values from previous works and increase the time horizon of reliable predictions.

The rest of the chapter is organized as follows. Section 4.1 explains the methodology, describing the techniques to model the glucose and calculate the models' accuracy and quality. The experimental results, including the dataset, configuration, and results, are shown in section 4.2. The conclusions are discussed in section 4.3.

4.1 Methodology

The proposed methodology is a three-step process:

- Data collection, preprocessing, and division.
- Data clustering and detection of a set of glucose profiles.
- Models training, creating models by evolutionary computation.

Figure 4.1 represents the diagram of the process. When a prediction is requested, a profile is detected, then the predictor is selected and applied to obtain the glucose prediction and eventually makes an insulin dosing recommendation. These steps are explained below.

4.1.1 Data Collection, Preprocessing and Division

Data from PwD is collected using CGM. Glucose measurements every five to fifteen minutes are recorded with the annotation of intakes and carbohydrate estimates and the insulin doses injected by the patients.

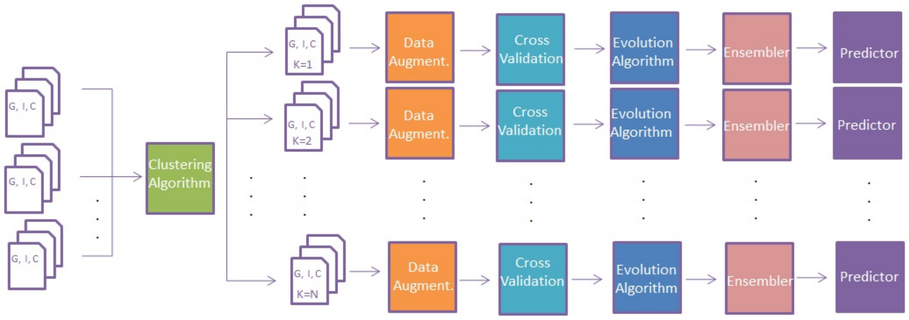


Figure 4.1: Flow diagram describing the data clustering and models training.

When collecting data with a CGM for long periods, it is usual to find some missing data. To solve this problem as proposed in [98, 99], the values are imputated, performing a correction of the values using segmented spline interpolation of degree 3, where the maximum number of consecutive imputated values is one hour [100]. In the case of insulin doses, a function to simulate the plasma insulin dynamics after subcutaneous injection of insulin based on the Berger model is used [101]. Both methods are implemented with the R free software environment for statistical computing and graphics version 3.6.1 [97].

Data is divided according to a criterion that creates different glucose profiles. The division criteria are based on the behavior of glucose time series under study, such as splitting time series into 4-hour non-overlapping slots or splitting, taking into account food intakes (i.e., breakfast, lunch, and dinner). In this research, we used the division criteria applied in 3.2.2. This criterion allows easy implementation of selection criteria in the prediction phase. The system will select the cluster (profile), knowing only the day and hour of the prediction time and consequently the predictor to apply to each situation. Figure 4.2 shows a diagram of the data collection, preprocessing, and division steps.

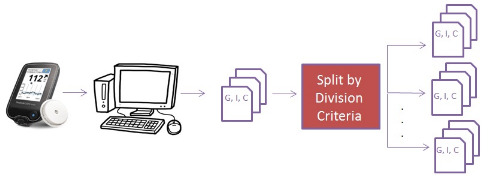


Figure 4.2: Flow diagram describing the data collection, preprocessing, and division.

4.1.2 Data Clustering

In this step, a clustering algorithm is applied to explore the data and get hidden information. The objective is to obtain glucose profiles to train specific models on each profile. So a cluster will include all the glucose data assigned to a profile. Here CHAID implemented with the IBM SPSS predictive analysis software version 23.0 [94] is applied to recursively divide the data concerning a target variable using multiple divisions between the different input variables: days of the week and time slots. A division must reach a threshold level of significance ($\alpha = 0.05$) using the independence test *F-Snedecor* between the nominal values of the target variable and the branches. If not, the node is not divided. The Bonferroni setting is used for the number of categorical values of the input variable, thus mitigating the bias towards entries with many discounts. The search ends when the algorithm can no longer join more branches or significant divisions. The last division is chosen as the solution. Note that typically the last division is not the most significant examined.

For the construction of the decision trees, we use the exact parameters of section 3.2.2. Once the classification is obtained, the groups of glucose values are selected as training datasets.

4.1.3 Models Training

This is the most extensive step. First, a data augmentation algorithm [102] that generates synthetic glucose time series is used in training datasets to develop meaningful information as well as significantly enhance data quality.

Next, models based on evolutionary algorithms such as GP or GE are created using cross-validation. In this case, GP is selected as the evolutionary algorithm to create models for the different datasets obtained in the previous step. Cross-validation of ten iterations is used to create the models (10-fold cross-validation), where each iteration is repeated ten times. Each model is generated using the open-source software HeuristicLab version 3.3.15 [103].

GP is an evolutionary computation technique that automatically solves problems without requiring the user to know the structure of the solution in advance. At the most abstract level, GP is a systematic, domain-independent method for getting computers to solve problems, automatically starting from a high level. In GP we evolve a population of computer programs. That is, generation by generation, GP stochastically transforms populations of programs into new, hopefully, better, populations of programs. GP, like nature, is a random process, and it can never guarantee results. However, it can lead to escape traps by which deterministic methods may be captured. Like nature, GP has been very successful at evolving novel and unexpected ways of solving problems. GP finds out how well a program works by running it and then compares its behavior to some ideal. This comparison is called fitness. Those programs that do well are chosen to produce new programs for the next generation.

The basic steps in a GP system are shown below:

```
Randomly create an initial population of programs from the available primitives (neither the
  growth nor the complete method);
while An acceptable solution is found or some other stopping condition is met (e.g., a maximum
  number of generations is reached) do
    Evaluation: execute each program and ascertain its fitness;
    Selection: select one or two programs from the population with a probability based on fitness
      to participate in genetic operators;
    Crossover: creating a child program by combining randomly chosen parts from two selected
      parent programs;
    Mutation: creating a new child program by randomly altering a randomly chosen part of a
      parent program chosen;
end
Return to the best-so-far individual;
```

Algorithm 2: Step-by-step of the Genetic Programming algorithm.

GP [104] is used for solving a SR problem. The models based on GP are generated using the time series of glucose, insulin, carbohydrates, and a set of features added in periods calculated using the following equation (being X: glucose G, insulin I, or carbohydrates C):

$$mean(X,t,range) = \frac{\sum_{t \in range}(X_t)}{n}, \quad range \in [t_1, ..., t_n] \tag{4.1}$$

The goal is to reduce the number of values used to generate the models. For each time t, the set of features $F(t)$ that describe the historical values F_{his} of the time series (values of glucose G, insulin I, and carbohydrates C) up to t, as well as the future values F_{fut} of insulin and carbohydrates, are defined using the following equations:

$$\begin{aligned} F(t) = F_{his}(G,t) \cup F_{his}(I,t) \cup F_{his}(C,t) \cup \\ F_{fut}(I,t) \cup F_{fut}(C,t) \cup \{G,I,C\} \end{aligned} \tag{4.2}$$

$$\begin{aligned} F_{his}(X,t) = \{&mean(X,t,[-15,0]), mean(X,t,[-30,-15]), \\ &mean(X,t,[-45,-30]), mean(X,t,[-60,-45]), \\ &mean(X,t,[-90,-60]), mean(X,t,[-120,-90]), \\ &mean(X,t,[-150,-120]), mean(X,t,[-180,-150]), \\ &mean(X,t,[-210,-180]), mean(X,t,[-240,-210])\} \end{aligned} \tag{4.3}$$

$$\begin{aligned} F_{fut}(X,t) = \{&mean(X,t,[0,15]), mean(X,t,[15,30]), \\ &mean(X,t,[30,45]), mean(X,t,[45,60]), \\ &mean(X,t,[60,75]), mean(X,t,[75,90]), \\ &mean(X,t,[90,105]), mean(X,t,[105,120]), \\ &mean(X,t,[120,135]), mean(X,t,[135,150]), \\ &mean(X,t,[150,165]), mean(X,t,[165,180]), \\ &mean(X,t,[180,195]), mean(X,t,[195,210]), \\ &mean(X,t,[210,225]), mean(X,t,[225,240])\} \end{aligned} \tag{4.4}$$

4.1.4 Selection of the Best Models

The best model of the ten repetitions is selected by the Akaike Information Criterion (AIC) [105]. This method is used to measure the relative quality of models based on the entropy of information. It provides a comparative estimate of the information lost when a model is used to represent the process that generates the data. The following equation defines the criterion:

$$\text{AIC} = n \cdot \log(\text{SE} + 1) + 2(m+1) \tag{4.5}$$

where n is the sample size, m is the number of model parameters and Square Error (SE) is the residual quadratic error, which is defined by:

$$\text{SE} = \sum_{i=1}^{n} (X_i - Y_i)^2 \tag{4.6}$$

where X_i is the real value i, and Y_i is the estimated value i in the prediction.

One model for each time horizon and fold is selected (the best model is the model with the lowest AIC value). Models for predicting glucose values at intervals of 30 minutes for a maximum of four hours are obtained (time horizons at 30, 60, 90, 120, 150, 180, 210, and 240 minutes). Then, an ensembler method is used to create the predictor. This technique uses multiple models made in the previous step and then combines them to produce more accurate solutions than a single model would.

Finally, the predictor is selected to forecast future glucose values. The data in this phase is classified into one of the clusters obtained in the training phase. The predictor that has been accepted for this cluster is selected and used to make the final prediction.

4.1.5 Parkes Error Grid

Parkes Error Grid (PEG) method [106] is used to evaluate the accuracy of the predictions for each time horizon. This method was published in 2000 as an alternative to CEG [57]. These methods were developed to calculate the clinical accuracy of CGM for the entire range of glucose values, using the differences between the reference values and the values measured by the measurement system. Analogously, it can be used to represent the differences between the values estimated in a prediction and the current or reference values in a graph with Cartesian coordinates, where the X-axis represents the reference values and the Y-axis the values of the prediction, and Y = X is the ideal prediction. The unique feature of this representation is that the graph is divided into five zones depending on the degree of accuracy of the glucose estimates. The difference between the two methods lies in the definition of the zones. In PEG, the zones are redefined based in the zones of CEG and in the limits established by 100 medical experts in diabetes in a survey carried out in the *American Diabetes Association Meetings* in June 1994. The new zones are defined as follows:

- Zone A: glucose values without effect on clinical action. The estimates are accurate.
- Zone B: glucose values with alteration of the clinical action, with little or no effect in the clinical treatment.
- Zone C: glucose values with alteration of the clinical action, with a probability of effects in the clinical treatment.
- Zone D: glucose values with alteration of the clinical action, with a probability of significant medical risk.
- Zone E: glucose values with alteration of the clinical action, probability of having dangerous consequences.

The goal is to maximize the predictions included in zones A and B and minimize those in zones C, D, and E. It should be noted that neither the data augmentation nor the ensemble model in the training phase has been applied in this work. These two techniques will be used in work presented in chapter 5.

4.2 Experimental Results

4.2.1 Dataset

Data was collected from a group of ten patients with T1DM. Table 3.1 shows the number of days and observations of data for each patient, as well as the mean, SD, and percentages of time where the patient has glucose levels below 70 mg/dl, above 250 mg/dl, and in the range [70, 180] mg/dl known as the time in range. Patients have a high SD. This is normal in PwD due to the disease and diabetes control.

4.2.2 Configuration

The set of terminals use the time series of glucose, insulin, carbohydrates and constants. Each value is accompanied by a real value weight that is initialized randomly using a Normal distribution, $N(\mu = 1, \sigma = 1)$, and is randomly mutated by adding a sampled value from $N(\mu = 0, \sigma = 0.05)$. Constants with real values are initialized randomly using a Uniform distribution $U(-20.0, 20.0)$, and the constant mutation adds a sampled value from $N(\mu = 0, \sigma = 1.0)$.

The set of functions used are: $\{+, -, *, /, log(x), exp(x)\}$, where the protected variants of the functions division and logarithm are used [104]. The models in the initial population are generated using Probabilistic Tree Creator [107] with a limit to the maximum depth and the maximum number of nodes allowed for the trees. The same depth and size restrictions are applied in the crossover and mutation operations. The crossover uses a sub-tree crossover operator. The mutation uses various operators that replace a complete sub-set of the tree with a tree initialized randomly, mutate all the tree's nodes, or mutate only one node of the selected tree randomly. The mutation operator selected randomly is executed after each crossover operation with a different mutation rate. A maximum number of generations and parents' proportional aptitude selection is used. The objective function in all cases is calculated with Pearson's COD R^2 between the real BG values and the values obtained with the model [108]. The selected prediction models are linearly scaled to minimize the sum of the quadratic errors between those values. The parameters of the GP are not tuned specifically for this task. Robust standard configurations are applied.

The parameters selected for each model are the same. A population size of 1500 individuals, a maximum tree depth of 11, a maximum number of generations and a maximum number of nodes of 100, and a mutation rate of 0.15 and crossover rate of 0.90 are selected.

In this work, two different methods are compared. The first method is referred to as *CHAID-GP* in the work. It consists of models created with the general method explained in section 4.1. The second method corresponds to previous approaches, referred to as *GP* in the rest of the work. In this case, the division is not used, and therefore no clustering algorithm is applied. *GP* models are created with the original dataset; thus, there are no customized models for different glucose profiles. The goal is to compare the new general method with the traditional method used as the reference method. Both methods are compared in terms of PEG to evaluate the accuracy of the predictions for each time horizon.

4.2.3 Results

Table 3.4 shows the results obtained with the CHAID algorithm. The final depth of the tree is 2, the number of nodes is 33, and the number of terminal nodes is 23. The first predictor (variable) used in the construction of the tree is the day of the week and the second is the time slot. Significant differences and associations are observed and discusses in sections 3.2.3 and 3.2.4.

A comparison between *GP* and *CHAID-GP* is made in figures 4.3 and 4.4. The ratio of predictions in zone A+B for the different clusters of data is shown in figure 4.3. In 154 out of 189 cases (81.48%), models created with *CHAID-GP* are more accurate (120 out of 189) or at least of equal accuracy (34 out of 189) than models created with *GP*. Moreover, in 16 cases *CHAID-GP* predictions are the best for all time horizons. Figure 4.4 shows the ratio among the values obtained in zone A+B for the time horizons of models created with *CHAID-GP* and *GP*. It is observed that in 67 out of 80 cases (83.75%), models created with *CHAID-GP* are more accurate (59 out of 80) or at least of equal accuracy (8 out of 80) than models created with *GP*. Again, *CHAID-GP* predictions are the best for all clusters in 2 cases.

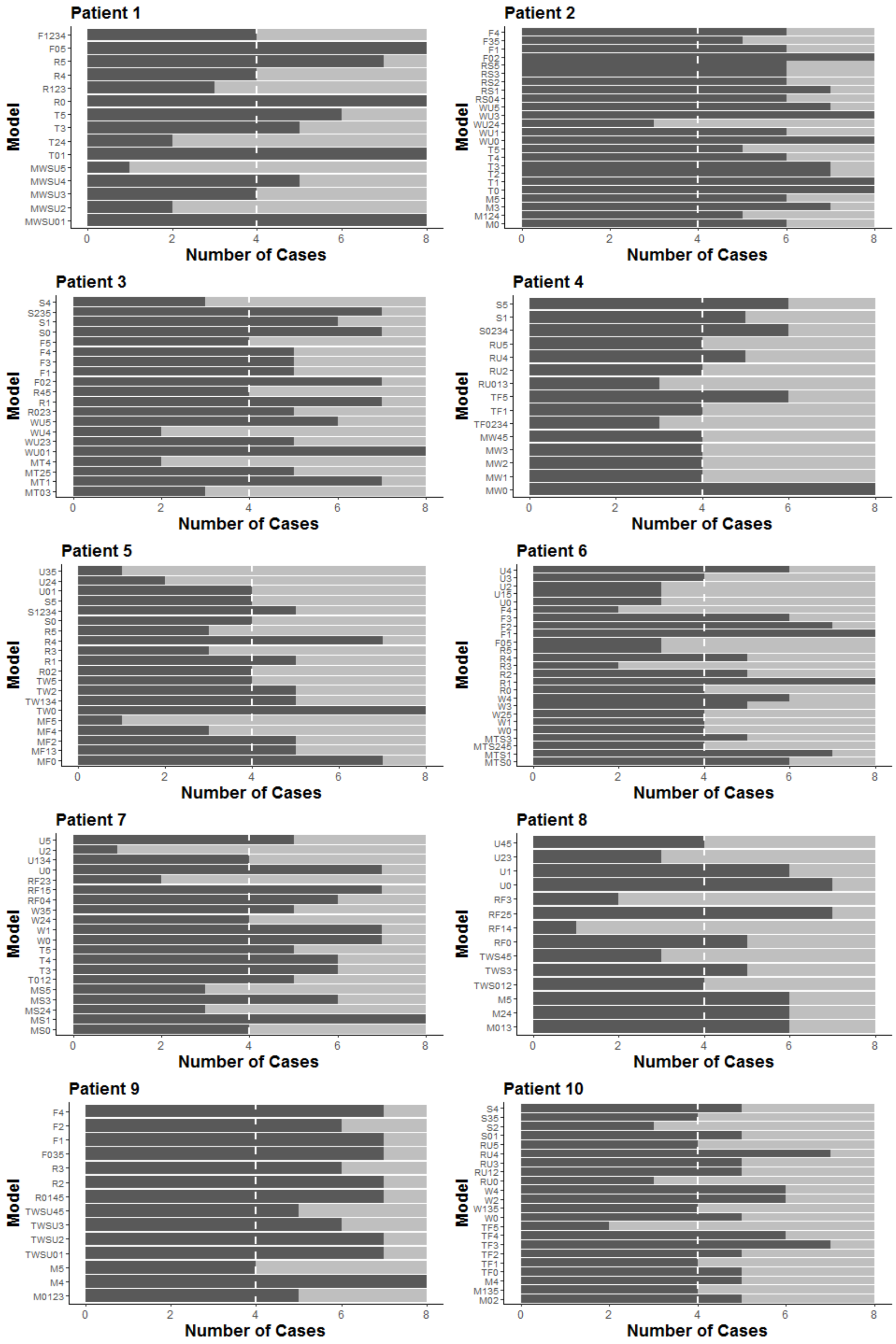


Figure 4.3: The ratio of models with better predictions (higher values in zone A+B) for different data clusters. Labels in the Y-axis identify clusters accordingly to table 3.4, columns day of the week + time slot. Dark-gray segments indicate that the best prediction is made by *CHAID-GP* while light-gray means a better model from *GP*.

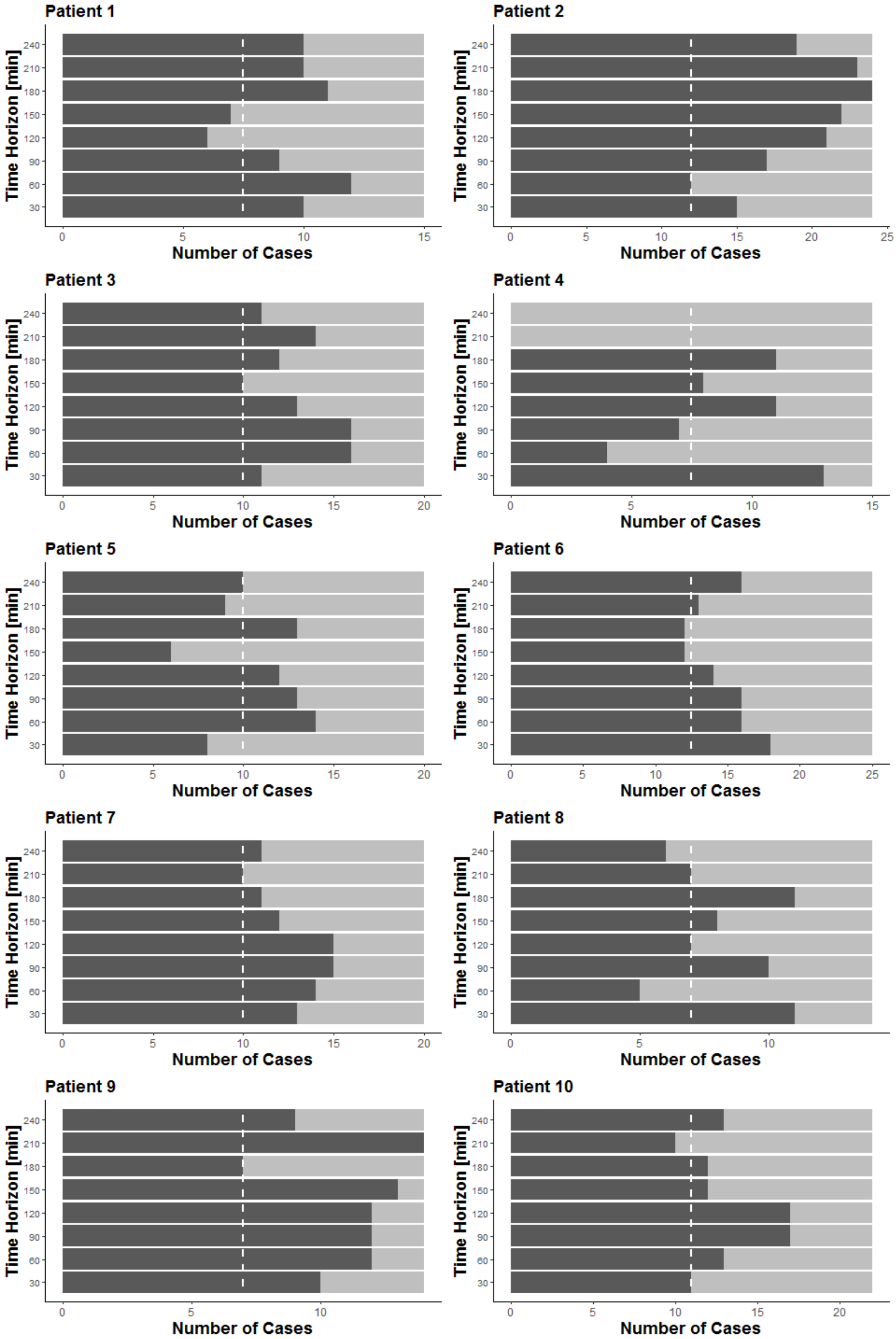


Figure 4.4: The ratio of models with better predictions (higher values in zone A+B) for the different time horizons of data. Dark-gray segments indicate that the best prediction is made by *CHAID-GP* while light-gray means a better model from *GP*.

Table 4.1 shows the predictions (in percentage) in the evaluation phase obtained for the zone A+B with PEG for all time horizons and patients (except 210 and 240 minutes included in table A.1). The accuracy of the models is analyzed under the assumption that the best model obtains a higher percentage of values in zone A+B and a lower percentage of values in zones C, D, and E. This assumption is a consequence of the meaning of the zones (see section 4.1.5). In general, the accuracy of predictions is better for shorter time horizons and gradually gets worse as the time horizon increases from 30 to 240 minutes. Note that models created with glucose values classified in categories with fewer elements obtained the best results. In 46 out of 60 cases (76.67%), models created with *CHAID-GP* are more accurate than models created with *GP*. It is significant that for patients 7 and 9, all models created with *CHAID-GP* are better than those models created with *GP*. For 180 minutes, all patients (except patient 6) have more accurate models created with *CHAID-GP*.

Table 4.1: Predictions (in percentage) obtained for the zone A+B with Parkes Error Grid for *GP* models, and the average percentage for *CHAID-GP* models. *Green* values indicate better solutions for *CHAID-GP*. The best result is highlighted in bold.

Model	t+30	t+60	t+90	t+120	t+150	t+180
Patient1						
GP	95.14	94.70	94.44	94.62	92.54	91.02
CHAID-GP	**96.32**	96.10	95.28	93.53	91.77	92.17
Patient2						
GP	89.81	90.89	87.33	83.59	71.75	72.32
CHAID-GP	90.45	90.36	90.43	89.77	91.03	**91.37**
Patient3						
GP	96.77	95.21	95.94	96.86	**98.08**	97.66
CHAID-GP	96.04	97.06	96.66	96.95	97.03	97.66
Patient4						
GP	91.19	64.60	92.94	89.87	92.50	93.42
CHAID-GP	93.74	93.13	92.16	91.98	92.64	**94.06**
Patient5						
GP	87.60	81.89	85.87	79.72	**89.28**	85.50
CHAID-GP	86.18	86.21	85.99	83.20	85.09	86.03
Patient6						
GP	89.20	85.61	85.68	85.79	88.62	**91.57**
CHAID-GP	90.46	88.65	89.22	88.11	86.92	90.03
Patient7						
GP	89.56	87.97	89.52	87.59	91.29	91.24
CHAID-GP	91.22	90.57	**92.76**	91.96	91.44	92.60
Patient8						
GP	88.95	92.94	92.38	**94.05**	93.41	86.33
CHAID-GP	91.27	92.50	93.84	90.55	92.69	88.84
Patient9						
GP	87.23	86.45	85.27	82.40	79.73	90.03
CHAID-GP	91.55	90.81	90.03	85.51	**92.17**	90.55
Patient10						
GP	**89.39**	80.98	80.87	78.53	84.79	86.77
CHAID-GP	86.91	84.08	85.71	83.65	86.66	88.06

4.3 Conclusions

In this work, a new general method to explore the glucose values and get hidden information obtaining glucose profiles to train customized models based on GP is applied. This new method is compared with the traditional one, where no clustering algorithm is used.

The main conclusions of this work can be summarized as follows:

- Significant differences (p-value < 0.05) and associations are observed among the glucose profiles classified using the independent variables day of the week and time slot.
- Glucose predictions with models created with GP are better for shorter time horizons and gradually

worsen as the time horizon increases from 30 to 240 minutes.

- Models created with glucose values classified in categories with fewer elements obtained the best results.
- In general, when using classified glucose values, the accuracy of the glucose values improves compared to those of models made with the original dataset.
- Significant differences found in the classification process can be helpful to correct and improve habits or therapies in patients and to obtain more accurate models through automatic learning techniques and AI.
- The results obtained will facilitate the mathematical modeling of glucose and can be used to create an individualized classifier for each patient.

Chapter 5

Contribution of Latent Glucose Variability Features

This chapter is an extension of the research line of chapter 4, where a methodology to obtain accurate predictors of glucose using a three-step process is proposed. We design a *Data-driven* modeling approach where data collected from CGM is used to solve a SR problem employing GP. CGM data includes information on IG, injected units of insulin, and ingested carbohydrate quantities. After a preprocessing step, for imputation, data is divided by time slots in the day and days of the week. Our method seeks patterns from the split dataset to train customized models for the different glucose profiles or patterns. In this chapter, we incorporate a set of GV measures that are obtained from the data.

Ensemble prediction models that take into account both the classified glucose profiles and the LGV (through different measures of glycemic average, GV and glycemic risk) are obtained using GP.

The main contributions of this work concerning chapter 4 are:

- We propose the use of several GV measures as new input features of the modeling engine. We call them LGV features since they are obtained from the original glucose values.
- We show how taking GV into account can lead to better predictions.
- We perform a study of the relative importance of the features.
- We test the experimental results for statistical validation with a Bayesian approach.

The rest of the chapter is organized as follows. Section 5.1 explains the methodology. The experimental results are shown in section 5.2 and conclusions are discussed in section 5.3.

5.1 Methodology

As an extension of the work proposed in chapter 4, in this work, we include different measures of glycemic average, GV, and glycemic risk to calculate the LGV of the patients. These measures are applied as input variables of the GP models to improve the accuracy in the prediction of the glucose values and study the contribution of those variables in the makeup of the models.

5.1.1 Data Collection, Preprocessing and Division

Figure 5.1 shows a diagram of the data collection, preprocessing, and division steps. Data from PwD is collected using CGM and insulin injection devices. Glucose measurements are recorded every five to fifteen minutes with the annotation of intakes and carbohydrate estimates. The insulin doses are injected manually by the patient or by the device. As figure 5.1 shows, this process can be extended to incorporate information from other devices, such as smartwatches, smartphones, or different types of insulin pumps.

Glucose data is imputated by performing a correction of the values using segmented spline interpolation of degree 3, where the maximum number of consecutive imputated values is one hour [100]. In the case of insulin doses, a function to simulate the plasma insulin dynamics after subcutaneous injection of insulin based on the Berger model is used [101].

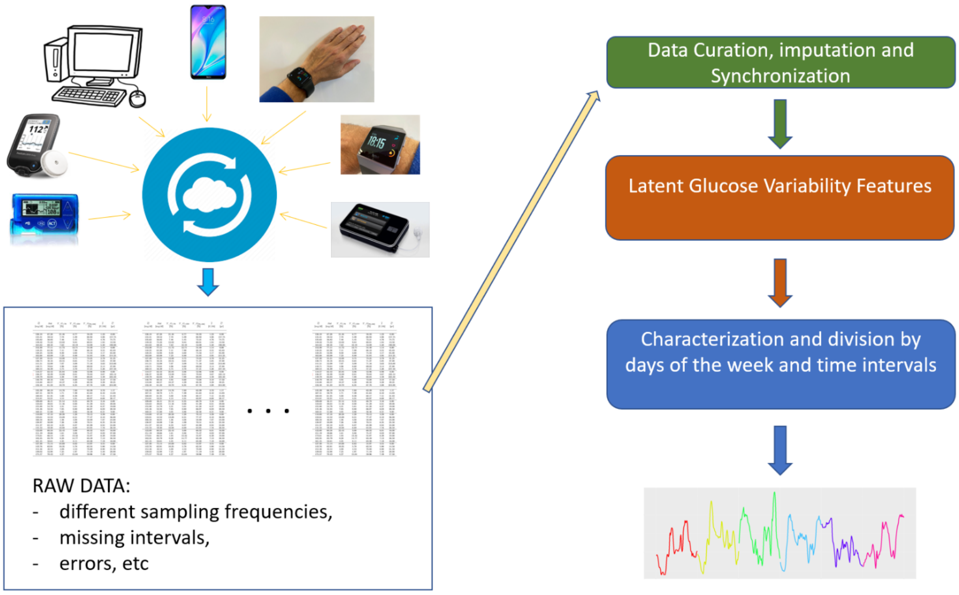

Data Curation, imputation and Synchronization
Latent Glucose Variability Features
Characterization and division by days of the week and time intervals
RAW DATA:
- different sampling frequencies,
- missing intervals,
- errors, etc
. . .

Figure 5.1: Flow diagram describing the data collection, preprocessing, and division.

Once the data is curated, organized, and synchronized by timestamps, a set of additional features are generated based on different measures of GV. The usual way of measuring GV is the HbA1c level, which, roughly speaking, indicates the percentage of blood red cells with glucose adhered to them during the last three months. In addition to the HbA1c level, there are different measures of glycemic averages and numerous measures that have been developed to evaluate the GV (see chapter 2). Due to the large number of measures that exist and the high degree of correlation among some of them, it isn't easy to determine which one is the most appropriate [38].

As we mentioned in section 5, we investigate if taking into account GV could improve the prediction of the models obtained with GP algorithms. For this purpose, we calculate different average measures of glycemia, measures of GV and measures of glycemic risk, named LV, that are used as additional input variables for the training step. We experimented with the 18 measures reported in table 5.1. One of the advantages of GP is that if there is a correlation between two variables, models include only the most important one. To check that, we also performed a study of the importance of the GV in the models. All measurements are calculated using 4-hour non-overlapping slots.

For AUC and PSTR the values below/above the limits of hypoglycemia and hyperglycemia are 70 mg/dl and 180 mg/dl, respectively. CONGA is calculated for one hour, metabolic control for LI of one hour, MAGE uses 0.5 times the standard deviation, and MV, the ideal glucose, is fixed to 120 mg/dl. All the necessary values were selected based on medical literature [33, 109].

Table 5.1: Glucose variability measures used. Measures of glycemic risk can also be considered as measures of glucose variability.

Measure	Acronym
Measures of Glycemic Average	
Arithmetic mean	mean
Area Under Curve	auc
Area Under Curve below the hypoglycemic limit	lauc
Area Under Curve above the hyperglycemic limit	hauc
Percentage Spent below/above the Target Range	pstr
Low Percentage Spent below the Target Range	lpstr
High Percentage Spent above the Target Range	hpstr
Measures of Glucose Variability	
Standard Deviation	sd
Coefficient of percentual Variation	cv
Continuous Overall Net Glycemic Action	conga
J Index	ji
Lability Index	li
Mean Amplitude of Glycemic Excursions	mage
Glycemic Excursions	ge
Low Mean Amplitude of Glycemic Excursions	lmage
High Mean Amplitude of Glycemic Excursions	hmage
M Value	mv
Measures of Glycemic Risk	
Average Daily Risk Range	adrr

After all the LV have been generated, data is divided according to criteria that create different glucose profiles. The division process splits data by days of the week and 4-hour non-overlapping time slots. Hence, there are seven categories for the variable day of the week and six for the variable time slot. The categories of the variables are represented by letters and numbers, as shown in table 3.3. These criteria permit a direct selection of the adequate model in the prediction phase. Knowing the day and hour of the prediction time, we can select the profile and consequently the predictor to apply in each situation.

5.1.2 Data Clustering

In work presented in chapter 3, we applied CHAID clustering algorithm to obtain glucose profiles and to train specific models on each profile. Each cluster includes all the glucose data assigned to each profile. In this work, we follow the same procedure, CHAID is applied to divide the data about the days of the week and time slots.

For the construction of the decision trees, we use the same parameters of section 3.2.2. Once the classification is obtained, the groups of glucose values are selected as training datasets.

5.1.3 Models Training

Figure 5.2 represents the diagram of the modeling process. First, a data augmentation algorithm [102] that generates synthetic glucose time series is used in training datasets to develop meaningful information as well as significantly enhance data quality. Next, models based on GP evolutionary algorithms are created using cross-validation. GP is implemented using the open-source software HeuristicLab version 3.3.15 [103].

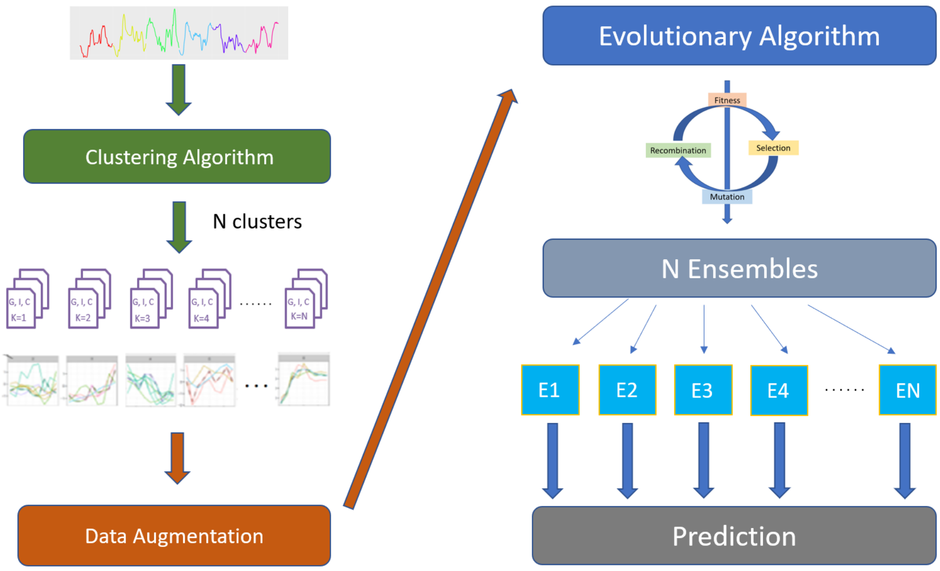


Figure 5.2: Flow diagram describing the data clustering and models training. For N clusters found in the classification process, N predictors are generated.

We tackle the problem as a SR problem. Models based on GP are generated using time series of glucose, insulin, carbohydrates, GV measures (see section 5.1.1) and a set of historical and future features generated by averaging (function *mean* in equations) and aggregating (function *agg* in equations) in periods of time calculated using equations 5.1, 5.2 and 5.3, where G is glucose, I represents insulin, and C carbohydrates:

$$mean(G,t,range) = \frac{\sum_{t\in range}(G_t)}{n}, \quad range \in [t_1, ..., t_n] \tag{5.1}$$

$$agg(I,t,range) = \sum_{t\in range}(I_t), \quad range \in [t_1, ..., t_n] \tag{5.2}$$

$$agg(C,t,range) = \sum_{t\in range}(C_t), \quad range \in [t_1, ..., t_n] \tag{5.3}$$

For each time step t, the set of features $F(t)$ (equation 5.4) that includes historical values ($F_{his}(G,t)$, $F_{his}(I,t)$, and $F_{his}(C,t)$), future values ($F_{fut}(I,t)$ and $F_{fut}(C,t)$), and LV of GV $F_{lv}(G,t,240)$, are generated:

$$\begin{gathered} F(t) = F_{his}(G,t) \cup F_{his}(I,t) \cup F_{his}(C,t) \cup \\ F_{fut}(I,t) \cup F_{fut}(C,t) \cup F_{lv}(G,t) \cup \{G, I, C\} \end{gathered} \tag{5.4}$$

$$\begin{aligned} F_{his}(G,t) = \{mean(G,t,[t_i,t_{+1}])\}, \\ F_{his}(I,t) = \{agg(I,t,[t_i,t_{+1}])\}, \\ F_{his}(C,t) = \{agg(C,t,[t_i,t_{+1}])\}, \\ t_i \in \{-240,-210,-180,-150,-120,-90,-75,-60,-45,-30,-15,0\} \end{aligned} \tag{5.5}$$

$$\begin{aligned} F_{fut}(I,t) &= \{agg(I,t,[t_i,t_+1])\}, \\ F_{fut}(C,t) &= \{agg(C,t,[t_i,t_+1])\}, \\ t_i &\in \{0,15,30,45,60,75,90,105,120\} \end{aligned} \tag{5.6}$$

$$\begin{aligned} F_{lv}(G,t,240) = \{&mean(G,t,240), auc(G,t,240), lauc(G,t,240), \\ &hauc(G,t,240), pstr(G,t,240), lpstr(G,t,240), \\ &hpstr(G,t,240), sd(G,t,240), cv(G,t,240), \\ &conga(G,t,240), ji(G,t,240), li(G,t,240), \\ &mage(G,t,240), ge(G,t,240), lmage(G,t,240), \\ &hmage(G,t,240), mv(G,t,240), adrr(G,t,240)\} \end{aligned} \tag{5.7}$$

All the selected parameters of the GP models are the same as those used in [110]. The best model of the ten repetitions is selected using the AIC [105] following the same process described in [110].

5.2 Experimental Results

5.2.1 Dataset

Data was collected from a group of ten patients with T1DM. Table 3.1 shows the number of days and observations of data for each patient, as well as the mean, SD and percentages of time where the patient has glucose levels below 70 mg/dl, above 250 mg/dl, and in the range [70,180] mg/dl (known as time in range). The glucose values of the patients present a high SD, which is usual in PwD due to the effects of the disease and the actions for diabetes management.

5.2.2 Configuration

In this study, four different GP variants were compared using PEG [106] to evaluate the accuracy of the predictions for each time horizon. Table 5.2 summaries the meaning of each algorithm. The parameters of each GP are the same as the parameters used in chapter 4 (see section 4.2.2). Furthermore, a deeper analysis for statistical significance was performed using a Kernel Density Estimation (KDE), the Nemenyi test [3], and the Bayesian model based on the Plackett-Luce distribution over rankings to analyze multiple algorithms in multiple problems [4]. The algorithms are named as follows:

- *GP*: It corresponds to previous approaches, where neither classified glucose values nor LV are used, and therefore no clustering algorithm is applied. Historical and future values are used as input variables. This is the baseline method.
- *CHAID-GP*: Profiled GP models are trained after a clustering phase. It uses the historical and future values, but it doesn't make use of the LV. This is the approach applied in [110].
- *LV-GP*: This algorithm uses the historical and future values and also the LV as input variables of the models (a contribution of this work).
- *LV-CHAID-GP*: Here, the historical and future values and the LV are considered as input variables for the models. Clustering with CHAID is also applied (a contribution of this work).

Table 5.2: Four different variants of the Genetic Programming approach were applied in this study. The differences among algorithms are the use or not of classified glucose values and the service or not of Latent Variables.

Algorithm Acronym	Acronym Meaning	Use of Classified Glucose Values	Use of Latent Variables
GP	*Genetic Programming*	No	No
CHAID-GP	*CHi-square Automatic Interaction* *Detection-Genetic Programming*	Yes	No
LV-GP	*Latent Variables-* *Genetic Programming*	No	Yes
LV-CHAID-GP	*Latent Variables-CHi-square Automatic* *Interaction Detection-Genetic Programming*	Yes	Yes

5.2.3 Results

A comparison between *GP* and the two algorithms *CHAID-GP* and *LV-CHAID-GP* is made based on the ratio of predictions in zone A+B for the different clusters of data. Figure 5.3 shows the results between *GP* and *LV-CHAID-GP* (see figure B.2 for results between *CHAID-GP* and *LV-CHAID-GP*). It plots the histograms obtained for each patient. The X-axis represents the eight cases corresponding to the eight prediction time horizons (from 30 to 240 minutes), and the Y-axis represents the clusters found using the clustering algorithm. The dark-gray segments indicate that the best prediction is made by *LV-CHAID-GP* and the white-gray that the best prediction is made by *GP*, the reference technique. For example, the algorithm has found 15 clusters for patient 1 (from MWSU01 to F1234), and 6 of the clusters have a better or at least comparable performance with *LV-CHAID-GP*. In 81.48% and 58.73% of the cases, models created with *CHAID-GP* and *LV-CHAID-GP* are more accurate (63.49% and 42.86%) or at least of equal accuracy (17.99% and 15.87%) than models created with *GP*. Moreover, in 16 and 26 cases, *CHAID-GP* and *LV-CHAID-GP* predictions are the best for all time horizons.

We also compared the ratio among the values obtained in zone A+B for the different time horizons (see figures B.1 and B.3). In 83.75% and 52.50% of the cases, models created with *CHAID-GP* and *LV-CHAID-GP* are more accurate (73.75% and 47.50%) or at least of equal accuracy (10.00% and 5.00%) than models created with *GP*. Further, in 2 and 5 cases, *CHAID-GP* and *LV-CHAID-GP* are the best for all clusters.

In general, both methods *CHAID-GP* and *LV-CHAID-GP* get better results than *GP*. It should be noted that there are more cases in *LV-CHAID-GP* than in *CHAID-GP* where the predictions are the best for all time horizons (26 vs. 16) and clusters (5 vs. 2).

Figures 5.4 to 5.6 show the PEG for all patients with all methods for three-time horizons: 30, 90 and 150 minutes (see figures B.5 to B.9 for 60, 120, 180, 210 and 240 minutes). The X-axis represents the reference values of glucose, and the Y-axis the prediction values, where Y=X is the ideal prediction. The figures are divided into five zones depending on the degree of accuracy of the glucose estimates (see section 4.1.5 and reference [106] for more detail). Each zone is identified with a different color. It should be noted that, as expected, when the time horizon increases, the points are more scattered, obtaining a greater number of points in undesirable zones (C, D, and E). This happens because as the time horizon increases, the predictions are more and more complex, and the accuracy of the models decreases. However, we can also see that *LV-CHAID-GP* is not giving predictions in the worst zone E (purple points) at all and only a few points in zone D (blue points), even for a 150-minute prediction horizon. It is also important to note that figures 5.4 to 5.6 include all the patients, which can give an idea of the importance of this result.

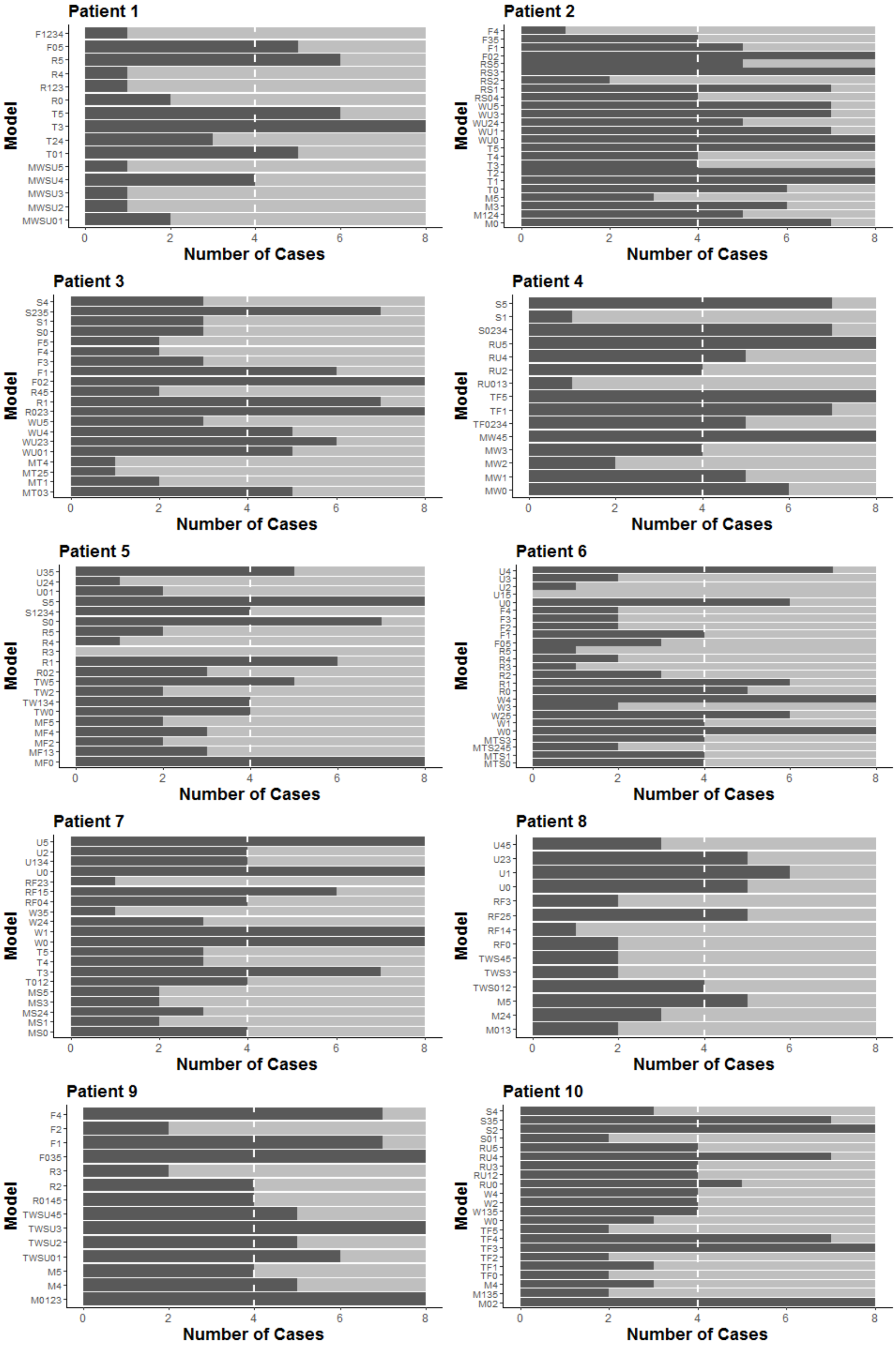


Figure 5.3: The ratio of models with better predictions (higher values in zone A+B) for different data clusters. Labels in the X-axis represent the number of cases. Each case represents a time horizon, and the label Y-axis identifies clusters found with the clustering algorithm according to the criteria day of the week and time slot. Dark-gray segments indicate that the best prediction is made by *LV-CHAID-GP* while light-gray means a better model from *GP*.

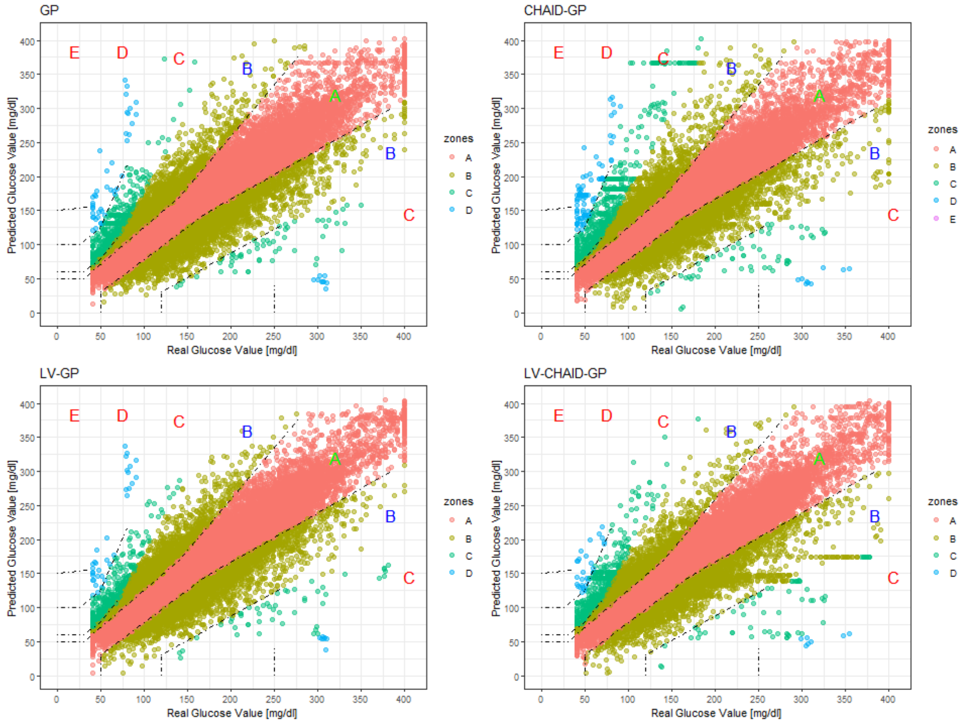


Figure 5.4: Parkes Error Grid from test data for all different models of all patients for fold-0 and time horizon of 30 minutes. The X-axis represents the reference values of glucose, and the Y-axis the values of the prediction. Each zone of the Parkes Error Grid is identified with a different color.

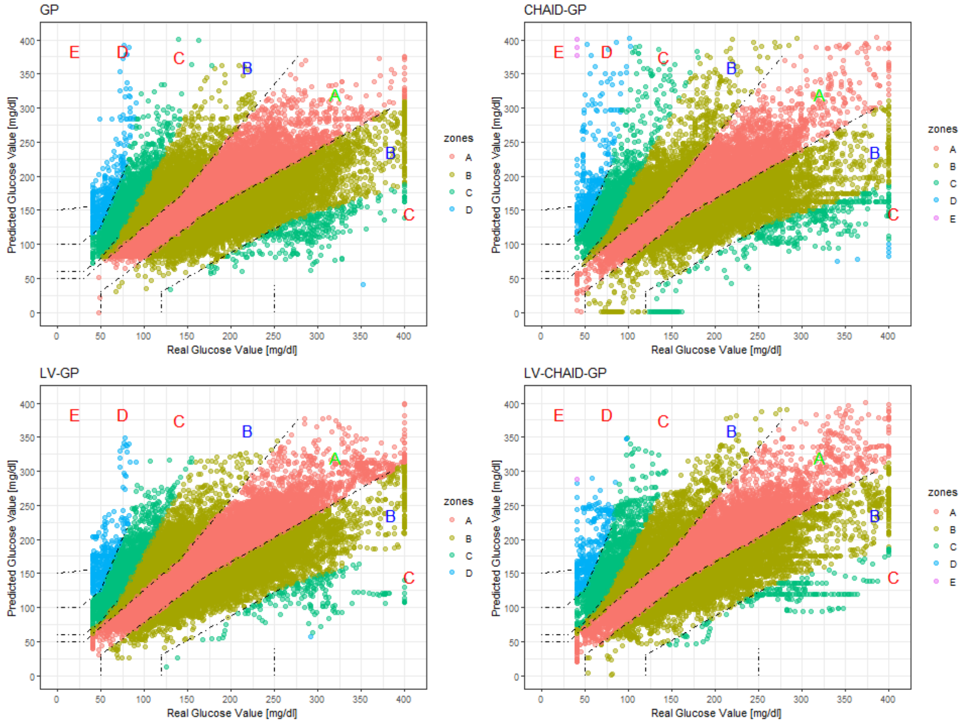


Figure 5.5: Parkes Error Grid from test data for all different models of all patients for fold-0 and time horizon of 90 minutes. The X-axis represents the reference values of glucose and the Y-axis the values of the prediction. Each zone of the Parkes Error Grid is identified with a different color.

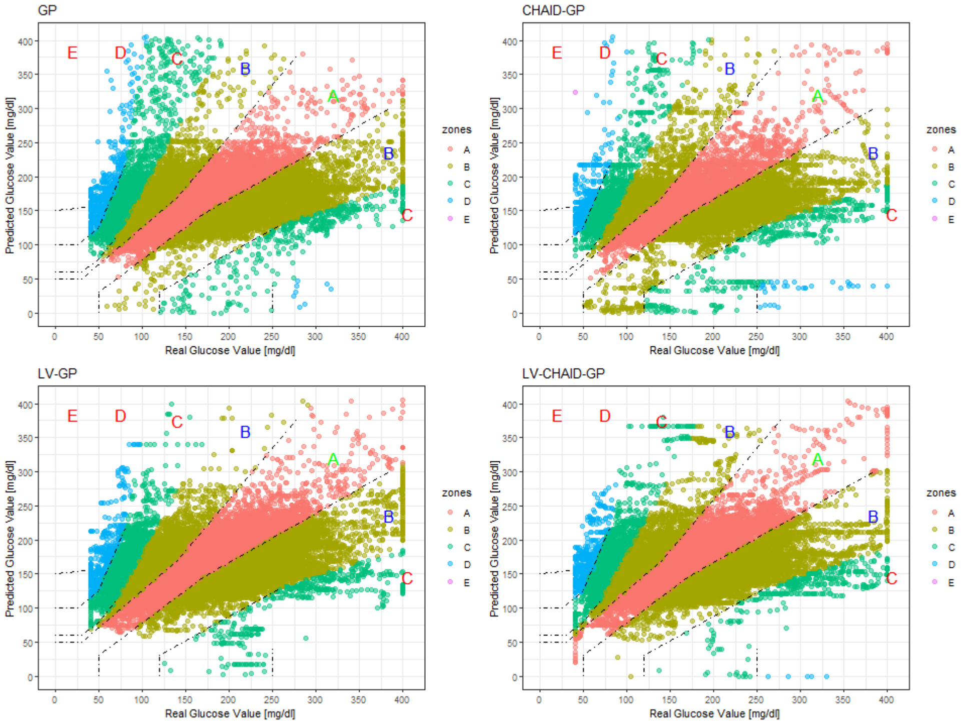


Figure 5.6: Parkes Error Grid from test data for all different models of all patients for fold-0 and time horizon of 150 minutes. The X-axis represents the reference values of glucose, and the Y-axis the values of the prediction. Each zone of the Parkes Error Grid is identified with a different color.

Table 5.3 shows the predictions (in percentage) in the evaluation phase obtained for the zone A+B with PEG for all time horizons (except 210 and 240 minutes included in table B.1) and patients. The accuracy of the models is analyzed under the assumption that the best model obtains a higher percentage of values in zone A+B and a lower percentage of values in zones C, D, and E.

In 65.00% of the cases, models created with *LV-GP* are more accurate, and in 53.33% of the cases, models made with *LV-CHAID-GP* are more accurate. Moreover, in 45.00% of the cases, *LV-GP* obtains the most accurate models, and in 31.67% of the cases, *CHAID-GP*. It is significant that for patients 7 and 9, all models created with *CHAID-GP* are better than those of models created with *GP*, and also for patient 6, all models created with *LV-CHAID-GP*. In general, for all patients and short-term horizons, *LV-GP*, *CHAID-GP* and *LV-CHAID-GP* have more accurate models than *GP*. We will analyze these results for statistical significance later.

Next, we analyze the variables that appear in the models to get an idea of what the models look like and the most significant variables to select for future studies. GP is computationally expensive, and, among other parameters, the number of input variables increases the computational cost. If we can locate the most important variables, we can create models similar to those created with the whole variables while being computationally less expensive. Figure 5.7 represents the 25 most significant relative importance variables out of a total of 83 variables defined with the equations in section 5.1.3 used to create the models for methods *LV-GP* and *LV-CHAID-GP* (see the whole variables in figure B.4). The importance is between 0 and 1, where 0 is a variable that never appears in the model and 1 is a variable that always appears in the model.

It should be noted that 68.00% of the cases are LGV features in *LV-CHAID-GP* and 32.00% in *LV-GP*. The MEAN, PSTR, and JI measures appear in the top ten positions. For the non-LV variables, glucose, insulin, and carbohydrates appear in the top five positions, which is not surprising since those have a more substantial influence on prediction problems for diabetes.

Table 5.3: Predictions (in percentage) obtained for the zone A+B with Parkes Error Grid for *GP* and *LV-GP* models, and the average percentage for *CHAID-GP* and *LV-CHAID-GP* models. *Green* values indicate better solutions compared to *GP*. The best result is highlighted in bold.

Model	t+30	t+60	t+90	t+120	t+150	t+180
Patient1						
GP	95.14	94.70	94.44	94.62	**92.54**	91.02
CHAID-GP	96.32	96.10	**95.28**	93.53	91.77	**92.17**
LV-GP	**99.49**	**97.41**	94.72	**93.87**	87.44	69.63
LV-CHAID-GP	98.04	96.06	92.74	90.76	80.82	78.08
Patient2						
GP	89.81	90.89	87.33	83.59	71.75	72.32
CHAID-GP	90.45	90.36	90.43	**89.77**	**91.03**	**91.37**
LV-GP	**99.35**	96.48	**94.87**	89.55	84.31	55.63
LV-CHAID-GP	99.00	**96.49**	91.52	88.59	84.29	79.56
Patient3						
GP	96.77	95.21	95.94	96.86	**98.08**	**97.66**
CHAID-GP	96.04	97.06	96.66	**96.95**	97.03	**97.66**
LV-GP	**99.81**	98.42	**97.70**	96.85	96.13	95.19
LV-CHAID-GP	99.69	**98.80**	96.40	93.84	92.45	91.20
Patient4						
GP	91.19	64.60	92.94	89.87	92.50	93.42
CHAID-GP	93.74	93.13	92.16	91.98	**92.64**	**94.06**
LV-GP	**98.85**	**96.61**	93.80	92.51	91.79	91.11
LV-CHAID-GP	98.34	95.82	**94.39**	**93.20**	92.17	90.78
Patient5						
GP	87.60	81.89	85.87	79.72	**89.28**	85.50
CHAID-GP	86.18	86.21	85.99	83.20	85.09	**86.03**
LV-GP	**97.25**	**92.41**	**88.99**	**86.48**	84.02	83.55
LV-CHAID-GP	95.28	89.18	84.02	77.44	75.43	72.92
Patient6						
GP	89.20	85.61	85.68	85.79	**88.62**	**91.57**
CHAID-GP	90.46	88.65	89.22	**88.11**	86.92	90.03
LV-GP	**97.70**	**93.44**	**90.38**	80.46	83.96	76.60
LV-CHAID-GP	96.24	90.78	85.79	82.85	80.39	78.95
Patient7						
GP	89.56	87.97	89.52	87.59	91.29	91.24
CHAID-GP	91.22	90.57	92.76	91.96	**91.44**	**92.60**
LV-GP	**99.35**	**96.57**	**94.04**	**92.34**	82.72	64.47
LV-CHAID-GP	97.82	93.90	88.42	84.81	85.45	85.36
Patient8						
GP	88.95	92.94	92.38	**94.05**	**93.41**	86.33
CHAID-GP	91.27	92.50	**93.84**	90.55	92.69	**88.84**
LV-GP	**98.23**	**95.10**	93.01	83.28	83.19	54.40
LV-CHAID-GP	88.77	91.09	89.80	89.47	90.47	88.17
Patient9						
GP	87.23	86.45	85.27	82.40	79.73	90.03
CHAID-GP	91.55	90.81	90.03	85.51	**92.17**	**90.55**
LV-GP	**97.91**	90.81	88.19	85.94	84.09	81.39
LV-CHAID-GP	96.64	**91.89**	**91.83**	**89.08**	83.82	84.54
Patient10						
GP	89.39	80.98	80.87	78.53	84.79	86.77
CHAID-GP	86.91	84.08	85.71	83.65	**86.66**	**88.06**
LV-GP	**97.28**	**92.84**	**87.87**	**84.70**	83.28	69.93
LV-CHAID-GP	97.17	92.52	85.65	80.68	76.51	73.79

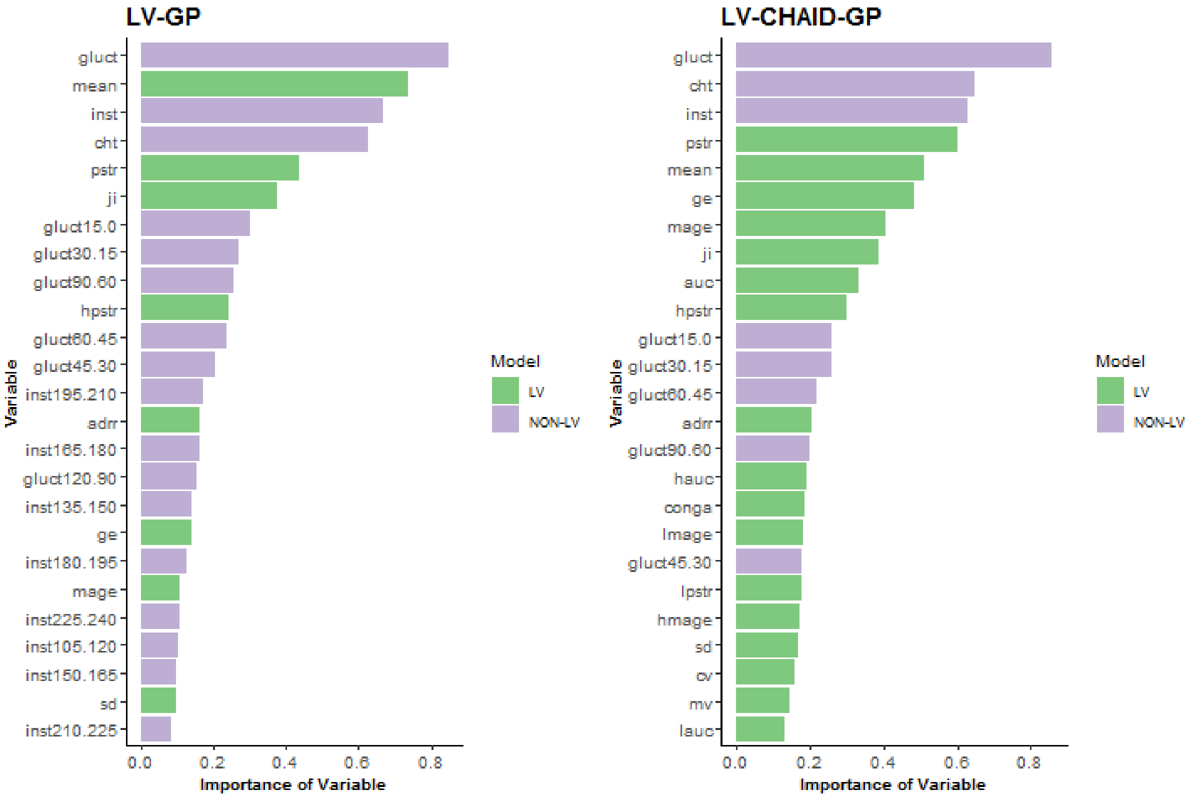


Figure 5.7: The relative importance of the first 25 variables that make up the models based on *LV-GP* and *LV-CHAID-GP*. The importance is between 0 and 1, where 0 is a variable that never appears in the model and 1 is a variable that always appears in the model. Latent Variables are represented in green, and historical and future variables in purple.

We apply a statistical assessment of the results. First, to select the suitable test, we analyze the distribution of the results, using a KDE of the distribution of the samples. The objective is to visualize if the data meets the conditions for a parametric test. Then, looking for statistically significant differences among results, we compare the performance of the algorithms proposed in this work. Figure 5.8 shows the distributions of the results for all methods and time horizons. Each algorithm is plotted with a different color. As we can see, the data is not distributed according to a Gaussian distribution, nor is the variance the same for all methods. Data distribution is multi-modal, and a non-parametric test is necessary. All the plots were obtained with [2].

Next, we select the Nemenyi test [3] to apply an all pairwise comparison to the results. We selected this test because it is a non-parametric test and has the advantage of having an associated plot to represent the comparison results. This test is based on the absolute difference of the average rankings of the predictors. For a significant level $\alpha = 0.05$, the test determines the Critical Difference (CD). If the difference between the average ranking of two methods is greater than CD, then the null hypothesis that the methods have the same performance is rejected. Figure 5.9 shows the graphical comparison of the CD for all methods and time horizons. Each algorithm is placed on an axis according to its average ranking. Then, those algorithms that show no significant difference are grouped using a horizontal line. The plot also shows the size of the CD required for considering two algorithms as significantly different. Results show significant differences for all time horizons (except for 150 minutes).

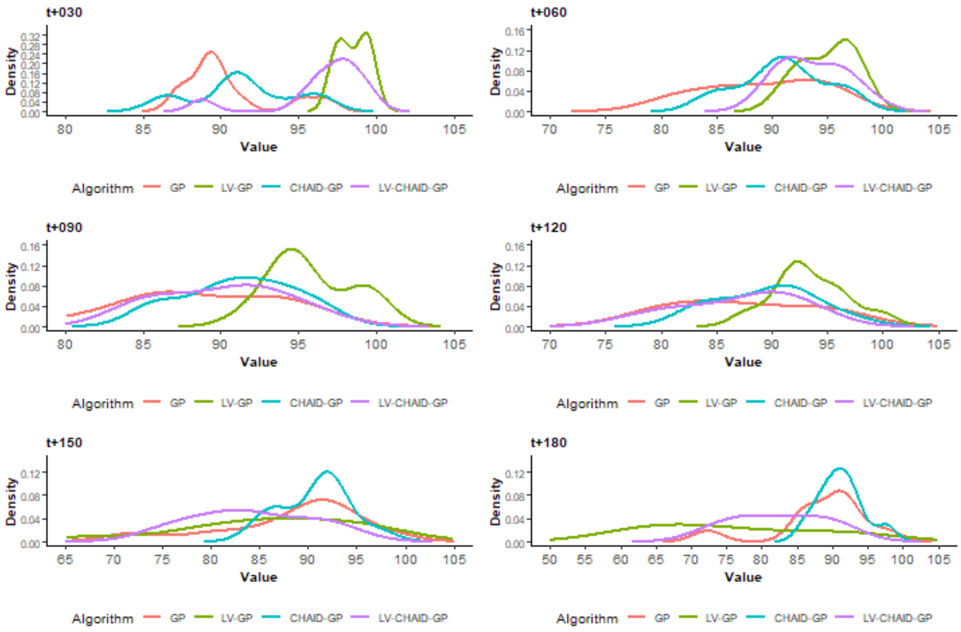


Figure 5.8: Density plots of the distribution of the results for all methods and time horizons (except 210 and 240 minutes). Each algorithm is plotted with a different color. The distributions are multi-modal. All the plots were obtained with [2].

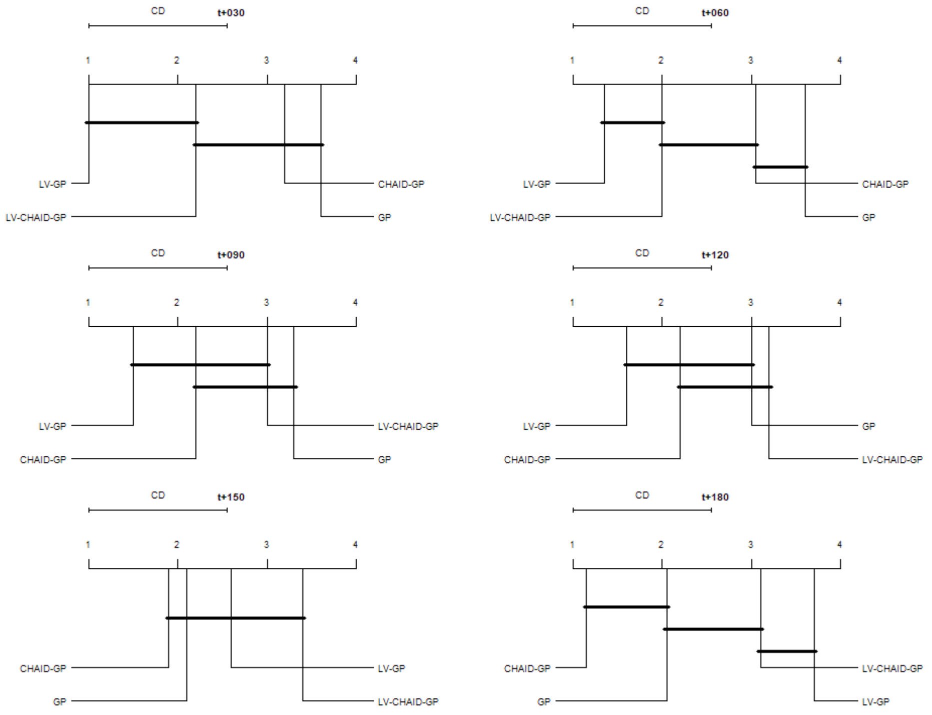


Figure 5.9: Nemenyi test for all methods and time horizons (except 210 and 240 minutes) using the graphical representation of [3]. Each algorithm is placed on an axis according to its average ranking. Those algorithms that show no significant difference are grouped using a horizontal line. The plot also shows the size of the Critical Difference required for considering two algorithms as significantly different.

Finally, to make a deeper statistical study of the results, we follow the Bayesian model of [4, 111] based on the Plackett-Luce distribution over rankings to analyze multiple methods in multiple time horizons. Figure 5.10 shows the Bayesian test for all methods and time horizons. The X-axis represents the probability of winning, with values between 0 and 1, where 0 is a solution with 0% of the likelihood of winning and 1 is a solution with 100% of the probability of winning. Y-axis represents each of the algorithms. The dot in the graph is the average probability of winning, and the horizontal line represents the size of the confidence interval of each test, ending in vertical lines. Figure 5.10 shows that *LV-GP* has the highest probability of being the best for 30, 60, 90, and 120 minutes, and for 30 minutes, there are significant differences between this method and the rest of the methods. For 150 and 180 minutes, *CHAID-GP* is the method with the highest probability of being the best. For 180 minutes, there are significant differences between this method and the rest of the methods.

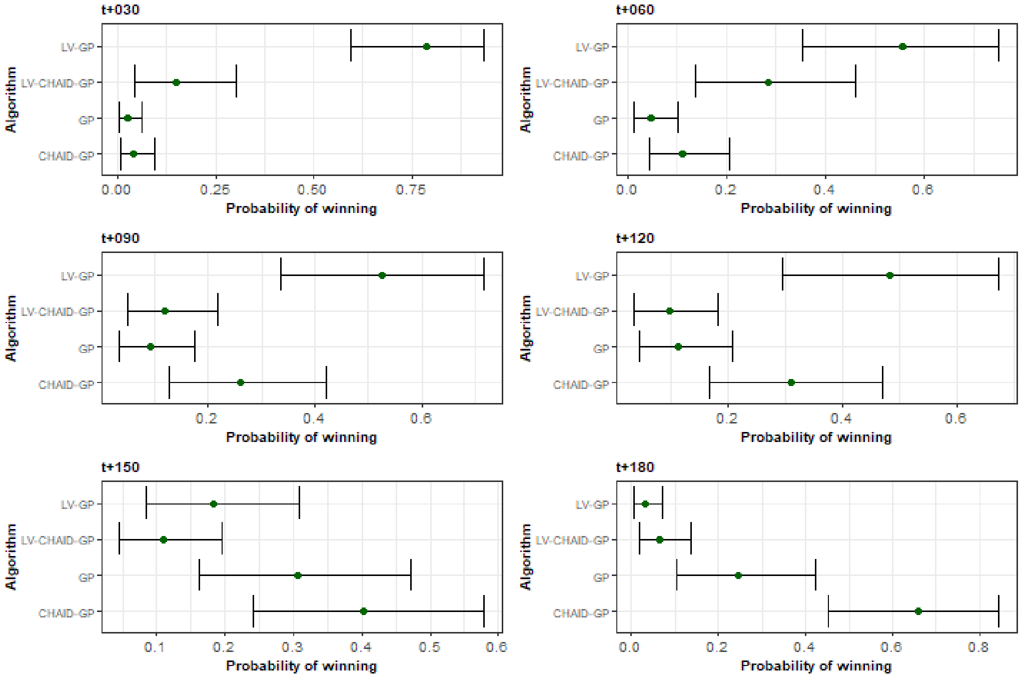


Figure 5.10: Bayesian model of [4] to analyze all methods and time horizons (except 210 and 240 minutes). The X-axis represents the probability of winning, with values between 0 and 1, where 0 is a solution with 0% of the probability of winning and 1 is a solution with 100% of the probability of winning. Y-axis represents each of the algorithms.

5.3 Conclusions

In this work, we propose the use of several GV measures as additional features for GP modeling of glucose levels in PwD. This work is an extended version of the work presented in chapter 4. In particular, we tested our proposal with a real dataset from ten patients from a Spanish Hospital that was also used in previous works. Four different GP variants were compared using PEG, two correspondings with earlier results (*GP* and *CHAID-GP*), and two new proposed in this work (*LV-GP* and *LV-CHAID-GP*) that make use of a set of LGV features. We explored models for eight different time predictions horizons, from 30 to 240 minutes, by steps of 30 minutes.

The main conclusions of this work can be summarized as follows:

- GV can be incorporated by generating models for different patterns of glucose profiles or by including

LV.

- Definitely, taking into account GV is important for developing good prediction models.
- Models created with LV improve the quality predictions and do not produce forecasts in the worst zone (E) and only a few points in the second-worst zone (D), even for a 150-minute prediction horizon.
- *CHAID-GP* and *LV-CHAID-GP* predictions are the best for all time horizons.
- *CHAID-GP* and *LV-CHAID-GP* are the best for all patients.
- In both cases, dangerous predictions are reduced concerning previous works.
- In general, for all patients and short-term horizons, *LV-GP*, *CHAID-GP* and *LV-CHAID-GP* have more accurate models than *GP*.
- The analysis of the relative importance of the variables reveals that MEAN, PSTR and JI measures are in the top-ranking positions.
- The non-LV features, i.e., glucose, insulin, and carbohydrates, appear in the top five positions of influence. These results also indicate the correctness and the coherence of the obtained models, which was expected.
- The statistical analysis was performed with a novel approach based on a Bayesian model and the Plackett-Luce distribution over rankings. It reveals that *LV-GP* (one of the approaches proposed in this work) has the highest probability of being the best for 30, 60, 90, and 120 minutes. For 150 and 180 minutes, *CHAID-GP* is the method with the highest probability of being the best.

Chapter 6

Glucose Prediction using Multi-objective GE

In this chapter, we investigate the benefits of applying a multi-objective approach for solving a SR problem utilizing GE. In particular, we extend previous work, obtaining mathematical expressions to model glucose levels in the blood of PwD. Here we use a multi-objective GE approach based on the NSGA-II, considering the RMSE and an ad-hoc fitness function as objectives. This ad-hoc function is based on the CEG analysis, which helps show the potential danger of mispredictions in PwD. In this work, we use two datasets to analyze two different scenarios: *What-if* and *Agnostic*, the most common in daily clinical practice.

In particular, we tackle the following research questions: is the multi-objective approach able to improve single-objective predictions? Does the multi-objective approach obtain clinically better solutions? Is it always better to use more information from the past, or can we get equivalent predictions with information from the last hour instead of the previous two?

As will be seen, the development of a problem-specific fitness function does not only considerably improve the quality and the robustness of the GE algorithm as a SR tool in the *What-if* scenario, as we show in [112] but also allows us to obtain good *Agnostic* prediction models using only information from the last hour previous to the time of prediction.

The best single-objective solutions may suffer from a significant number of erroneous predictions that could lead to incorrect treatments that may be dangerous to the patient's health. Later, we will show that the multi-objective strategy improves this situation by reducing the number of wrong predictions and the severity of incorrect treatments.

The rest of the chapter is organized as follows. In section 6.1, we explain the multi-objective approach. Section 6.2 gives details about the dataset and discusses the results, comparing it with previous work and analyzing the contributions of the multi-objective approach. We finish the chapter with the conclusions in section 6.3.

6.1 Methodology

The work that we present here extends those works based on GE and explores the use of a multi-objective approach, based on the NSGA-II [61], applied to actual data from PwD. We are interested not only in the problem-specific but also in the performance of the multi-objective implementation. We integrate the different possibilities of using GE for short-term and medium-term glucose prediction in PwD, for the *Agnostic* and *What-if* scenarios, and prediction horizons of 30, 60, 90, and 120 minutes, using information from 60 and 120 minutes before the time of prediction.

The complete workflow can be seen in figure 6.1. Actual data is gathered using different electronic devices such as CGM, smart bands, and CSII (insulin pump), as well as annotations by the patients. These data are processed to form the different datasets (combining two historical windows and four prediction horizons) used in the two scenarios studied in this research. In sections 6.1.1 and 6.1.2 we thoroughly explain the two scenarios. Still, we can summarize that in the *What-if* case, the model predicts future glucose values by taking into account not only past values of the three variables (glucose, carbohydrates, and insulin) but also future carbohydrate intake and insulin injections from the present until the prediction horizon. In the *Agnostic* scenario, the model predicts future glucose values based only on past values. For both scenarios, we train and test, using cross-validation, a GE model and a multi-objective GE model.

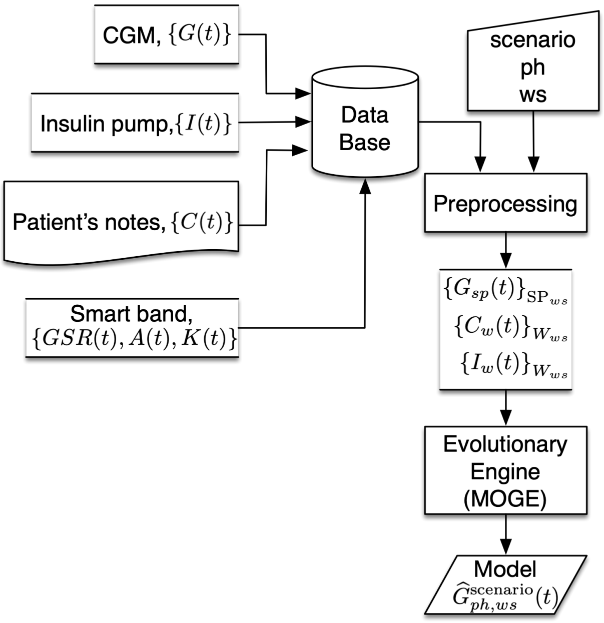

CGM, $\{G(t)\}$
Insulin pump,$\{I(t)\}$
Patient's notes, $\{C(t)\}$
Smart band, $\{GSR(t), A(t), K(t)\}$
Data Base
scenario
ph
ws
Preprocessing
$\{G_{sp}(t)\}_{\mathrm{SP}_{ws}}$
$\{C_w(t)\}_{W_{ws}}$
$\{I_w(t)\}_{W_{ws}}$
Evolutionary Engine (MOGE)
Model
$\widehat{G}^{\text{scenario}}_{ph,ws}(t)$

Figure 6.1: Workflow illustrating the generation of models for the glucose prediction problem.

One of the advantages of using GE as a modeling approach is that we can tackle the different possibilities mentioned in the paragraph above using the same engine and only making changes in the grammar for each prediction horizon, scenario, and information configuration. Each variable, feature, or attribute used in the model can be easily included in the grammar. With this procedure, we can study each feature or group of features to the quality of the solution.

6.1.1 What-if Scenario or Insulin-Carbohydrate Recommendation

As mentioned in the introduction, the *What-if* scenario models may have access to input events, so we can predict the BG level after m minutes, supposing that the patient eats a meal with C grams of carbohydrates and I units of insulin are injected in t minutes from the time of prediction. In other words, our objective is to construct an insulin-carbohydrate recommendation tool. Thus, we look for predictive models that help us in evaluating those recommendations. In this scenario, models can use the following data:

- IG using a Medtronic CGM device that gives us observations every five minutes.
- Notes of estimated carbohydrate units ingested, taken by each patient.
- Insulin injected using an insulin infuser device from Medtronic, which registers injections of both basal and bolus insulin every five minutes.

Once all the information has been collected, we imputate the data using cubic splines and match all the events to the closest timestamp to construct a set of matched time series corresponding to glucose, insulin, and carbohydrate values. We also process the set of features available at the time of modeling.

Using the cubic spline technique in time series for imputation is the most widely used strategy in medical literature, and physical science research [113, 114].

At each time point, t, data from the previous two hours are available for prediction. We process these data to define a set of new features as we did in [16]. To this aim, we first define two sets: the set of time windows in which we evaluate new features, $W_{120} = \{0-0, 0-30, 31-60, 61-90, 91-120\}$ minutes, and the set of sample times previous to the current prediction time, $\mathrm{SP}_{120} = \{0, 15, 30, 45, 60, 75, 90, 105, 120\}$ minutes. The new features[1] are:

- The set of glucose values measured sp minutes previous to prediction time: $\{G_{sp}(t)\}_{\mathrm{SP}_{120}} = \{G(t-sp)\}_{sp\in\mathrm{SP}_{120}}$,
- The set of the sums (in grams) of the carbohydrates ingested in window w: $\{C_w(t)\}_{W_{120}} = \{\sum_{i\in w} C(t-i)\}_{w\in W_{120}}$,
- The set of the sums of the units of insulin injected in window w: $\{I_w(t)\}_{W_{120}} = \{\sum_{i\in w} I(t-i)\}_{w\in W_{120}}$.

Notice that $G_0(t) = G(t)$, $C_{0-0}(t) = C(t)$, and $I_{0-0}(t) = I(t)$ are the actual values at prediction time for glucose, carbohydrates, and insulin, respectively. We also define the set of prediction horizons, i.e., the sample times in future to predict the BG, $\mathrm{PH} = \{30, 60, 90, 120\}$ minutes, so that $\widehat{G}_{ph}(t)$ with $ph \in PH$ is the predicted BG ph minutes ahead in time, whereas $I_{ph}(t)$ and $C_{ph}(t)$ are the values of insulin and carbohydrates ph minutes ahead in time. Hence, our prediction problem can be stated as finding an expression for the predicted BG level, $\widehat{G}_{sp}(t)$, given by equation (6.1), that minimizes the objective functions RMSE and F_{CLARKE}.

$$\widehat{G}^{\mathrm{What-if}}_{ph,120}(t) = f_{t,ph}(\{G_{sp}(t)\}_{\mathrm{SP}_{120}}, \{I_w(t)\}_{W_{120}}, \{C_w(t)\}_{W_{120}}, I_{ph}(t), C_{ph}(t)) \tag{6.1}$$

6.1.2 Agnostic Scenario or Glucose Prediction in Absence of Event Information

Recently, with the appearance of new smart devices, incorporating more predictive variables into models has been raised to improve the accuracy of the models. For example, the activity bracelets on the market provide accurate information on many variables, including exercise, sleep, heart rate, body temperature, caloric consumption, and more. When all this information is incorporated into the dataset, it becomes much more complicated to make *What-if* models since the number of possible combinations of the variables and assumptions involved in an event is enormous, and its usefulness would be minimal. However,

[1] When constructing prediction models that help in the recommendation, we can use variables (features) that include the information involved in the recommendation process and thus be able to use them effectively. This does not mean that we use information from the future.

it is possible to produce *Agnostic* models in which there is no access to information on future events in the prediction phase. These types of models need to predict those events implicitly. For example, the models must identify fasting periods or physical exercise.

Our GE-based model generation engine allows us to make these types of models without substantial changes to our methodology. It only requires adjusting the grammars to reflect access only to the variables available. In the case of equation (6.1), we only need to eliminate access to future values, i.e., $C_{t+H}(t)$ and $I_{t+H}(t)$, so the general equation for the models would be expressed by equation (6.2).

$$\widehat{G}^{\text{Agnos}}_{ph,120}(t) = f_{t,ph}(\{G_{sp}(t)\}_{\text{SP}_{120}}, \{I_w(t)\}_{W_{120}}, \{C_w(t)\}_{W_{120}}) \tag{6.2}$$

In this work, we solve the *Agnostic* problem in relationship with the data described in [115]. This dataset includes information on BG levels from a CGM, insulin doses from an insulin pump, self-reported information, and data from an activity band. More details can be found in the previous work. We similarly processed the data to that described in section 6.1.1 to obtain aggregated features with information every 15 minutes, including BG level, insulin doses, carbohydrate estimates, GSR, skin temperature, air temperature, and acceleration. The influence of these variables in PwD is documented in [116, 117, 118]. However, the analysis of their particular impact is left for future collaborations with medical staff since it is beyond the scope of the multi-objective analysis made in this work.

Although the *Agnostic* scenario may be considered a subset of the *What-if* scenario, in this work, we consider them as different problems since we use additional data.

In addition to the analysis, we compare two options regarding the information available at the time of prediction. On the one hand, we use the events from the previous two hours to construct the models as expressed in equation (6.2). On the other hand, we use only the last hour before the time of prediction. As we will see in the experimental section, this comparison is helpful for the applicability of the models since most of the time, the patient has little or no information from the past. Incorporating the information of the variables obtained from the smart devices, we will search for models by applying (6.3) with window size, $ws \in \{60, 120\}$ minutes.

$$\begin{aligned}\widehat{G}^{\text{Agnos+}}_{ph,ws}(t) = f_{t,ph}(&\{G_{sp}(t)\}_{\text{SP}_{ws}}, \{I_w(t)\}_{W_{ws}}, \{C_w(t)\}_{W_{ws}}, \\ &\{\text{GSR}_{sp}(t)\}_{\text{SP}_{ws}}, \{A_{sp}(t)\}_{\text{SP}_{ws}}, \{K_{sp}(t)\}_{\text{SP}_{ws}})\end{aligned} \tag{6.3}$$

Where we incorporate the variables GSR, skin temperature, and magnitude of acceleration to those defined in equation (6.1):

- The set of GSR values sp minutes before t: $\{\text{GSR}_{sp}(t)\}_{\text{SP}_{ws}} = \{\text{GSR}(t - sp)\}_{sp \in \text{SP}_{ws}}$,
- The set of skin temperatures measured sp minutes before t: $\{\text{K}_{sp}(t)\}_{\text{SP}_{ws}} = \{\text{K}(t - sp)\}_{sp \in \text{SP}_{ws}}$,
- The set of accelerations measured sp minutes before t: $\{\text{A}_{sp}(t)\}_{\text{SP}_{ws}} = \{\text{A}(t - sp)\}_{sp \in \text{SP}_{ws}}$.

6.1.3 Fitness Functions

Again, GE allows us to use the same configuration for both scenarios and both algorithms: single-objective (GE) and multi-objective (MO-GE). GE relies on RMSE to guide the search. RMSE is a common fitness function when adjusting data in SR problems. Equation (6.4) shows its definition.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(G(t_i + ph) - \widehat{G}_{ph,ws}(t_i))^2} \tag{6.4}$$

In the case of MO-GE, we use a second objective function, denoted as F_{CLARKE}, which was defined in [119]. This function follows the CEG criterion used to test the clinical significance of differences between a glucose measurement technique and venous BG reference measurements [57]. F_{CLARKE} is defined in equation (6.5).

$$F_{\text{CLARKE}} = \sum_{i=1}^{N} w_i \tag{6.5}$$

To compute w_i we used the next equation:

$$w_i = \begin{cases} 100 & \text{if } (\widehat{G}_{ph,ws}(t_i) \geq 180) \wedge (G(t_i + ph) \leq 70) \\ 100 & \text{if } (\widehat{G}_{ph,ws}(t_i) \leq 70) \wedge (G(t_i + ph) \geq 180) \\ 10 & \text{if } (180 \geq \widehat{G}_{ph,ws}(t_i) \geq 70) \wedge (G(t_i + ph) \geq 240) \\ 10 & \text{if } ((180 \geq \widehat{G}_{ph,ws}(t_i) \geq 70) \wedge (G(t_i + ph) \leq \frac{175}{3}))\vee \\ & ((\widehat{G}_{ph,ws}(t_i) \geq \frac{6}{5} \times G(t_i + ph)) \wedge (70 \geq G(t_i + ph) \geq \frac{175}{3})) \\ 1 & \text{if } (\widehat{G}_{ph,ws}(t_i) \geq G(t_i + ph) + 110) \wedge (290 \geq G(t_i + ph) \geq 70) \\ 1 & \text{if } (\widehat{G}_{ph,ws}(t_i) \leq \frac{7}{5} \times G(t_i + ph) - 182) \wedge (180 \geq G(t_i + ph) \geq 130) \end{cases} \tag{6.6}$$

In this regard, CEG considers a grid divided into five zones (A to E) depending on the severity of the misprediction. The values that fall within zones A and B are clinically exact and/or acceptable, and thus the clinical treatment will be correct. We consider A and B as a single category with no contribution to equation (6.5). Values in zone C can be dangerous in some situations. Although less dangerous than D and E zones, we should also try to minimize predictions in these zones, so predicted points in these zones contribute a value of 1 to equation (6.5). Finally, zones D and E represent potentially dangerous areas since the prediction is far from acceptable, and the suggested treatment will be different from the correct treatment. Each prediction in zone D adds 10 to equation (6.5), while predictions in zone E add 100. The inequalities and conditions of equation (6.6) delimit these zones according to [57].

RMSE is a standard linear regression metric that measures the raw quality of a model. Intuitively, it can be expected that the greater the RMSE, the greater the F_{CLARKE}. Nevertheless, a suitable model in terms of RMSE can, at the same time, be dangerous from a medical point of view. For instance, an error of 20 mg/dl for an expected value of 50 mg/dl is worse than the same error for 100 mg/dl since, in the first case, the patient will be in a hypoglycemic situation. In contrast, in the second case, the patient is within the normal glucose range. These kinds of problems are not well identified by RMSE. On the contrary, F_{CLARKE} can amplify the effect of those dangerous situations. Similarly, a suitable model in terms of F_{CLARKE}, with all predictions in zones A and B, could reach significant values of RMSE. Therefore, these objective functions are not entirely orthogonal, but they help identify good models in terms of both raw and medical qualities.

Previous experimentation showed poor RMSE results of F_{CLARKE} in single-objective (GE). Therefore, for the sake of space and the clarity of the discussion, we have discarded these experiments.

6.1.4 Multi-Objective Grammatical Evolution

In this work, we propose a multi-objective approach to GE [120]. We use the same process and grammars of [121], where the interested reader can find more details about applying GE for the creation of models in this scenario. As is well known, the GE method is powered by an evolutionary computation algorithm, usually an adapted implementation of a GA or a Particle Swarm Optimization (PSO) algorithm. There exist some other multi-objective implementations such as [122]. However, we use our library, which is publicly available through GitHub.

Here we also use a bi-objective approach, using equations (6.4) and (6.5) as fitness functions. As an evolutionary engine, we apply NSGA-II, which is perhaps the most effective way of optimizing and searching for solutions to multi-objective problems with evolutionary computation when dealing with two or three objectives. One of the important issues for selecting a multi-objective approach is to study the fitness functions, i.e., the objectives. Although the objectives may work in the same direction, it is not desirable that both functions have similar features. Fitness functions should guide the algorithm through the search space differently, although with a common search. This is the case for equations (6.4) and (6.5), where both try to improve the quality of the solution by simultaneously minimizing error and obtaining solutions with 100% of the predictions in zones A and B.

Figure 6.2 represents an extract of the grammar in Backus-Naur Form (BNF) format designed to find a predictive model for future BG levels. This is a typical grammar for SR adapted to the variables and features for each dataset. It is a recursive grammar, and the operators are reduced to addition, subtraction, and multiplication, based on the conclusions of [123] and our previous experimental experience. Despite the size of the search space being infinite because it is a recursive grammar, the algorithm can efficiently explore the search space.

```
func> ::= <expr>

<expr> ::= (<expr> <op> <expr>) | (<cte> <op> <expr>) | <var>

<var> ::= <varch> | <varins> | <vargl>

<op> ::= + | - | *

# Glucose
<vargl> ::= G_{t-120}(t) | G_{t-105}(t) | G_{t-90}(t) | G_{t-75}(t) | G_{t-60}(t) |
            G_{t-45}(t) | G_{t-30}(t) | G_{t-15}(t) | G(t)

# Carbohydrates
<varch> ::= C_{t-120}(t) | C_{t-90}(t) | C_{t-60}(t) | C_{t-30}(t) | C(t) | C_{t+H}(t)

# Insulin
<varins> ::= I_{t-120}(t) | I_{t-90}(t) | I_{t-60}(t) | I_{t-30}(t) | I(t) | I_{t+H}(t)

<cte> ::= <factor> * <digit>
<factor> ::= 1 | 0.1 | 0.01 | 0.001 | 0.0001
<digit> ::= 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10
```

Figure 6.2: Grammar for Blood Glucose prediction.

Notice that grammar is the tool that allows knowledge of the problem to be included in the optimization process. For instance, the grammar proposed sets the particular glucose, carbohydrate, and insulin variables to be used and the operands and precision factors available to produce model expressions. This way, the search is guided to a particular part of the solutions space.

6.2 Experimental Results

We carry out two types of experiments testing the two scenarios explained in sections 6.1.1 and 6.1.2: the *What-if* and *Agnostic* scenarios. Hence, we use two different datasets containing the information and variables required by each scenario.

To find the models that will later be tested, we divided the data into two sets, training and test, using the k-fold cross-validation technique. This technique generally results in a less biased estimate of the model's performance than a simple train/test division. After shuffling the data, the total dataset is divided into k subsets (in this article, k = 10). Over k iterations, the models are tested on one subset and trained on the other (k-1) subsets. The final results are the union of the k iterations.

We could expect that the variance across these small subsets would contribute a biased estimation, but, in the literature, we can find research that does not support this idea [124]. It is also interesting to note that here each point (value to be predicted) consists of a BG value plus a historical window with glucose, carbohydrate, and insulin data. This experimental setup is illustrated in figure 6.3.

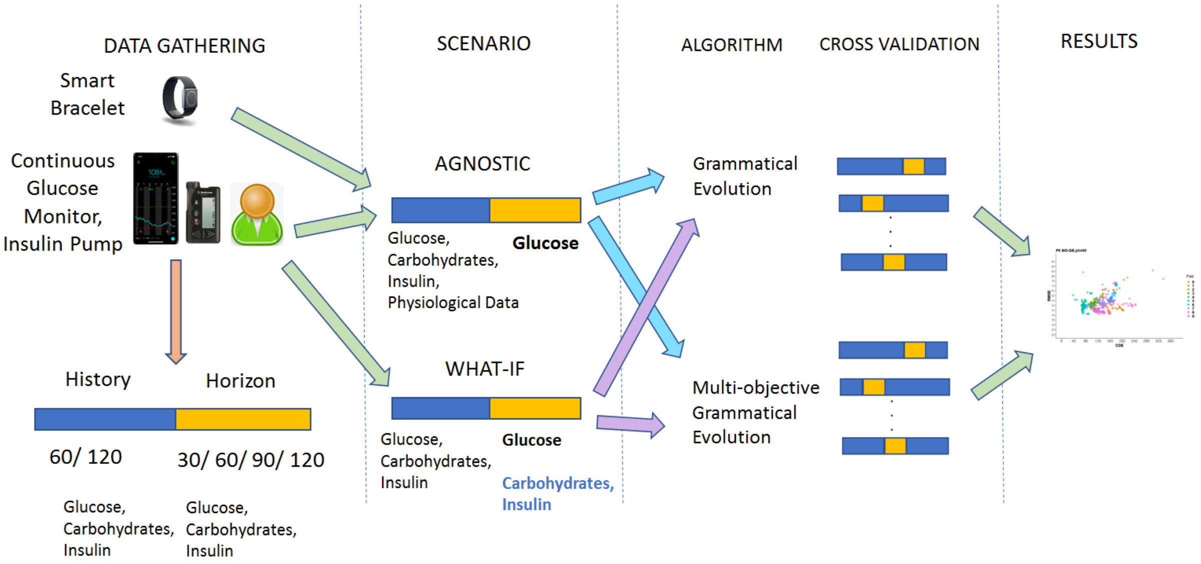

DATA GATHERING
SCENARIO
ALGORITHM
CROSS VALIDATION
RESULTS
Smart Bracelet
Continuous Glucose Monitor, Insulin Pump
AGNOSTIC
Glucose, Carbohydrates, Insulin, Physiological Data
Glucose
WHAT-IF
Glucose, Carbohydrates, Insulin
Glucose
Carbohydrates, Insulin
Grammatical Evolution
Multi-objective Grammatical Evolution
History
Horizon
60/ 120
30/ 60/ 90/ 120
Glucose, Carbohydrates, Insulin
Glucose, Carbohydrates, Insulin

Figure 6.3: Workflow illustrating the prediction problem setup.

We want to emphasize that obtaining data from real patients is difficult because this is sensitive information that usually requires special authorization from both the patients and the medical authorities. In addition, patients must be committed to the research for data completion. It is important to wear the devices correctly while taking notes and registering any unexpected event affecting the data. Besides, it is usual to discard part of the data collected because of mistakes made by patients. Therefore, studies usually deal with small datasets, as shown in [125, 126, 127, 128], with 10, 11, 7 and 7 T1DM patients, respectively, with different gender distributions.

6.2.1 Datasets

Ten T1DM patients have been selected for the *What-if* scenario, based on conditions of reasonable glucose control. To be included in the study, three primary needs must be met: (1) at least one year since T1DM diagnosis; (2) absence of diagnosis of major psychiatric disorder; (3) no severe breakthrough disease in the last six months. The Ethics Committee approved the study of the Hospital Príncipe de Asturias in Alcalá de Henares, Spain, and all patients signed a prior informed consent. Data acquisition was carried out by two nurses from the Endocrinology and Nutrition Service at the hospital.

Data from patients were acquired over multiple weeks using the devices described in section 6.1. Log entries were stored in five-minute intervals. In this dataset, we have at least ten full days of data for each patient. These days are not necessarily consecutive. Each log entry contains the date and time, the BG value, the amount of insulin (injected via insulin pump), and the number of carbohydrate intake estimated by the patients. The population characterization is female (80%), average age 42.30±11.07, years of disease 27.20±10.32, years with insulin pump therapy 10.00±4.98, weight 64.78±13.31 kg and HbA1c average of 7.27±0.50%. The average number of days with data is 44.80±30.73. Figure 6.4 describes the glucose levels of patients from this dataset. The upper figure shows the percentage of time the patient has a very low glucose level (<54 mg/dl in dark-red), low ([54,70) mg/dl in red), in range ([70,180] mg/dl in green), high ((180,250] mg/dl in yellow), and very high (>250 mg/dl in dark-yellow). The numbers (from top to bottom) represent the percentage of time the patient has a glucose level >250 mg/dl, in range [70,180] mg/dl and <70 mg/dl. The lower figure shows the IQR of glucose, where the mean values are represented with red dots. This is a common way of evaluating the quality of BG management in PwD. The greater the time in range ([70-180] mg/dl), the better.

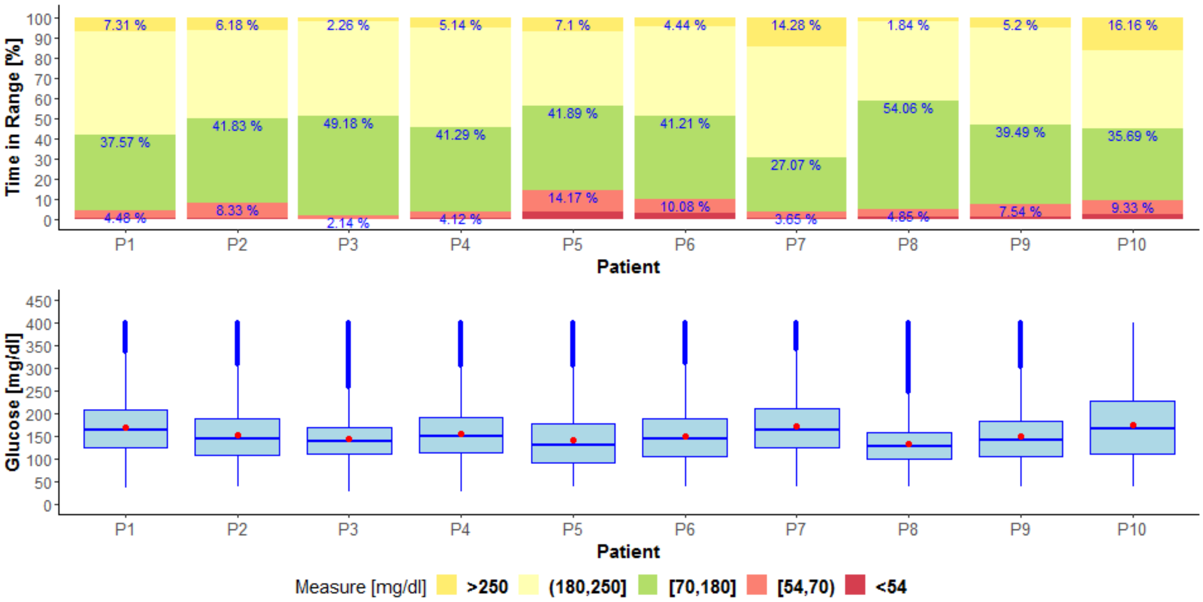


Figure 6.4: Histograms and boxplots describe the glucose level of the patients from the dataset used in the *What-if* scenario. The upper figure shows the percentage of time the patient has a very low glucose level (<54 mg/dl in dark-red), low ([54,70) mg/dl in red), in range ([70,180] mg/dl in green), high ((180,250] mg/dl in yellow), and very high (>250 mg/dl in dark-yellow). The numbers (from top to bottom) represent the percentage of time the patient has a glucose level >250 mg/dl, in range [70,180] mg/dl and <70 mg/dl. The lower figure shows the Inter-Quartile Ranges of glucose, where the mean values are represented with red dots.

The six patients of the dataset used in the *Agnostic* scenario were selected from the OhioT1DM dataset for BG Level Prediction: update 2020 [115]. This dataset contained twelve patients with T1DM and was first released in 2018 with half its current size, containing data for only six patients. We selected the new patients incorporated in this update because the sensor band used in these new patients (Empatica Embrace) registers some of the *Physiological* variables with more precision than the sensor band used in the old dataset (Basic Peak).

To protect the data contributors and ensure that the data are used only for research purposes, we had to sign a Data Use Agreement with the University of Ohio before using the dataset in our research.

Data were acquired over multiple weeks using insulin pump therapy with CGM. Patients wore Medtronic 530G or 630G insulin pumps and used the Medtronic Enlite CGM. They also wore the Empatica Embrace device, which reported life-event data via a custom smartphone app and provided *Physiological* data from a fitness band. Log entries were stored in five-minute intervals. In this dataset, we have at least 53 complete days of data for each patient. These days are not necessarily consecutive nor the same days for all the patients. Each log entry contains the date and time, the BG value, the amount of insulin, and the amount of carbohydrate intake as estimated by the patients. Additionally, the dataset includes self-reported mealtimes with carbohydrate estimates; self-reported times of exercise, sleep, work, stress, and illness; and data from the Empatica Embrace band, which includes one-minute aggregations of GSR, skin temperature, and magnitude of acceleration. Both bands indicated when they detected that the wearer was asleep, and this information is included when available. However, not all data contributors wore their sensor bands overnight. The population, in this case, has 83% of male patients with an average age between 33.33±16.33 and 53.33±16.33 years. The average number of days with data is 56.33±2.06. Figure 6.5 shows the glucose level of the patients from this dataset in the same way as in the *What-if* scenario.

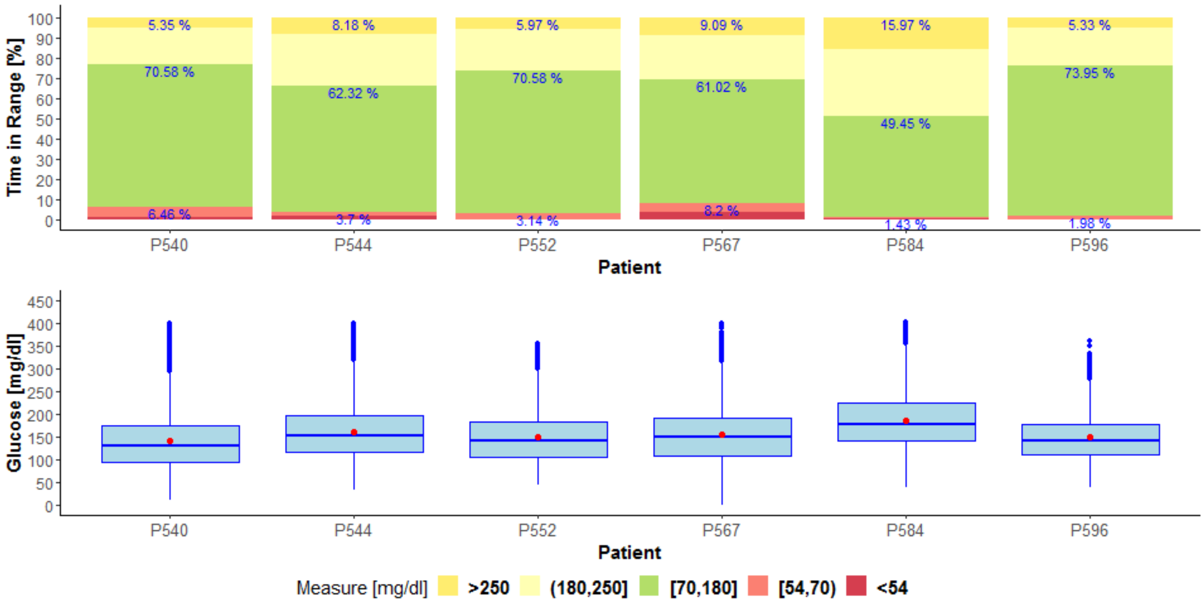


Figure 6.5: Histograms and boxplots describe the glucose level of the patients from the dataset used in the *Agnostic* scenario. The upper figure shows the percentage of time the patient has a very low glucose level (<54 mg/dl in dark-red), low ([54,70) mg/dl in red), in range ([70,180] mg/dl in green), high ((180,250] mg/dl in yellow), and very high (>250 mg/dl in dark-yellow). The numbers (from top to bottom) represent the percentage of time the patient has a glucose level >250 mg/dl, in range [70,180] mg/dl and <70 mg/dl. The lower figure shows the Inter-Quartile Ranges of glucose with mean values as red dots.

6.2.2 Configuration

The implementation of both GE and MO-GE algorithms is done in Java, and the code is publicly available at the GitHub repository called JECO, which stands for Java Evolutionary COmputation library [129]. As stated above, the multi-objective approach uses the NSGA-II as an optimization engine. Table 6.1 summarizes the configuration of the evolutionary engine for the MO-GE approach, which is based on our previous experience, and on a set of preliminary experiments where we studied the convergence of the algorithm as well as the wrapping value. In particular, little use was made of wrapping in the initial experiments. On the contrary, the algorithms tended to generate solutions of moderate size. Therefore, we set the maximum number of wrappings to a small value, evolving some complex expressions. To maintain a consistent computational effort, the same configuration was used for GE, which only considers RMSE as a fitness function. Finally, for each patient, ten runs were executed with the training data. The best model obtained was later used to calculate its performance on the test data.

Parameter settings of the multi-objective GE algorithm	
Grammar	gr120bvr.bnf
Objectives=2	F_{CLARKE} RMSE
Normalized data	No
Initialization	Random 50% and sensible 50%
Genetic operators	
Tournament size	2
Population size	400
Crossover probability	0.75
Mutation probability	0.01
Chromosome length	300
Number of generations	400
Maximum number of wraps	5
Elites	2

Table 6.1: Configuration of the evolutionary engine.

In addition to the GE models, we have also fitted ARIMA(p,d,q) models to estimate glucose values. Equation (6.7) presents the expression of such a model where $\epsilon(t)$ is the random error at time t, and p, q, and d integers called the orders of the model. This model only includes glucose values and does not use exogenous variables such as insulin doses or carbohydrates.

$$\widehat{G}_{ph}^{\text{arima}}(t) = \sum_{i=1}^{p} \alpha_i G(t-i) + \left(\epsilon(t) + \sum_{i=1}^{q} \theta_i \epsilon(t-i)\right) + \sum_{i=0}^{d} \phi_i t^i \tag{6.7}$$

We have tested ARIMA models of different orders in $0 < p,q \leq 12$, with 12 being the number of samples in one hour for Ohio patients to mimic the same time windows as the GE models. The best results have been for $p = q = 5$. We disregard the integrative component, that is, $d = 0$, because previous experiments show no improvement for $\phi_i \neq 0$. With every new glucose sample, the 12 model's coefficients are estimated using maximum likelihood in a 4-hour time window using the last 48 samples (including the last one) of the univariate glucose time series, $G(t)$. Once the model is fitted, it is warmed up using the last 24 samples, and then it estimates the glucose prediction for the four prediction horizons in PH.

6.2.3 Results

For the *What-if* scenario, in order to compare the results between the multi-objective (MO-GE) and the single-objective (GE) approaches, we have plotted in figures 6.6 to 6.8 the RMSE and CDE values for each solution model for several patients under different prediction horizons (see C.1 to C.15 for the rest of the patients). CDE is the sum of points in zones C, D, and E for each model. Since F_{CLARKE} is a weighted sum of the number of points in those areas, we prefer to represent CDE because it allows us to compare how similar values of RMSE may present different numbers of dangerous mispredictions, accounted for by CDE. This representation allows the comparison of solutions in terms of both the raw quality of a model, given by the RMSE value, and the medical quality of the model, given by the number of points in dangerous regions. It is important to note that a solution with two predictions in zone C is acceptable and better than one with two prediction points in zones D or E.

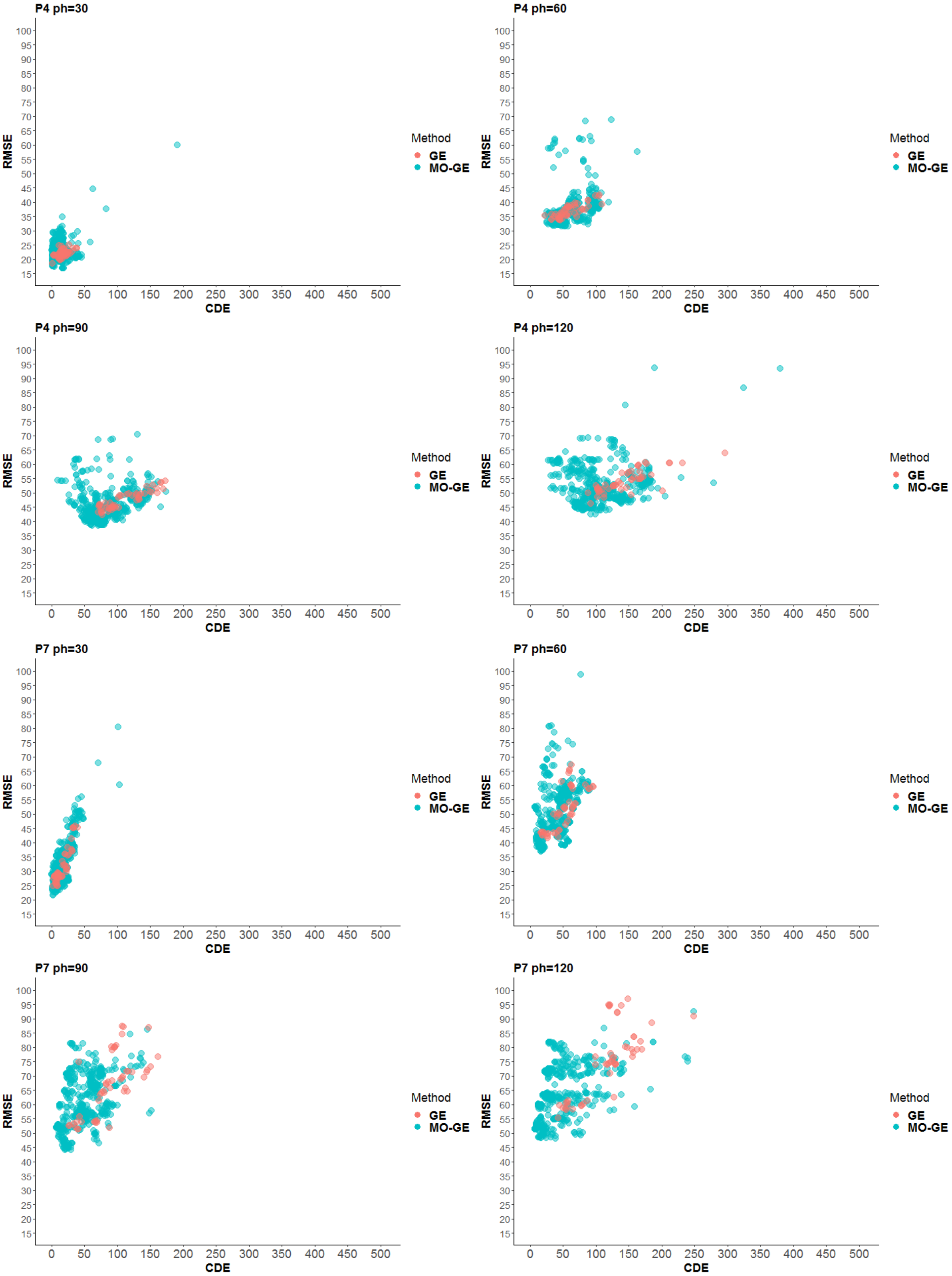


Figure 6.6: Comparison between GE and MO-GE (red and green points) for patients 4 and 7 in the *What-if* scenario for all prediction horizons (30 and 90 min in left column, and 60 and 120 min in right column) and both historical values.

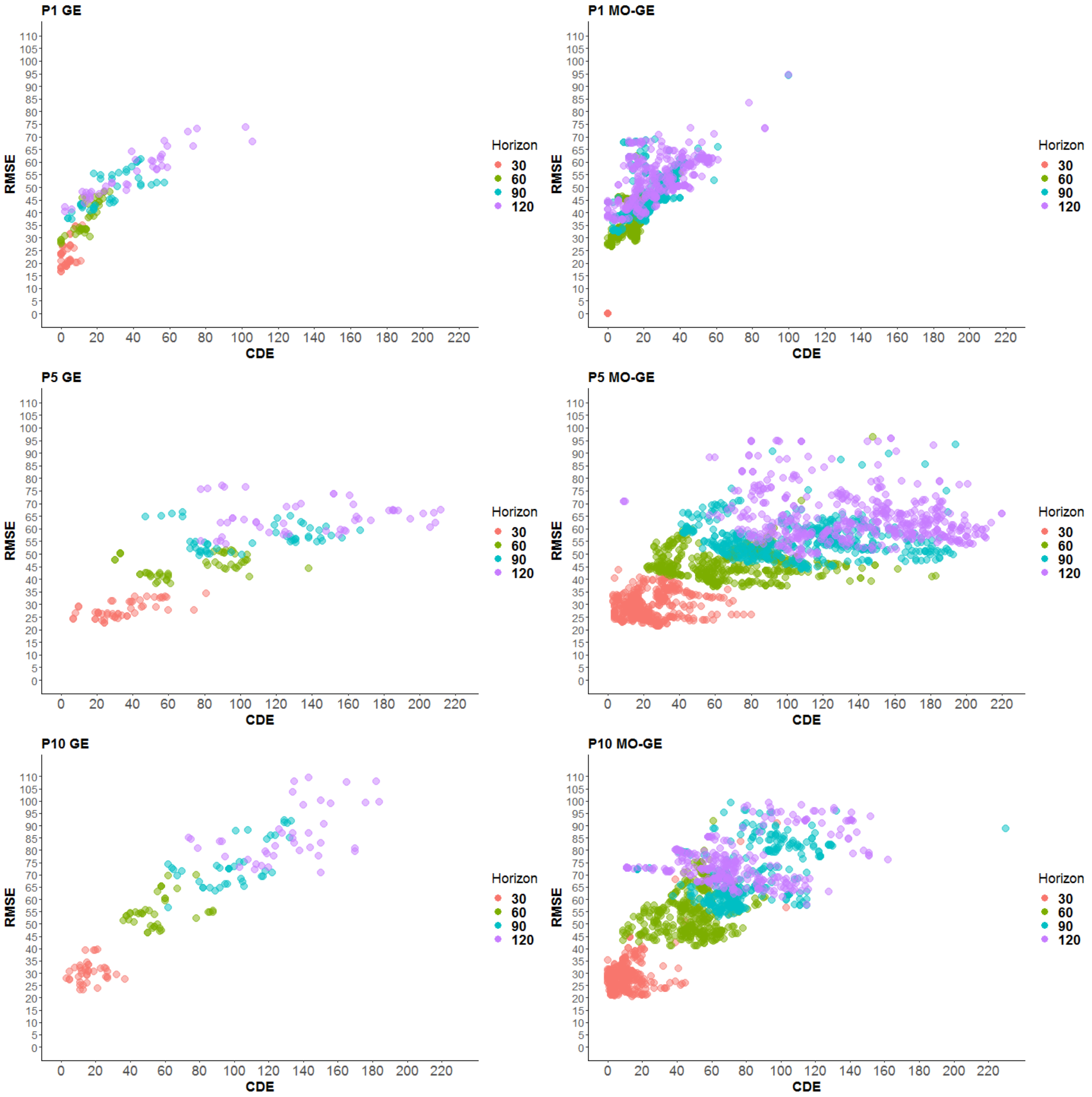


Figure 6.7: Root Mean Square Error and CDE values for all prediction horizons (30 min in red points, 60 min in green points, 90 min in blue points, and 120 min in purple points) for patients 1, 5, and 10 (rows one, two and three) in the *What-if* scenario for GE and MO-GE (left and the right column) and both historical values.

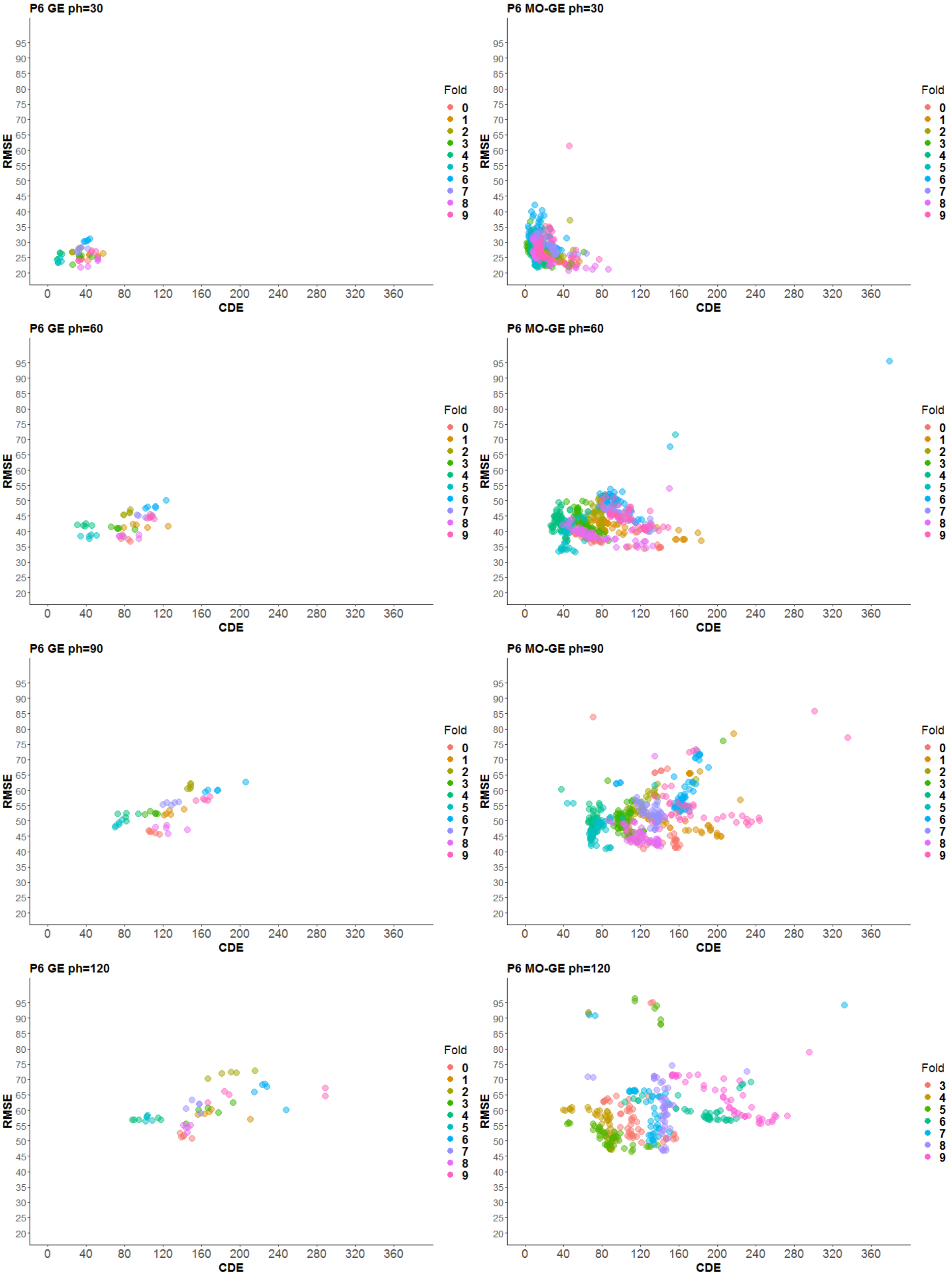


Figure 6.8: All folds (each color represents a different fold) for patient 6 in the *What-if* scenario for GE and MO-GE (left and right column) and all prediction horizons (row one for 30 min, row two for 60 min, row three for 90 min and row four for 120 min) and both historical values.

Figure 6.6 shows the solutions obtained for patients 4 and 7 (see C.1 to C.4 for patients 1, 2, 3, 5, 6, 8, 9 and 10) with both a single-objective (GE) evolution approach using the RMSE value as fitness function (red points) and the solutions obtained with the MO-GE method (green points). As can be seen, the use of F_{CLARKE} as an additional objective improves the fitness of the solutions, not only for the new fitness function but also in terms of RMSE. We can also observe that solutions developed by MO-GE dominate all the solutions generated by GE. Moreover, the distribution of the solutions suggests that our MO-GE is more robust than the GE since most solutions it finds are close to the approximation of the Pareto front. This is observed for all four prediction horizons. Similar results were obtained with the rest of the patients in the dataset. From a medical point of view, this result measures the method's robustness since patient 4 is very different from patient 7 according to the descriptive statistics. Patient 7 has the worst glucose control since the patients' glucose values lie within a healthy range only for a brief time (27%) compared to medical recommendations ($\geq$ 80%). The average glucose level is also poor for this patient since it is the highest in the group (176.33 mg/dl). Nevertheless, we can see in figure 6.6 that GE and MO-GE reach reasonable solutions for the four prediction horizons in both patients, regardless of the *a priori* difficulty of the dataset.

We want to highlight that MO-GE can reach the optimum solution for patient 1, that is, the solution with both objectives equal to 0 and, hence, with neither error nor dangerous predictions.

Figure 6.7 shows the distribution of the solutions in the multi-objective space with both approaches, GE and MO-GE, for three patients: 1, 5 and 10 (see figures C.5 and C.6 for patients 2, 3, 4, 6, 7, 8 and 9). As expected, the error and the number of points in dangerous prediction zones increase with the prediction horizon. However, solutions are restricted to a limited area of the graph (maximum RMSE by a maximum number of points in zones C, D, and E).

It is also important to note that, although some non-dominated solutions with a deficient number of points in zones CDE, those are not necessarily the best ones. For example, let us analyze the solutions from MO-GE for the 120-minute horizon, patient 5, with the lowest value of CDE in the graph (the magenta-colored points close to the Y-axis around point (71, 7)). This solution is exact. However, the wrong predictions could be hazardous for the patients since, as the RMSE indicates, the deviation from the correct prediction must be very high. When selecting the final model or predictor, the decision-maker should take these characteristics into account. Another interesting feature that can be extracted from figure 6.7 is that the longer the prediction horizon, the worse the RMSE value, which is consistent with the intuitive idea of the difficulty of prediction for long horizons. Nevertheless, this situation does not happen with CEG. Notice that a small set of solutions for the 120-minute horizon in patients 5 and 10 are close to the CDE value of 10, which is relatively small. This means that, despite those solutions having a high RMSE value (around 70), the predictions are well located in terms of the grid defined by CEG, which provides a good CDE value. In addition, it is observed that the variability of solutions for a prediction horizon of 120 minutes is the highest, and the variance of the data along the horizontal axis decreases as the prediction horizon gets closer. This means that good CDE values can be reached for every horizon (as well as poor ones). The variability is reduced with the reduction of the prediction horizon. These facts prove that the two objectives are not correlated and, hence, the multi-objective approach is correct.

Figure 6.8 presents an additional analysis where we can see the distribution of the solutions in the multi-objective space for patient 6 with both approaches, single-objective (GE) and multi-objective (MO-GE) with the four prediction horizons PH=$\{30, 60, 90, 120\}$ minutes (see figures C.7 to C.15 for the rest of the patients). Each color represents the solutions obtained with one of the folds of the 10-fold cross-validation strategy for a prediction horizon. It should be noted that some of the folds are more difficult to solve than others, and a deeper study of the data should be done to improve the algorithm, for instance, by detecting some pattern and determining when the MO-GE works better, as proposed in [130].

We analyzed the solutions quantitatively by comparing the 40 different instances (ten patients by four different time horizons) for both GE and MO-GE methods. In all cases, solutions obtained with the MO-GE method dominate solutions obtained with GE.

Table 6.2 shows the aggregated results for both the single-objective and multi-objective approaches, GE and MO-GE, respectively. It also shows the results obtained with the ARIMA approach. The percentage of predictions in zones A and B (A+B column), C, D, and E are depicted for each time horizon. The bold figures highlight the best method since we maximize the points in zones A+B and minimize the issues in dangerous zones C, D, and E. The figures reported in bold in columns A+B point out the method that has obtained the maximum value for each time horizon. In contrast, the figures reported in bold in columns C, D, and E point out the method that has obtained the minimum value for each time horizon. Again, the MO-GE algorithm reaches the best performance, reducing predictions in the most dangerous zones D and E for short-term predictions (30 and 60 minutes). For medium-term predictions (90 and 120 minutes), the sum of D and E points are very similar, on average. However, MO-GE finds better solutions considering the distribution of all points in the different zones of the CEG. Results were tested for statistical significance. We found significant differences in the number of points in zones D and E, where the multi-objective approach reduces the most dangerous predictions.

Concerning the technique with a more significant percentage of points in the A+B zone, MO-GE is placed first for short-term horizons, whereas GE achieves better results for medium-term horizons. ARIMA gets the worst results in terms of points in zones A+B and in terms of dangerous zones D and E for all horizons except 30 minutes, where it gets better results than the GE approach.

Algorithm	Horizon	A+B	C	D	E	Horizon	A+B	C	D	E
GE	30	92.81	**1.26**	5.31	0.60	90	**92.96**	**1.24**	5.22	**0.56**
MO-GE	30	**95.70**	2.38	**0.95**	0.97	90	91.77	2.37	**4.66**	1.20
ARIMA	30	95.48	1.59	2.47	**0.46**	90	82.26	7.84	5.26	4.64
GE	60	93.19	**1.23**	5.00	0.59	120	**90.76**	**1.35**	6.96	**0.91**
MO-GE	60	**94.39**	1.85	**3.24**	**0.52**	120	90.35	2.93	**5.01**	1.70
ARIMA	60	87.65	4.84	4.74	2.77	120	78.26	9.47	5.80	6.47

Table 6.2: Percentage of predictions for the *What-if* scenario and each zone of the Clarke Error Grid metric. Aggregated data of all the patients in the four-time horizons. The figures reported in the bold point out the best method for each horizon.

For the *Agnostic* scenario, we construct models with GE and MO-GE for the four prediction horizons PH={30, 60, 90, 120} minutes. We also compare models with access to the information of WS={60, 120} minutes before the time of predictions (see equation 6.3). In figures 6.9 and 6.10, we analyze the differences in the multi-objective approach (MO-GE) when compared to the single-objective (GE) for patients 567 and 584, taking into account the WS values for each figure (see figures C.16 to C.19 for patients 540, 544, 552 and 596). As in the previous section, these figures represent solutions in the multi-objective space, and each point represents a solution referenced by its coordinates (CDE, RMSE).

As can be seen, there are some cases where solutions obtained with historical values of ws=60 minutes with GE are dominated by solutions generated with MO-GE, as seen in figure 6.9 for P584 ws=60 and ph=30 minutes. There are other cases where solutions obtained with MO-GE are dominated by solutions generated with GE, as seen in figure 6.9 for P567 with ws=60 and ph=120 minutes. The solutions are similar when using historical values of ws=120 minutes. Figure 6.11 shows the results for patient 552 for both historical values (see figures C.20 to C.22 for the rest of the patients). In this case, we find solutions in the Pareto front for historical values and the different time horizons, so there is no dominance between methods. Similar results were obtained with the rest of the patients in the dataset.

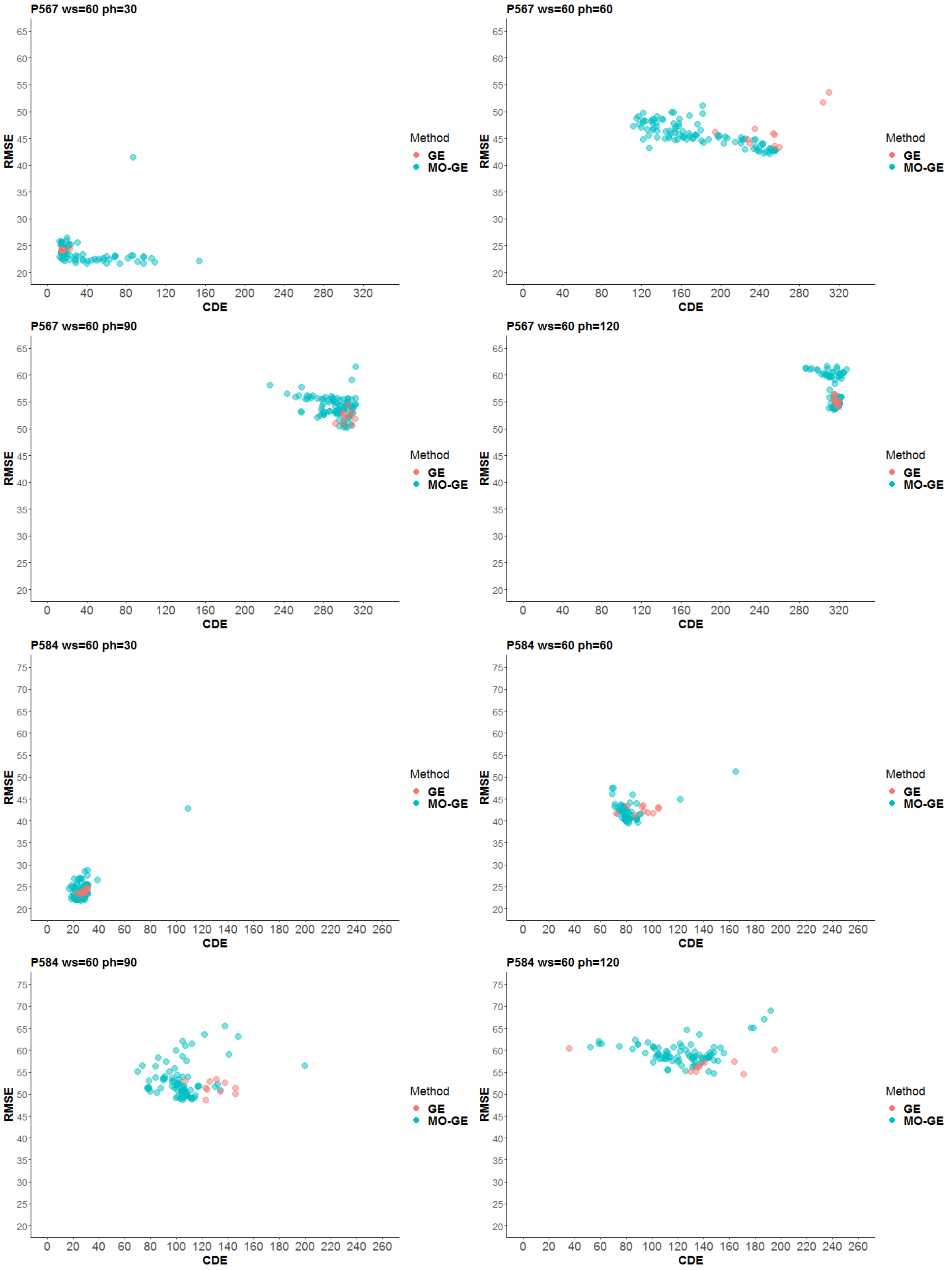


Figure 6.9: GE vs. MO-GE (red and green dots) for patients 567 and 584 (rows 1,2 and 3,4) in the *Agnostic* scenario for all prediction horizons (30 and 90 min in left column, and 60 and 120 min in right column) and historical values of 60 min.

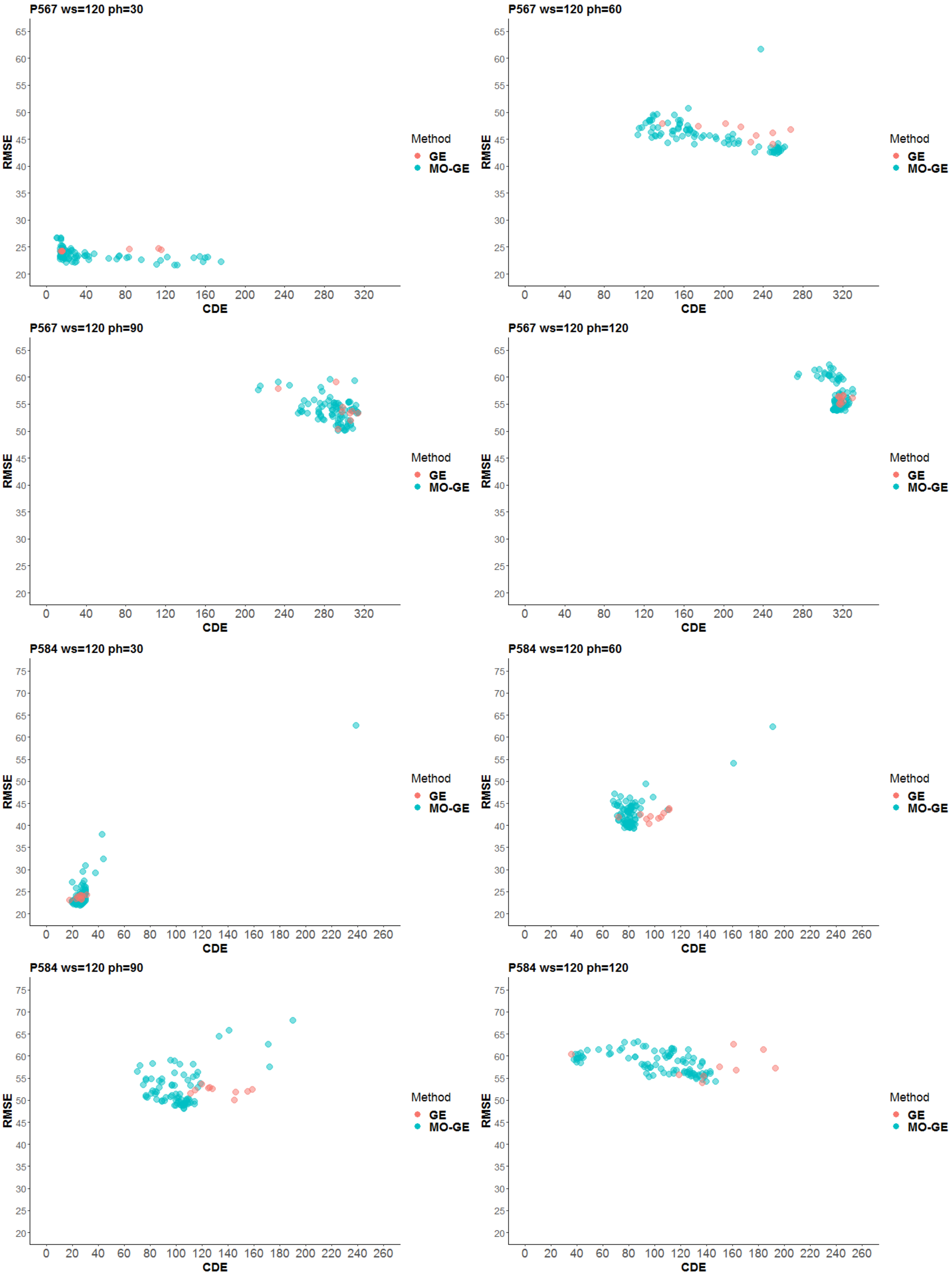


Figure 6.10: GE vs. MO-GE (red and green dots) for patients 567 and 584 (rows 1,2 and 3,4) in the *Agnostic* scenario for all prediction horizons (30 and 90 min in left column, and 60 and 120 min in right column) and historical values of 120 min.

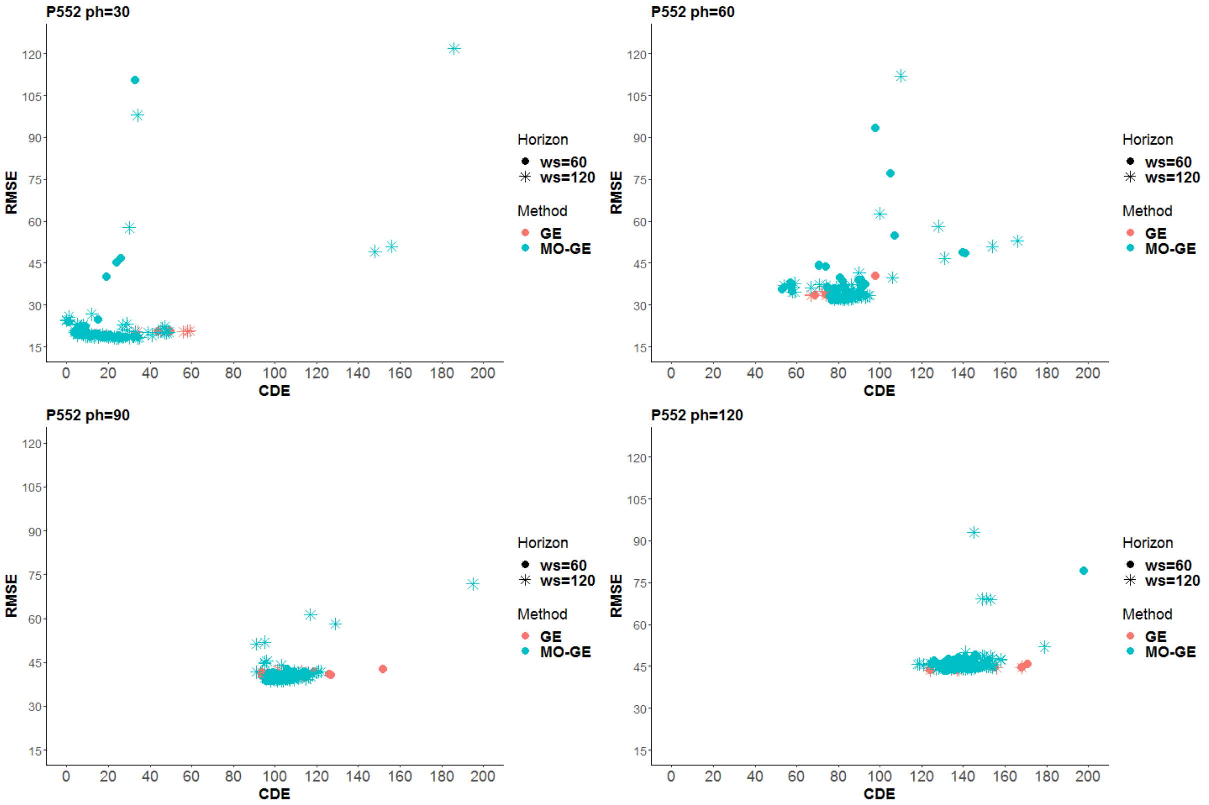


Figure 6.11: Solutions coming from historical values (60 min in dot shape and 120 min in star shape) for patient 552 in the *Agnostic* scenario for all prediction horizons (30 and 90 min in left column, and 60 and 120 min in right column) and both methods GE and MO-GE (red and green dots).

As well as in the *What-if* scenario, we compare the 24 different instances (six patients by four different time horizons) for both GE and MO-GE methods and the two historical values. For historical values of ws=60 minutes, in 18 out of 24 cases (75%), the solutions obtained with MO-GE dominate the solutions obtained with GE. For historical values of ws=120 minutes, in 19 out of 24 cases (79.17%), the solutions obtained with MO-GE dominate the solutions obtained with GE.

Table 6.3 shows the aggregated results for ws=120 minutes for the single-objective (GE), the multi-objective (MO-GE) and the ARIMA approaches. For each time horizon, the percentage of predictions in zones A and B (A+B column), C, D, and E are depicted, and the bold figures highlight the best method. As in the *What-if* scenario, the MO-GE algorithm reaches the best performance (in all horizons except for 30 minutes, where ARIMA gets the best results) and returns the best global solutions.

Algorithm	Horizon	A+B	C	D	E	Horizon	A+B	C	D	E
GE	30	89.89	4.95	3.63	1.53	90	88.79	5.30	**4.21**	1.70
MO-GE	30	95.95	1.62	2.40	**0.39**	90	**90.39**	**3.27**	5.17	**1.18**
ARIMA	30	**97.12**	**0.78**	**1.68**	0.42	90	79.63	8.77	5.41	6.19
GE	60	89.10	5.29	3.93	1.67	120	88.08	4.98	**4.40**	1.54
MO-GE	60	**91.68**	**3.23**	**3.85**	**1.05**	120	**91.77**	**1.72**	6.07	**0.44**
ARIMA	60	88.58	4.33	4.27	2.82	120	75.90	10.82	5.89	7.39

Table 6.3: Percentage of predictions for *Agnostic* scenario with ws=120 min and each zone of the Clarke Error Grid metric. Aggregated data of all the patients in the four-time horizons. The figures reported in the bold point out the best method for each horizon.

Similarly, MO-GE achieves the lowest percentage of points in zone E for all horizons. But if we look at the joint zone D+E, we find a different situation: MO-GE is only placed first in the 60-minute horizon, GE is the best for the medium-term (90 and 120 minutes), and the winner for the very short-term (30 minutes) is ARIMA.

It is interesting to analyze the complexity of the solutions in terms of the number of parameters and length of the solutions. Figure 6.12 shows the results obtained for both the *What-if* and *Agnostic* scenarios. The figure represents the average RMSE for each number of parameters found in the solutions for both the GE and MO-GE methods (left-hand side of the figure) and also about the length of the models (right-hand side of the figure). We have observed that solutions obtained with GE have a greater length, a greater number of parameters, and a higher RMSE than solutions obtained with MO-GE in both *What-if* and *Agnostic* scenarios. GE solutions have high variability in both the number of parameters and the length of the solutions. The variability is lower for MO-GE. Therefore, solutions obtained with MO-GE are less complex than solutions obtained with GE, and solutions obtained in the *What-if* scenario are more robust in terms of RMSE.

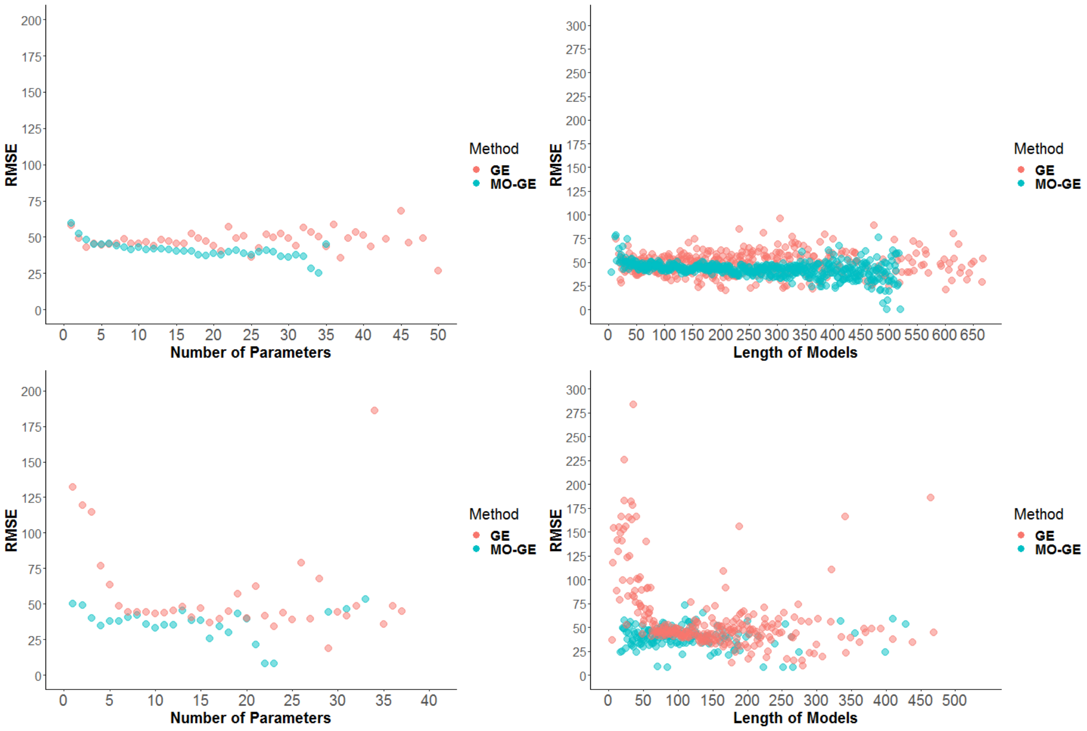
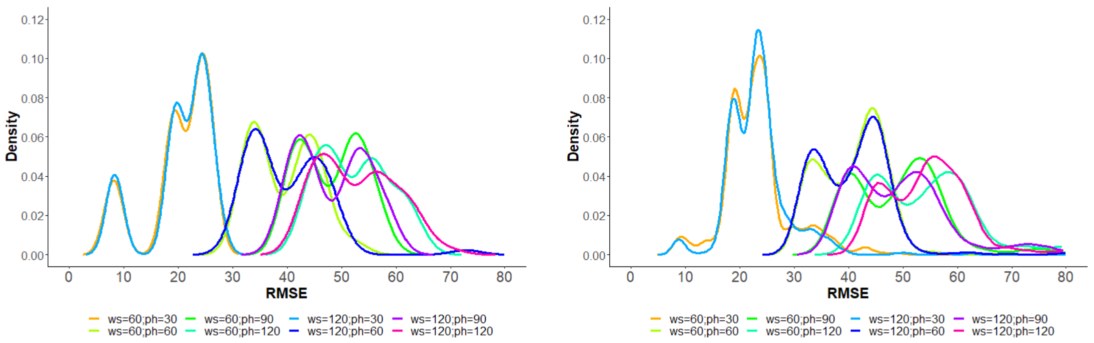


Figure 6.12: Analysis of the complexity of the solutions based on two criteria: the number of parameters and the length of the solutions. First row shows Root Mean Square Error vs. the number of parameters (left) and Root Mean Square Error vs. the length of the solutions (right) in the *What-if* scenario. The second row shows the results in the *Agnostic* scenario.

To study how the different historical values WS={60, 120} minutes contributed to the models, a deeper analysis would be required to assess the statistical significance of the results. To carry out this task, we first created density plots using a KDE for the distribution of the samples. The objective is to visualize whether the data meets the conditions for a parametric test, which is not the case. Figure 6.13 shows the results obtained with GE and MO-GE for the RMSE objective in the *Agnostic* scenario. Using Gaussian distribution, the variance is the same for all the cases. Data distribution is non-unimodal, and a non-parametric test is necessary. Similar results have been obtained with CDE but are not shown for the sake of space. All the plots were obtained following the method explained in [2].

Then, we followed the Bayesian model of [4, 111] based on the Plackett-Luce distribution over historical values and time horizons, taking into account the two methods and objective functions. We used a significance level α=0.05, with 20 Monte Carlo chains and 4000 simulations. Figure 6.14 shows the probability of being the best method, denoted as *probability of winning*, and its SD for the results obtained with GE and RMSE as objective functions. First, it can be seen that the prediction horizon ph=30 minutes is the best since both configurations with this horizon reach the highest probabilities of winning. Also, it can be seen that the historical value of ws=60 minutes has the highest probability. Even so, as the confidence interval overlaps with the results obtained with the historical value of ws=120 minutes, there is no statistical evidence that one method is better than the other. All the intervals for the rest of the time horizons overlap in the same way, and similar results have been obtained for the rest of the cases.



(a) GE

(b) MO-GE

Figure 6.13: Density plots of the Root Mean Square Error distribution for GE and MO-GE result for all the time horizons in the *Agnostic* scenario. The distributions are non-unimodal, and a non-parametric test is recommended.

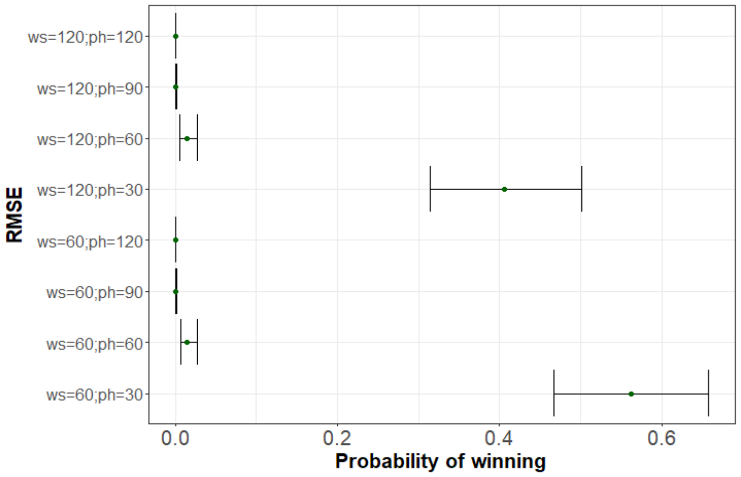


Figure 6.14: Bayesian model of [4] to analyse the GE results with Root Mean Square Error as objective function for the *Agnostic* scenario.

The experiments were performed on an Intel(R) Core (TM) i7-7700CPU at 3.60 GHz with 16 GB RAM Memory on Windows 10. Experiments were run with eight threads in parallel to benefit from all the core threads of the computer without affecting performance and no other task running at the same time. The average time to obtain a model with GE is 6.51 minutes and 3.10 hours for the MO-GE approach, almost 29 times slower.

6.3 Conclusions

In this work, which extends the work presented at the EvoStar 2020 Conference [112], we investigate the benefits and drawbacks of a multi-objective implementation of GE in the generation of models for BG prediction in PwD. In particular, we implement a multi-objective GE based on the classic NSGA-II, which is guided by two fitness functions: RMSE and F_{CLARKE}, which is a function designed following the CEG metric.

We have produced experimental evidence in two different scenarios: on the one hand, the *What-if* scenario, which takes into account future input data from the prediction horizon to predict glucose values. Data from ten actual PwD were studied in this scenario, as in [112]. On the other hand, we have reviewed the *Agnostic* scenario, where no information about future events is available. In this case, data from six actual patients from the OhioT1DM dataset [115] were studied.

Our results show that the multi-objective approach produces better models, reducing the number of predictions in the most dangerous zones of the CEG metric for both scenarios. These results are achieved because the multi-objective approach has a better ability than the single-objective approach to traverse the different areas of the search space defined by the F_{CLARKE} objective function. This is an essential conclusion since medical criteria are included in this function, which penalizes the models' most dangerous mispredictions. In addition, we have also found that GE can obtain good results in terms of CEG for both scenarios, despite it not being considered an objective function. However, the multi-objective approach can be regarded as safer since the CEG metric is explicitly included, and a decision-maker could examine non-dominated models and decide which one best fits a patient.

In addition to the new experimental results, we have performed a statistical analysis of the results for the *Agnostic* scenario. Our tests show no statistical evidence of the significant difference between using 60 and 120 minutes historical values in terms of better models. Hence, using historical values of 60 minutes is recommended since the complexity of the model is lower and the algorithms' execution time is shorter, requiring only half of the data.

Chapter 7

Conclusions and Future Works

In this thesis, the state-of-the-art models applied in the prediction of BG levels in T1DM patients have been improved in terms of accuracy and effectiveness. The time horizon of reliable predictions has also been increased for a predicting horizon up to two hours, usually the time needed to decide if the dose of insulin after a meal has been accurate and adequate.

In chapter 3, we demonstrated that using statistical techniques, it is possible to extract a small number of glucose profiles that characterize the majority of glucose patterns in patients with T1DM. Significant differences and dependencies have been observed among glucose profiles classified according to the variables day of the week and time slot. The groups found have been different for each patient, demonstrating the need for an individualized study. The results obtained indicate that the techniques applied can facilitate the mathematical modeling of glucose and may be used to create an individualized classifier for each patient that classifies glucose profiles according to the variables day of the week and time slot. The patients' glucose values can be predicted using this classifier by knowing what day the patient is in and in what time slot, obtaining more accurate models.

In the work presented in chapter 4, a methodology to obtain accurate predictors of glucose using a three-step process has been proposed improving the predictions of glucose values from previous works and increasing the time horizon of reliable predictions. In general, when using classified glucose values in models created with traditional GP, the accuracy of the prediction of the glucose values has improved compared to those of models made with the original dataset. Significant differences (p-value < 0.05) and associations have been observed among the glucose profiles classified using the independent variables day of the week and time slot. These significant differences found in the classification process can be helpful to correct and enhance habits or therapies in patients and to obtain more accurate models through automatic learning techniques and AI.

A new methodology has been created in chapter 5 to obtain accurate predictors of glucose, applying several latent GV measures (calculated through different measures of glycemic average, GV and glycemic risk) as new input features of the modeling engine. Four different GP variants were compared using PEG, two correspondings with previous work (*GP* and *CHAID-GP*), and two new proposed in this thesis (*LV-GP* and *LV-CHAID-GP*) that make use of a set of LGV features. Taking into account GV is important for developing good prediction models. Models created with LV have improved the quality predictions and have not produced predictions in the worst zone (E) and only a few points in the second-worst zone (D), even for a 150-minute prediction horizon. *CHAID-GP* and *LV-CHAID-GP* predictions were the best for all time horizons. *CHAID-GP* and *LV-CHAID-GP* were the best for all patients. In both cases, dangerous predictions were reduced concerning previous works. In general, for all patients and short-term horizons, *LV-GP*, *CHAID-GP* and *LV-CHAID-GP* have been more accurate models than *GP*. The analysis of the relative importance of the variables reveals that MEAN, PSTR and JI measures were in the top-ranking positions. The non-LV features, i.e., glucose, insulin, and carbohydrates, appeared in the top five positions of influence. These results also indicated the correctness and the coherence of the obtained models, which was expected. The statistical analysis was performed with a novel approach based on a Bayesian model and the Plackett-Luce distribution over rankings. It reveals that *LV-GP* (one of the approaches proposed in this thesis) has the highest probability of being the best for 30, 60, 90, and 120 minutes. For 150 and 180 minutes, *CHAID-GP* is the method with the highest probability of being the best.

The benefits of applying a multi-objective approach based on NSGA-II to solve an SR problem via GE have been investigated in chapter 6. We have produced experimental evidence in two different scenarios. On the one hand, the *What-if* scenario, which takes into account future input data from the prediction horizon to predict glucose values. On the other hand, we have studied the *Agnostic* scenario, where no information about future events was available. The multi-objective approach has produced better models, reducing the number of predictions in the most dangerous zones of the CEG metric for both scenarios. This is an essential conclusion since medical criteria are included in this function, which penalizes the models' most dangerous mispredictions. In addition, we have also found that GE can obtain good results in terms of CEG for both scenarios, despite it not being considered an objective function. However, the multi-objective approach can be regarded as safer since the CEG metric is explicitly included, and a decision-maker could examine non-dominated models and decide which one best fits a patient. Our tests showed no statistical evidence of the significant difference between using historical values of 60 and 120 minutes in terms of better models. Hence, using historical values of 60 minutes is recommended since the complexity of the model is lower and the algorithms' execution time is shorter, requiring only half of the data.

The models created in this thesis would, after clinical testing, be directly applicable to daily clinical practice and could be integrated into mobile and web applications. The recommendations generated could help the user to decide the insulin dose and the actions to take. The joint action of the models and digital technologies will make it possible to automate patient therapies, thus improving quality of life and patient autonomy. We aim to facilitate individualized mathematical modeling for each patient, which could allow clinicians to find significant differences and may eventually lead to more accurate models using ML and AI techniques.

Now, we are developing a framework to generate and deploy models created with the multi-objective approach for both the *What-if* and *Agnostic* scenarios. These models will be available online through two applications hosted on the following web pages: *glucmodel.ucm.es* and *glucnet.ucm.es*. These models could be used to design insulin or carbohydrate recommendation systems. In addition, our models could help to test possible treatment modifications without putting the patient's health at risk. *Agnostic* models could be useful when more smart devices are available. Specifically, we are working on a smartwatch application that generates alarm signals when the patient is in risky situations.

Our procedures and models have been evaluated by the medical staff collaborating on the project. They have concluded that these can be useful for correcting and improving patients' lifestyles and therapies. Subject to further clinical validation and regulatory approval, the applications we are developing have the potential to be helpful for the daily management of diabetes to improve glycemic control and increase patients' quality of life and autonomy. Our technique can be applied to other areas of medicine where a similar set of physical variables are available for measurement. For example, alerts for potentially dangerous heart rates can be generated by developing models based on the historical values measured by smart devices.

We can obtain white-box models that can potentially be interpreted in terms of the variables used by applying our modeling technique. Although perhaps not explainable in terms of *Physiological* aspects of the body, the advantage being that we do not need an initial model, and the search is not limited to *Previously Adopted* models.

Our future work will be focus on the refinement of the models. Datasets from actual patients include many different input variables whose influence on the model's changes. Therefore, an analysis of each variable's contribution could help create more precise models, even for different times of day (e.g., morning, afternoon, and night).

Also, we will explore other combinations of evolutionary computation techniques with fuzzy logic and neural network approaches. According to the nature of the data, different clustering algorithms will be applied (e.g., K-means, K-shape, RNN). We will explore splitting the dataset into others combinations such as different meals of the day (breakfast, lunch, and dinner) to improve the quality of the glucose patterns. The glucose profiles will be used with the multi-objective approach taking into account LGV features to improve the current accuracy and effectiveness of the models applied in the prediction of BG levels in T1DM patients.

Chapter 8

Publications

The publications generated throughout this thesis are the following:

- S. Contador, J. M. Velasco, O. Garnica, and J. I. Hidalgo, “Glucose forecasting using genetic programming and latent glucose variability features”. Applied Soft Computing Journal (2021), vol. 110, p. 107609. Ed Elsevier. IF: 6.725. Q1: 11/112 (Computer Science, Interdisciplinary Applications).
- S. Contador, J. M. Colmenar, O. Garnica, J. M. Velasco, and J. I. Hidalgo, “Blood glucose prediction using multi-objective grammatical evolution: analysis of the agnostic and what-if scenarios”. Genetic Programming and Evolvable Machines Journal (2021). Ed. Springer. IF: 1.781. Q2, 260/693 (Computer Science Applications).
- S. Contador, J. M. Velasco, O. Garnica, and J. I. Hidalgo, “Profiled glucose forecasting using genetic programming and clustering”. pp. 529–536, 03. Symposium On Applied Computing. Brno, Czech Republic (2020). CORE B.
- S. Contador, J. M. Colmenar, O. Garnica, and J. I. Hidalgo, “Short and medium term blood glucose prediction using multi-objective grammatical evolution”. pp. 494–509. 04. XXIII European Conference on the Applications of Evolutionary and Bio-inspired Computation. Sevilla, Spain (2020). CORE B.

Appendices

Appendix A

Additional Material for Chapter 4

Table A.1: Predictions (in percentage) obtained for the zone A+B with Parkes Error Grid for Genetic Programming models, and the average percentage for *CHAID-GP* models. *Green* values indicate better solutions for *CHAID-GP*. The best result is highlighted in bold.

Model	t+30	t+60	t+90	t+120	t+150	t+180	t+210	t+240
Patient1								
GP	95.14	94.70	94.44	94.62	92.54	91.02	89.35	89.17
CHAID-GP	**96.32**	96.10	95.28	93.53	91.77	92.17	92.00	92.00
Patient2								
GP	89.81	90.89	87.33	83.59	71.75	72.32	78.89	83.36
CHAID-GP	90.45	90.36	90.43	89.77	91.03	**91.37**	90.38	90.76
Patient3								
GP	96.77	95.21	95.94	96.86	**98.08**	97.66	96.60	96.72
CHAID-GP	96.04	97.06	96.66	96.95	97.03	97.66	96.84	97.10
Patient4								
GP	91.19	64.60	92.94	89.87	92.50	93.42	92.05	91.72
CHAID-GP	93.74	93.13	92.16	91.98	92.64	**94.06**	92.79	92.59
Patient5								
GP	87.60	81.89	85.87	79.72	**89.28**	85.50	84.63	84.28
CHAID-GP	86.18	86.21	85.99	83.20	85.09	86.03	84.99	85.51
Patient6								
GP	89.20	85.61	85.68	85.79	88.62	**91.57**	86.65	86.86
CHAID-GP	90.46	88.65	89.22	88.11	86.92	90.03	89.69	89.69
Patient7								
GP	89.56	87.97	89.52	87.59	91.29	91.24	89.15	89.23
CHAID-GP	91.22	90.57	**92.76**	91.96	91.44	92.60	92.24	91.52
Patient8								
GP	88.95	92.94	92.38	**94.05**	93.41	86.33	91.35	91.54
CHAID-GP	91.27	92.50	93.84	90.55	92.69	88.84	92.36	92.21
Patient9								
GP	87.23	86.45	85.27	82.40	79.73	90.03	78.73	85.44
CHAID-GP	91.55	90.81	90.03	85.51	**92.17**	90.55	90.37	90.97
Patient10								
GP	**89.39**	80.98	80.87	78.53	84.79	86.77	83.10	83.17
CHAID-GP	86.91	84.08	85.71	83.65	86.66	88.06	85.91	85.44

Appendix B

Additional Material for Chapter 5

Table B.1: Predictions (in percentage) obtained for the zone A+B with Parkes Error Grid for *GP* and *LV-GP* models, and the average percentage for *CHAID-GP* and *LV-CHAID-GP* models. *Green* values indicate better solutions compared to *GP*. The best result is highlighted in bold.

Model	t+30	t+60	t+90	t+120	t+150	t+180	t+210	t+240
Patient1								
GP	95.14	94.70	94.44	94.62	**92.54**	91.02	89.35	89.17
CHAID-GP	96.32	96.10	**95.28**	93.53	91.77	**92.17**	**92.00**	**92.00**
LV-GP	**99.49**	**97.41**	94.72	**93.87**	87.44	69.63	62.01	68.34
LV-CHAID-GP	98.04	96.06	92.74	90.76	80.82	78.08	79.98	80.14
Patient2								
GP	89.81	90.89	87.33	83.59	71.75	72.32	78.89	83.36
CHAID-GP	90.45	90.36	90.43	**89.77**	**91.03**	**91.37**	**90.38**	**90.76**
LV-GP	**99.35**	96.48	**94.87**	89.55	84.31	55.63	44.98	71.89
LV-CHAID-GP	99.00	**96.49**	91.52	88.59	84.29	79.56	81.51	79.08
Patient3								
GP	96.77	95.21	95.94	96.86	**98.08**	**97.66**	96.60	96.72
CHAID-GP	96.04	97.06	96.66	**96.95**	97.03	**97.66**	**96.84**	**97.10**
LV-GP	**99.81**	98.42	**97.70**	96.85	96.13	95.19	89.20	86.39
LV-CHAID-GP	99.69	**98.80**	96.40	93.84	92.45	91.20	94.09	93.36
Patient4								
GP	91.19	64.60	92.94	89.87	92.50	93.42	92.05	91.72
CHAID-GP	93.74	93.13	92.16	91.98	**92.64**	**94.06**	**92.79**	**92.59**
LV-GP	**98.85**	**96.61**	93.80	92.51	91.79	91.11	90.05	89.37
LV-CHAID-GP	98.34	95.82	**94.39**	**93.20**	92.17	90.78	90.22	89.83
Patient5								
GP	87.60	81.89	85.87	79.72	**89.28**	85.50	84.63	84.28
CHAID-GP	86.18	86.21	85.99	83.20	85.09	**86.03**	**84.99**	**85.51**
LV-GP	**97.25**	**92.41**	**88.99**	**86.48**	84.02	83.55	58.80	44.39
LV-CHAID-GP	95.28	89.18	84.02	77.44	75.43	72.92	78.39	77.32
Patient6								
GP	89.20	85.61	85.68	85.79	**88.62**	**91.57**	86.65	86.86
CHAID-GP	90.46	88.65	89.22	**88.11**	86.92	90.03	**89.69**	**89.69**
LV-GP	**97.70**	**93.44**	**90.38**	80.46	83.96	76.60	70.08	67.94
LV-CHAID-GP	96.24	90.78	85.79	82.85	80.39	78.95	82.37	84.33
Patient7								
GP	89.56	87.97	89.52	87.59	91.29	91.24	89.15	89.23
CHAID-GP	91.22	90.57	92.76	91.96	**91.44**	**92.60**	**92.24**	**91.52**
LV-GP	**99.35**	**96.57**	**94.04**	**92.34**	82.72	64.47	74.50	80.59
LV-CHAID-GP	97.82	93.90	88.42	84.81	85.45	85.36	85.38	82.77
Patient8								
GP	88.95	92.94	92.38	**94.05**	**93.41**	86.33	91.35	91.54
CHAID-GP	91.27	92.50	**93.84**	90.55	92.69	**88.84**	**92.36**	**92.21**
LV-GP	**98.23**	**95.10**	93.01	83.28	83.19	54.40	53.00	72.96
LV-CHAID-GP	88.77	91.09	89.80	89.47	90.47	88.17	88.71	89.03
Patient9								
GP	87.23	86.45	85.27	82.40	79.73	90.03	78.73	85.44
CHAID-GP	91.55	90.81	90.03	85.51	**92.17**	**90.55**	**90.37**	**90.97**
LV-GP	**97.91**	90.81	88.19	85.94	84.09	81.39	75.60	87.56
LV-CHAID-GP	96.64	**91.89**	**91.83**	**89.08**	83.82	84.54	84.72	87.77
Patient10								
GP	89.39	80.98	80.87	78.53	84.79	86.77	83.10	83.17
CHAID-GP	86.91	84.08	85.71	83.65	**86.66**	**88.06**	**85.91**	**85.44**
LV-GP	**97.28**	**92.84**	**87.87**	**84.70**	83.28	69.93	80.75	65.53
LV-CHAID-GP	97.17	92.52	85.65	80.68	76.51	73.79	79.63	77.58

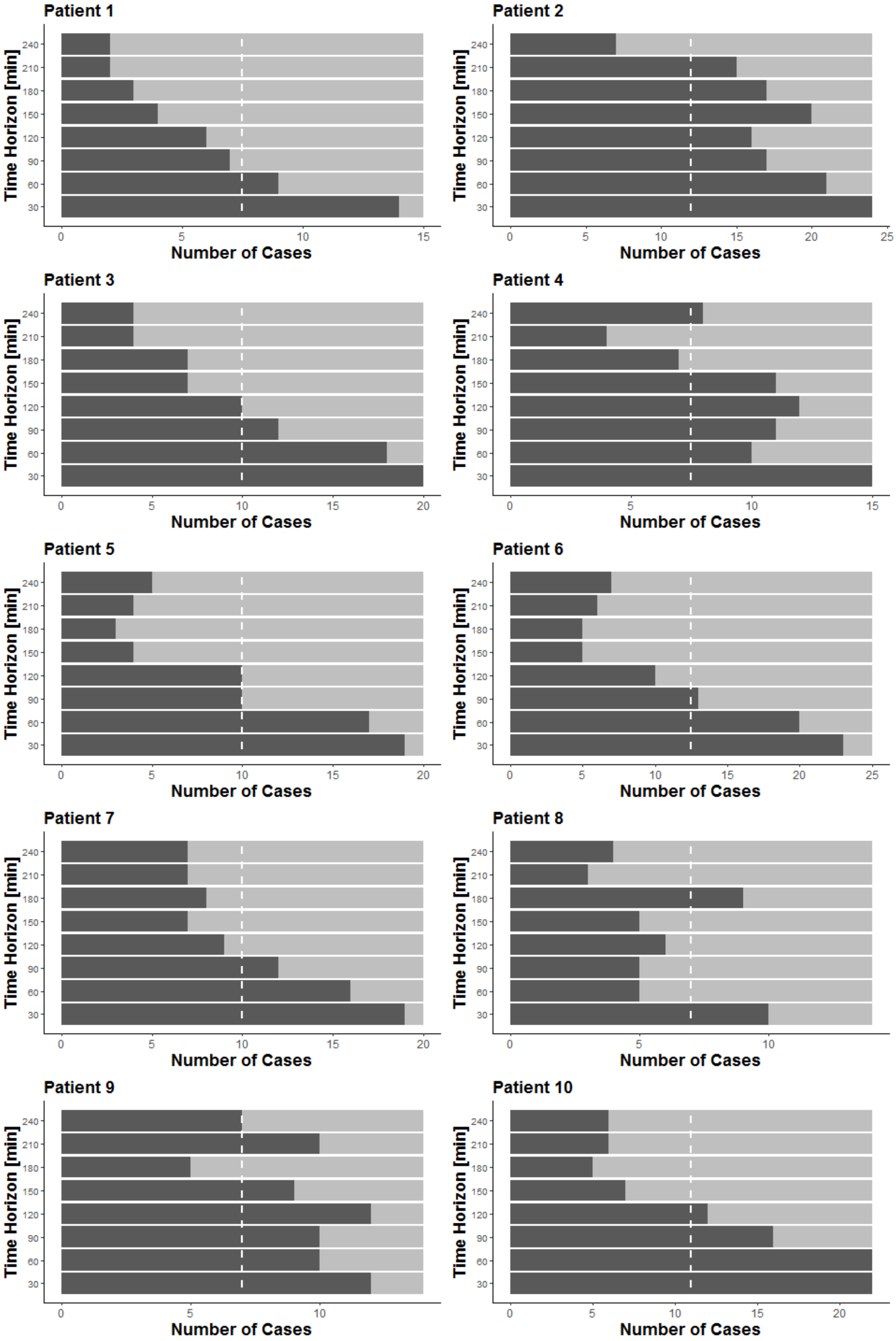


Figure B.1: The ratio of models with better predictions (higher values in zone A+B) for the different time horizons of data. Dark-gray segments indicate that the best prediction is made by *LV-CHAID-GP* while light-gray means a better model from *GP*.

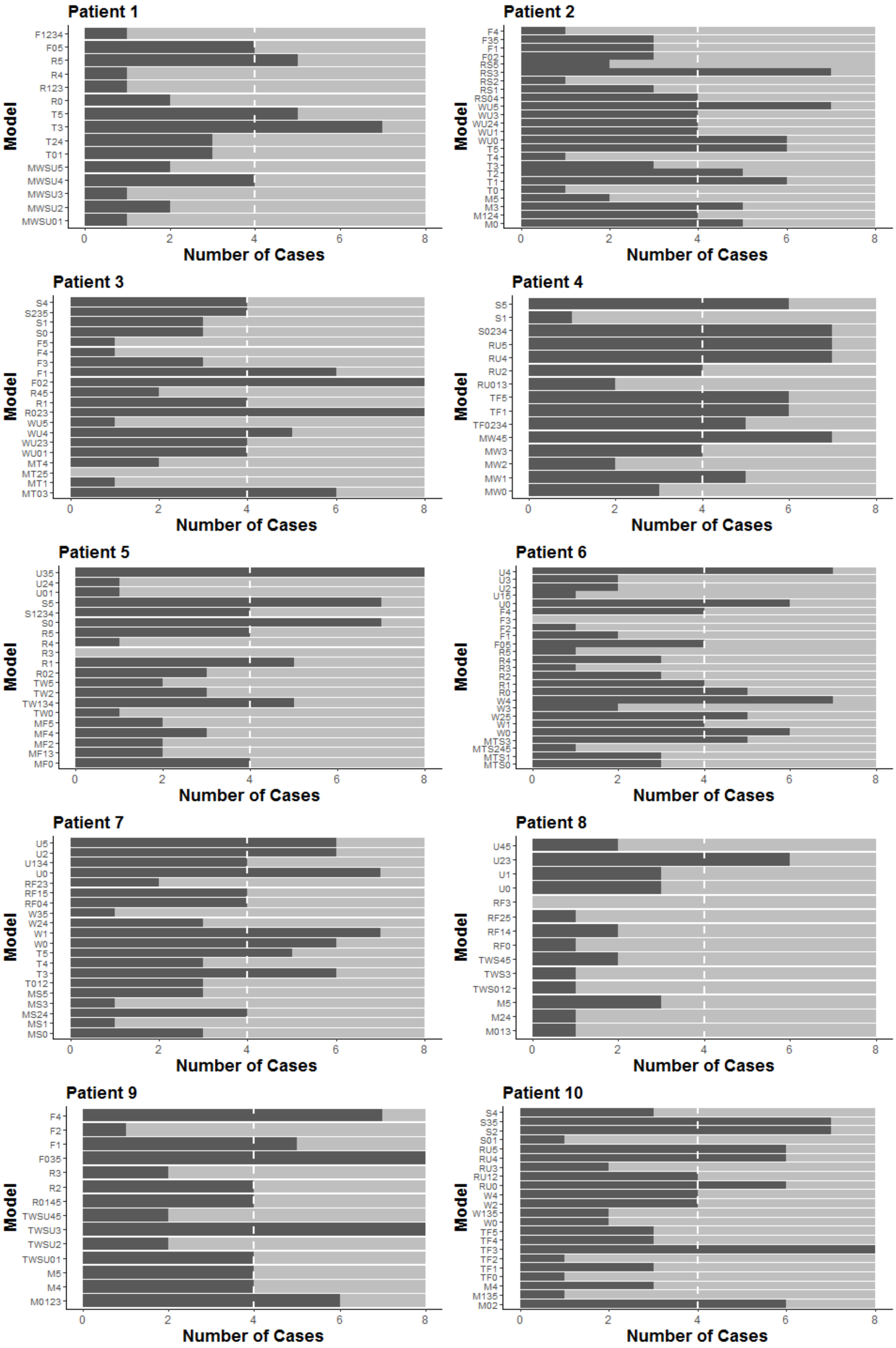
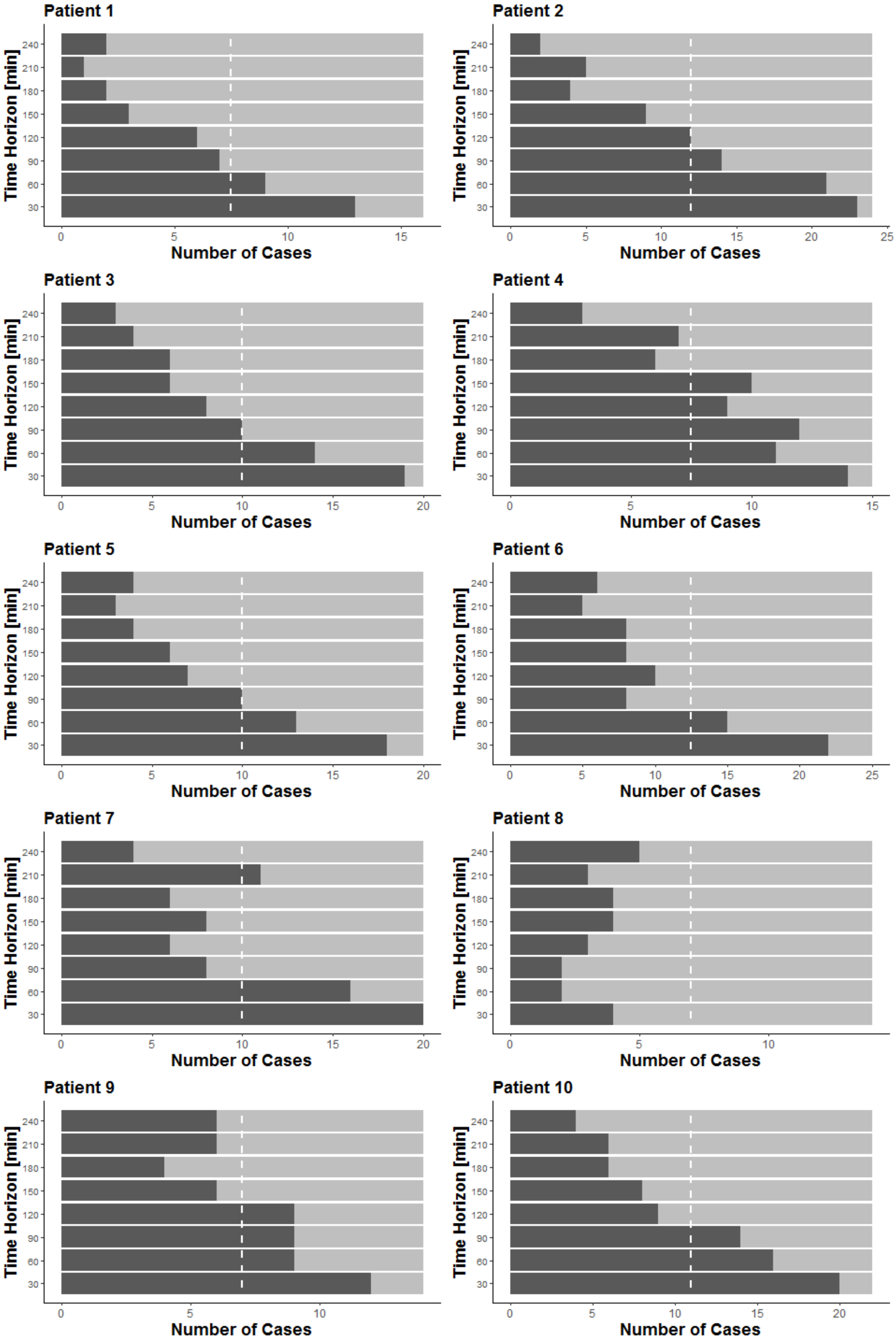


Figure B.2: The ratio of models with better predictions (higher values in zone A+B) for different data clusters. Labels in the X-axis represent the number of cases. Each case represents a time horizon, and the label Y-axis identifies clusters found with the clustering algorithm according to the criteria day of the week and time slot. Dark-gray segments indicate that the best prediction is made by *LV-CHAID-GP* while light-gray means a better model from *CHAID-GP*.



Patient 1

Patient 2

Patient 3

Patient 4

Patient 5

Patient 6

Patient 7

Patient 8

Patient 9

Patient 10

Figure B.3: The ratio of models with better predictions (higher values in zone A+B) for the different time horizons of data. Dark-gray segments indicate that the best prediction is made by *LV-CHAID-GP* while light-gray means a better model from *CHAID-GP*.

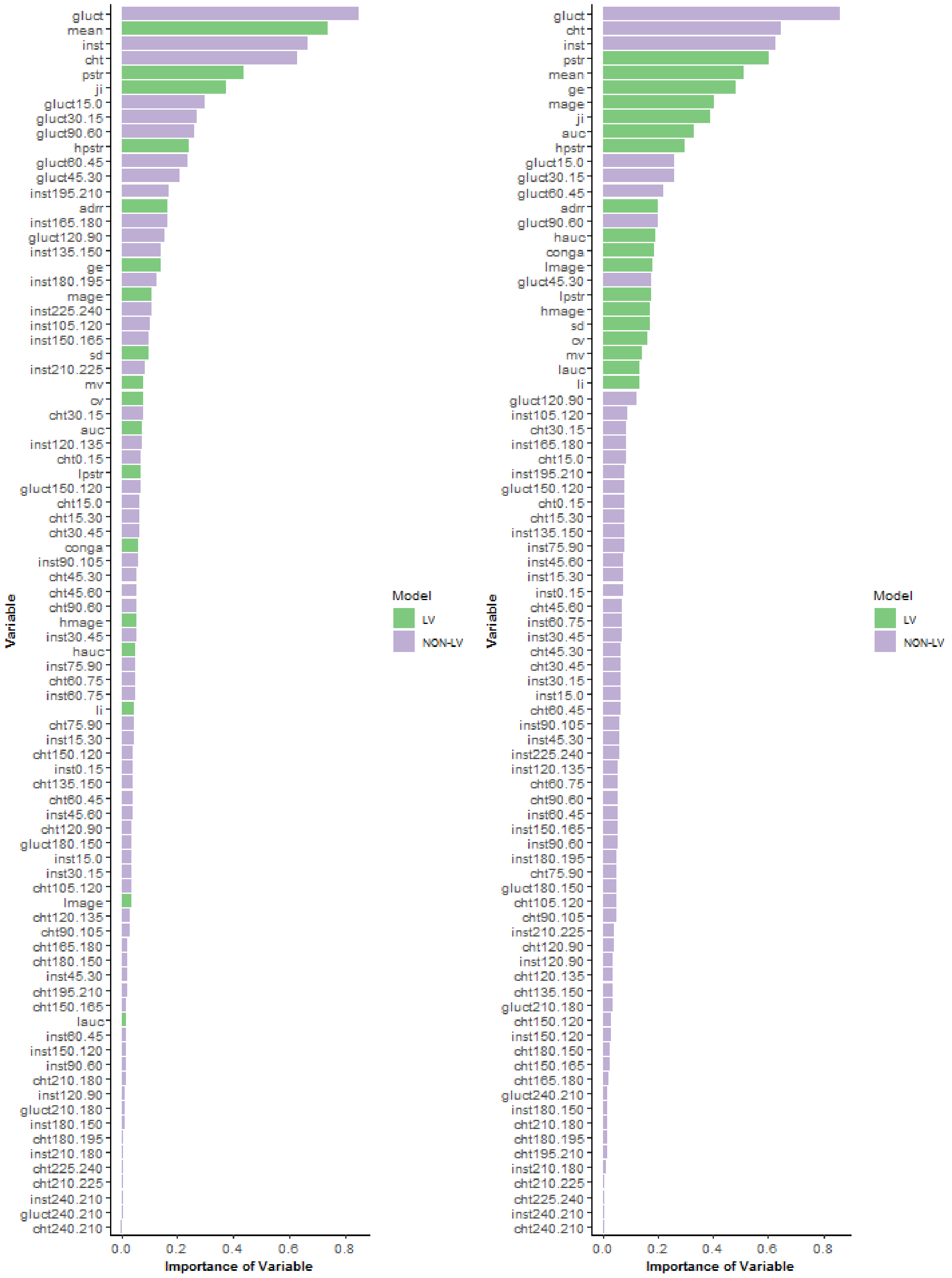


Figure B.4: The relative importance of all variables that make up the models is based on *LV-GP* and *LV-CHAID-GP*. The importance is between 0 and 1, where 0 is a variable that never appears in the model and 1 is a variable that always appears in the model. Latent Variables are represented in green, and historical and future variables in purple.

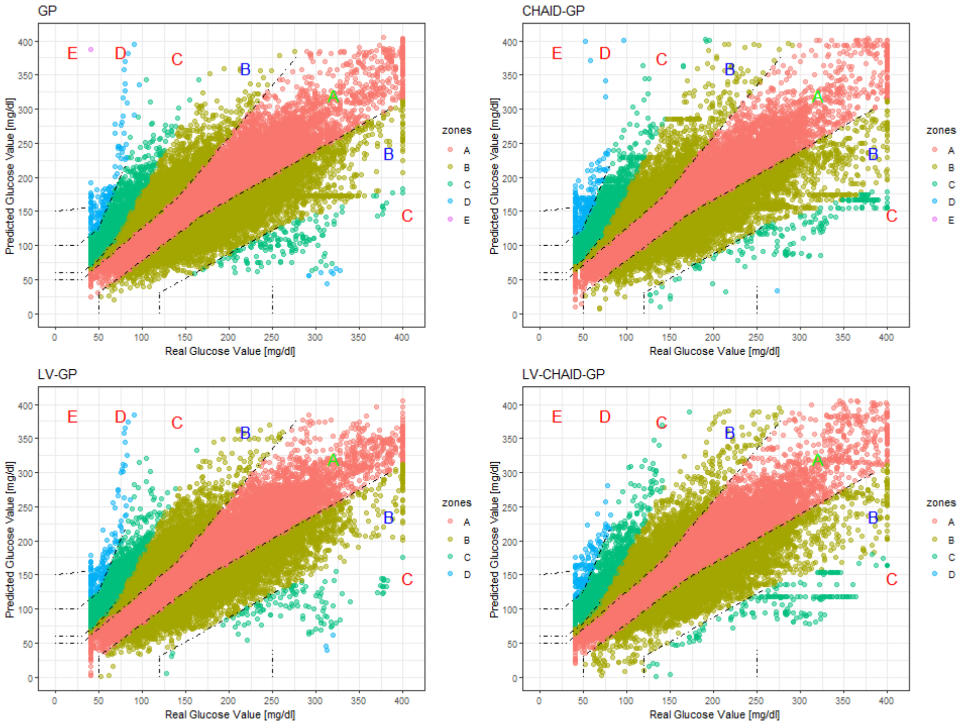


Figure B.5: Parkes Error Grid from test data for all different models of all patients for fold-0 and time horizon of 60 minutes. The X-axis represents the reference values of glucose, and the Y-axis the values of the prediction. Each zone of the Parkes Error Grid is identified with a different color.

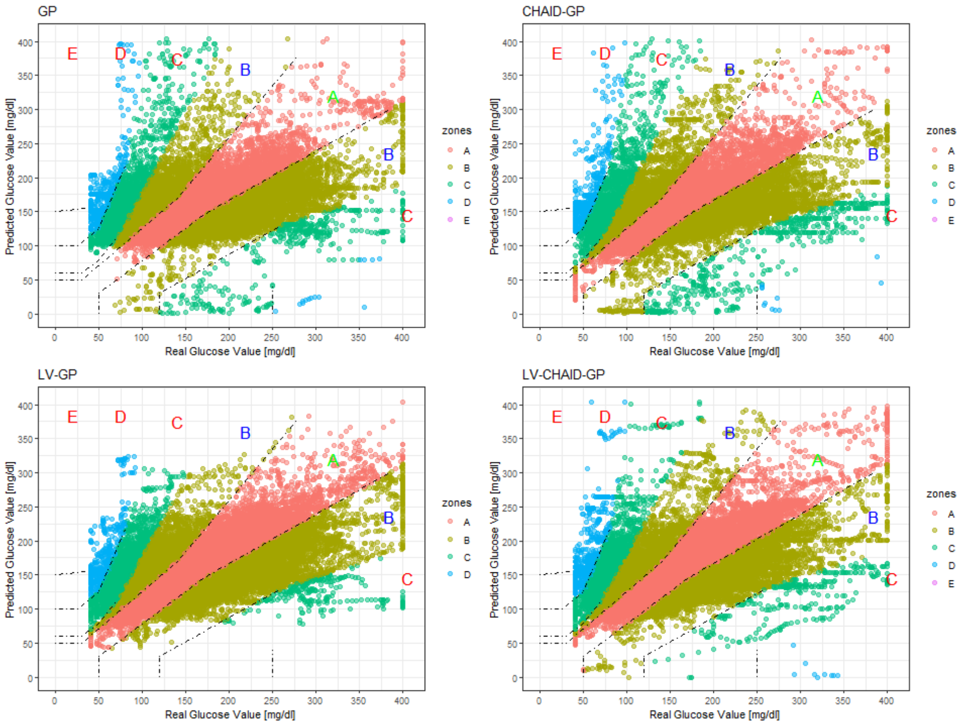


Figure B.6: Parkes Error Grid from test data for all different models of all patients for fold-0 and time horizon of 120 minutes. The X-axis represents the reference values of glucose, and the Y-axis the values of the prediction. Each zone of the Parkes Error Grid is identified with a different color.

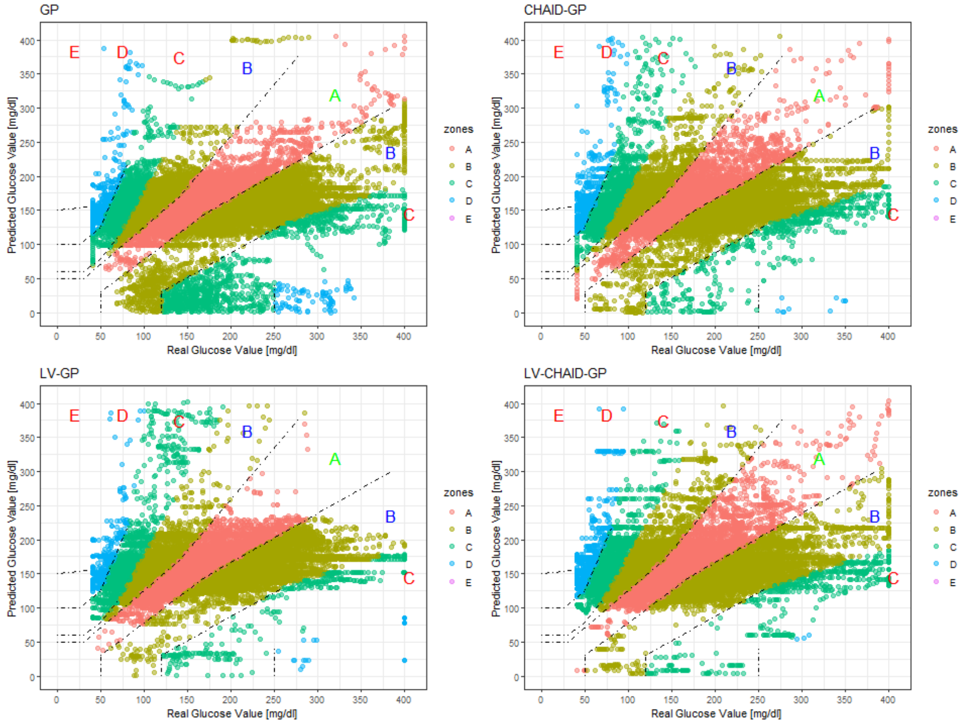


Figure B.7: Parkes Error Grid from test data for all different models of all patients for fold-0 and time horizon of 180 minutes. The X-axis represents the reference values of glucose, and the Y-axis the values of the prediction. Each zone of the Parkes Error Grid is identified with a different color.

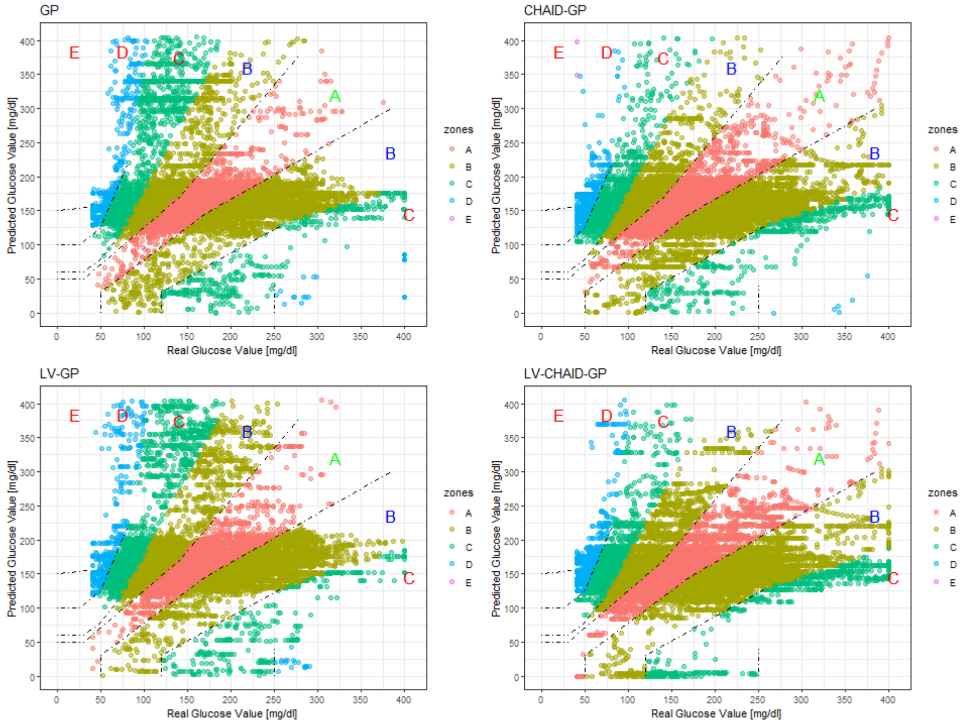


Figure B.8: Parkes Error Grid from test data for all different models of all patients for fold-0 and time horizon of 210 minutes. The X-axis represents the reference values of glucose, and the Y-axis the values of the prediction. Each zone of the Parkes Error Grid is identified with a different color.

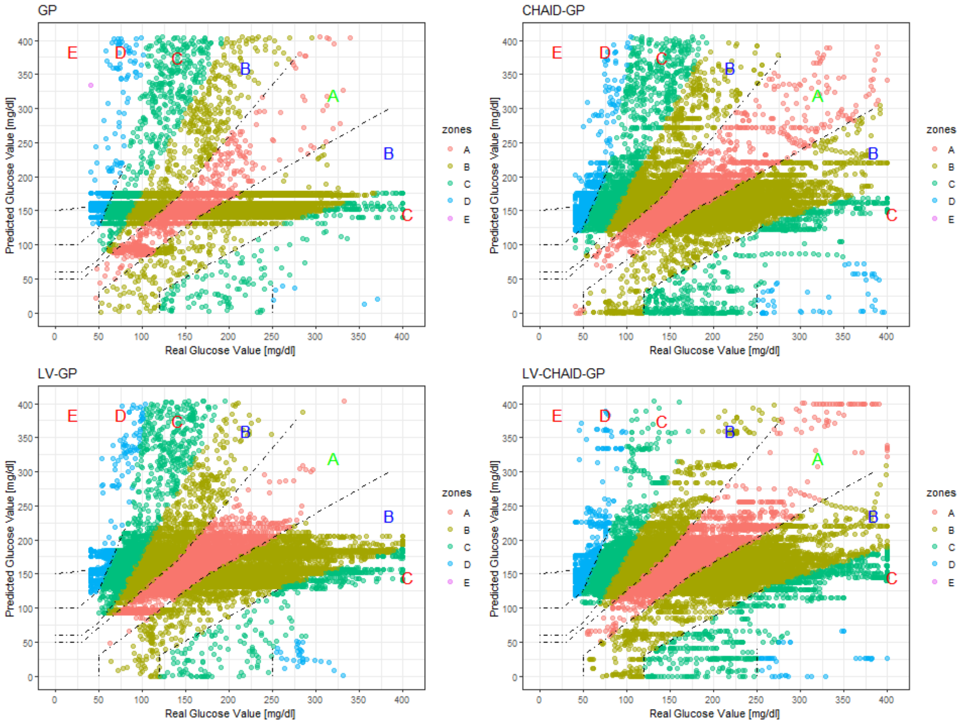


Figure B.9: Parkes Error Grid from test data for all different models of all patients for fold-0 and time horizon of 240 minutes. The X-axis represents the reference values of glucose, and the Y-axis the values of the prediction. Each zone of the Parkes Error Grid is identified with a different color.

Appendix C

Additional Material for Chapter 6

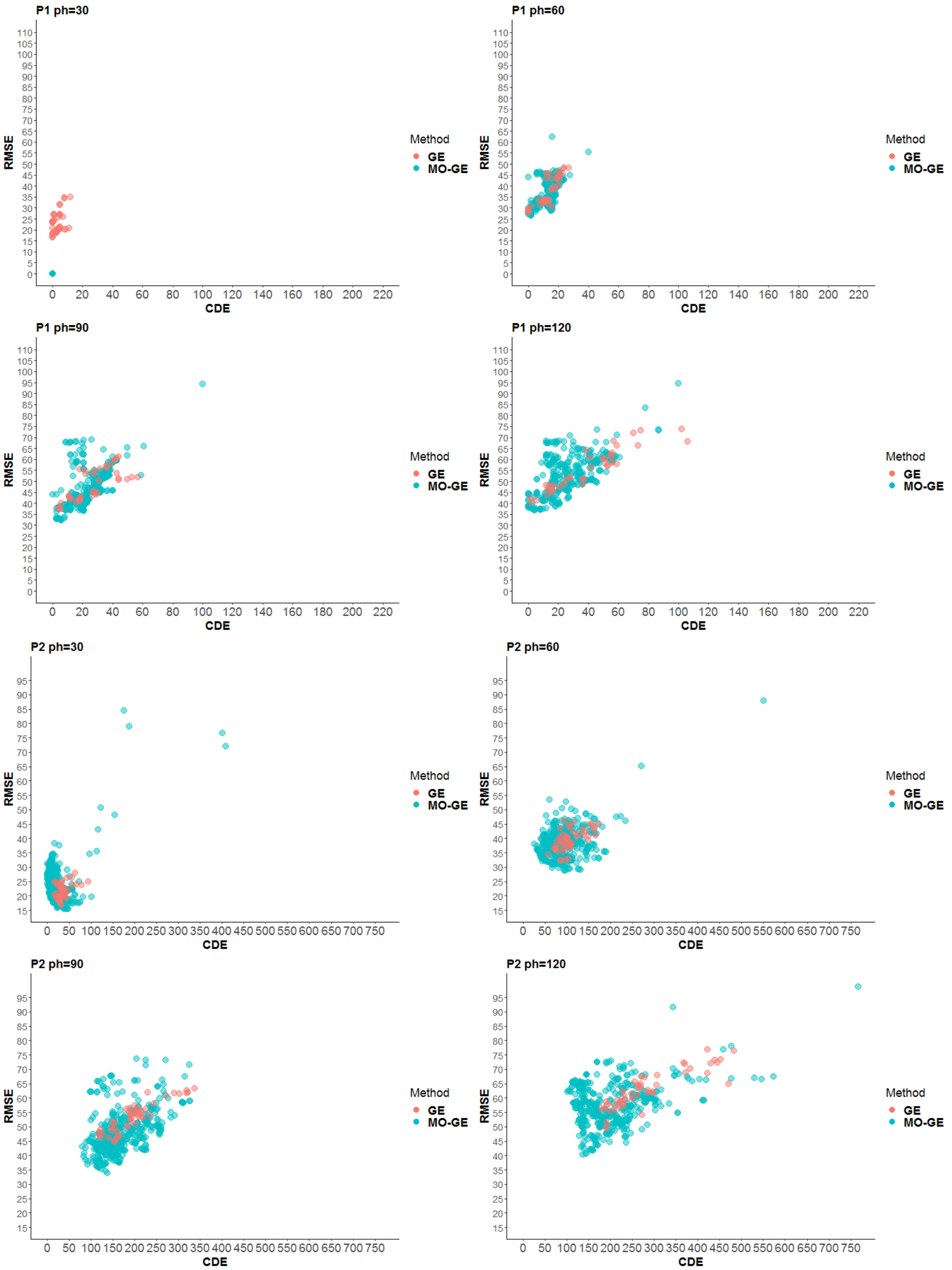


Figure C.1: Comparison between GE and MO-GE (red and green points) for patients 1 and 2 in the *What-if* scenario for all prediction horizons (30 and 90 min in left column, and 60 and 120 min in right column) and both historical values.

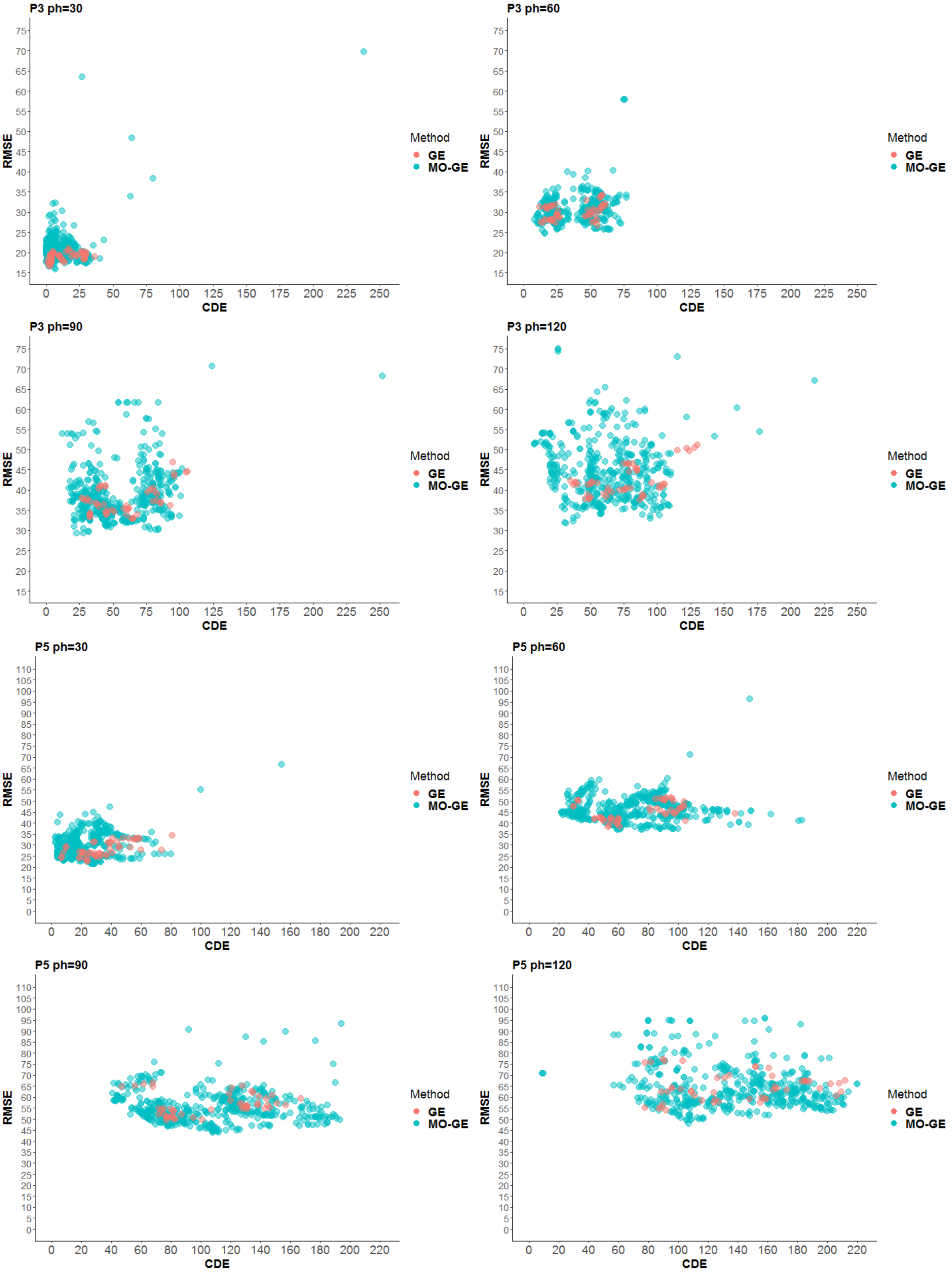


Figure C.2: Comparison between GE and MO-GE (red and green points) for patients 3 and 5 in the *What-if* scenario for all prediction horizons (30 and 90 min in left column, and 60 and 120 min in right column) and both historical values.

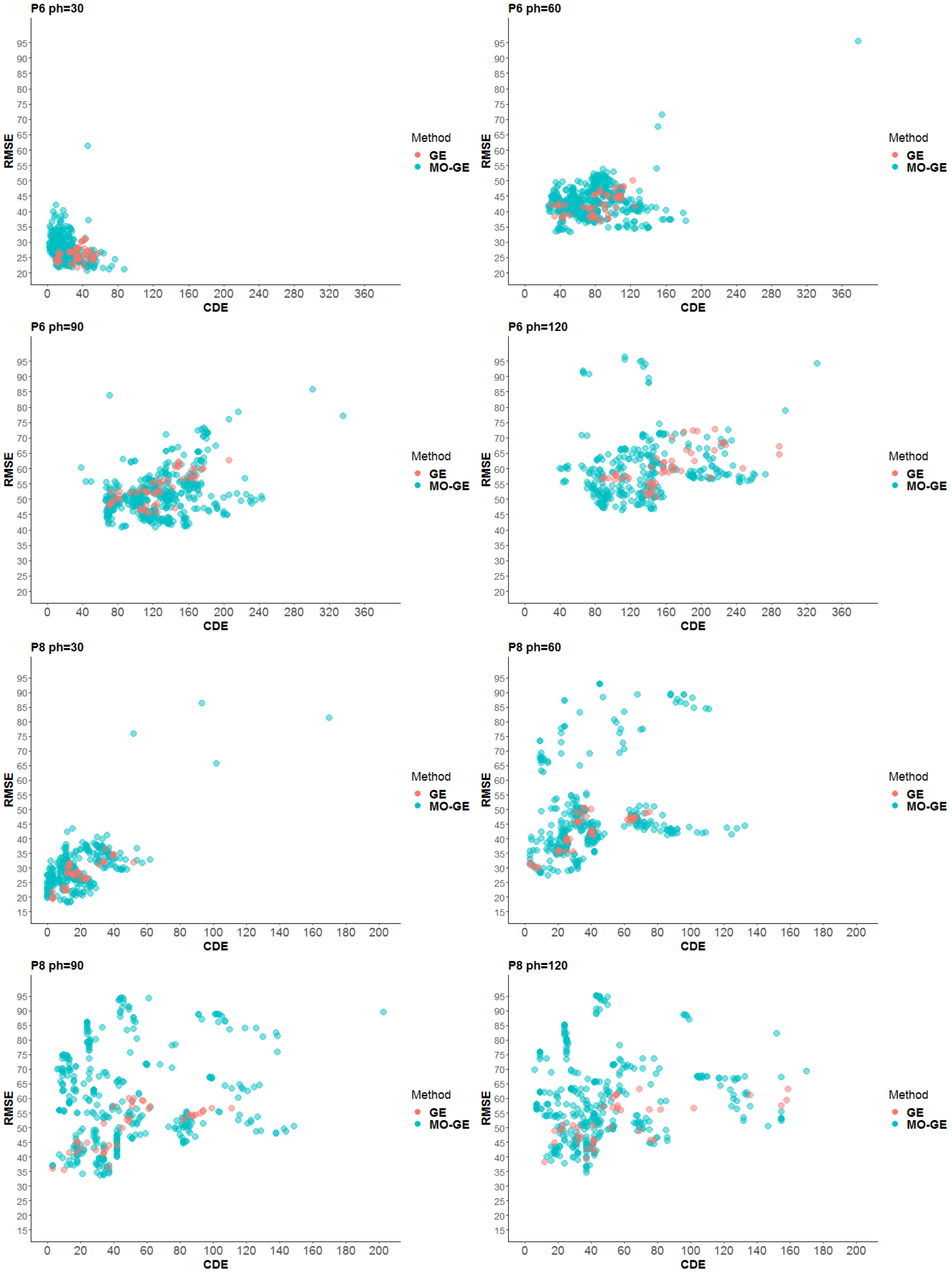


Figure C.3: Comparison between GE and MO-GE (red and green points) for patients 6 and 8 in the *What-if* scenario for all prediction horizons (30 and 90 min in left column, and 60 and 120 min in right column) and both historical values.

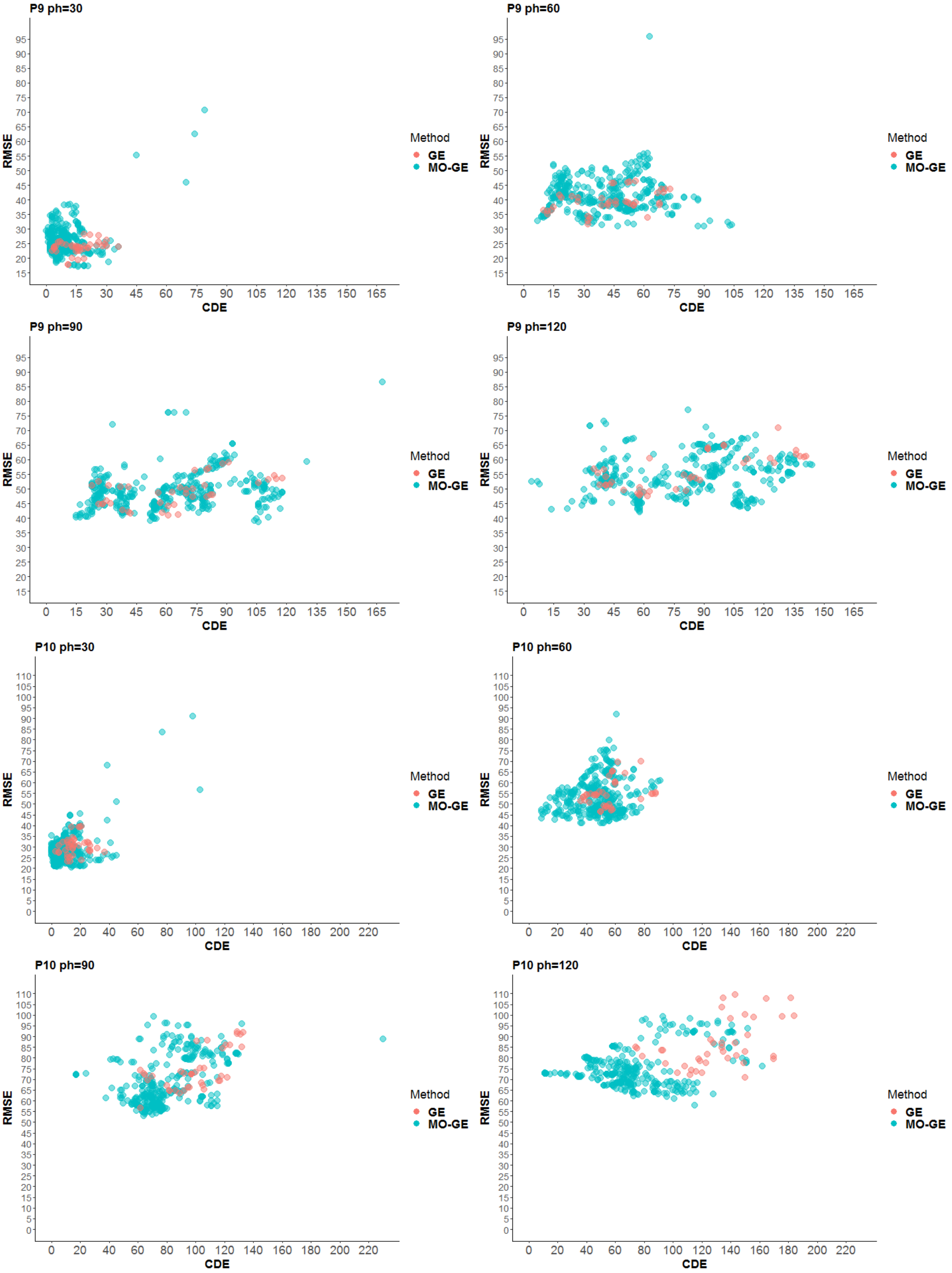


Figure C.4: Comparison between GE and MO-GE (red and green points) for patients 9 and 10 in the *What-if* scenario for all prediction horizons (30 and 90 min in left column, and 60 and 120 min in right column) and both historical values.

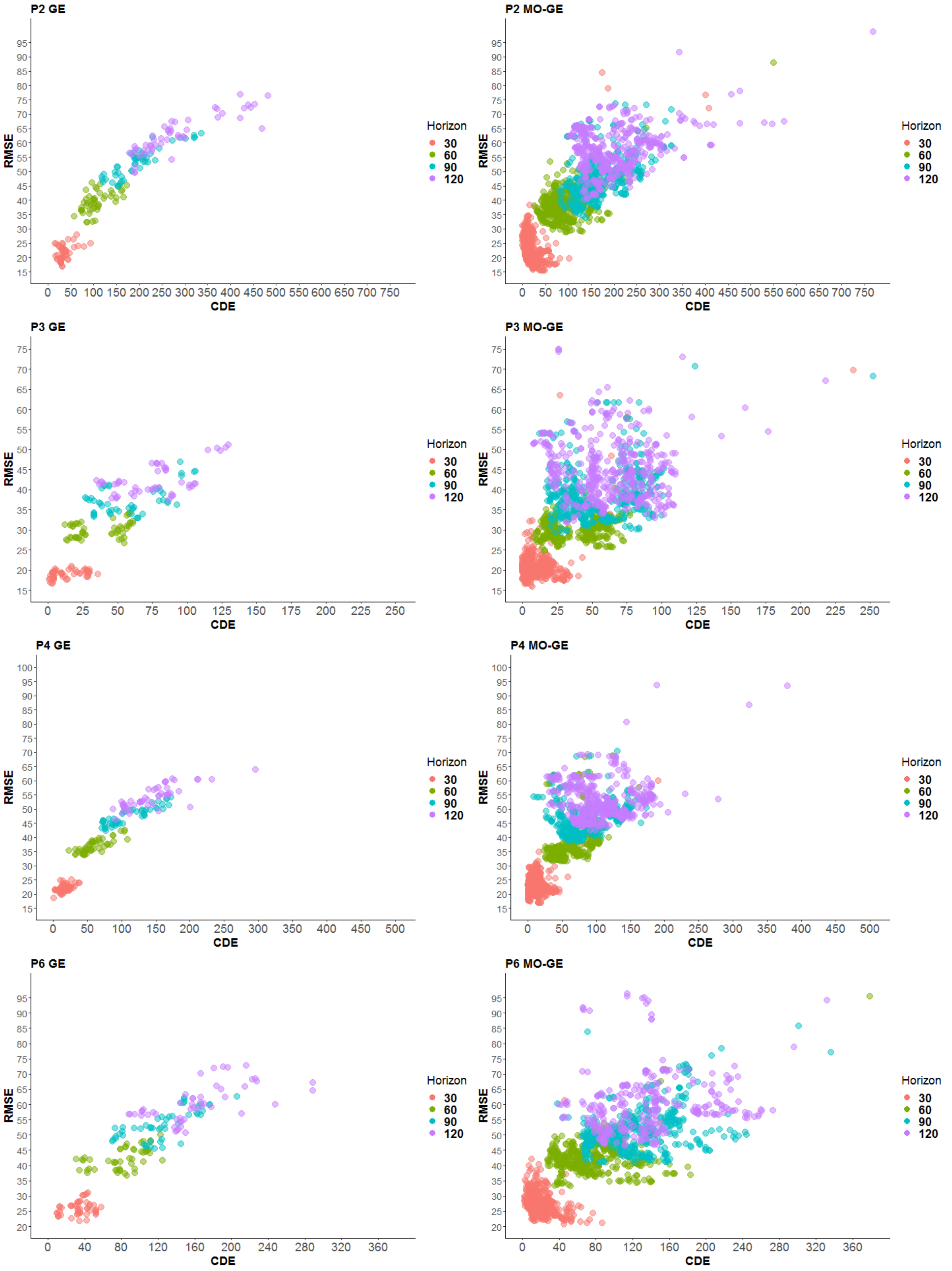


Figure C.5: Root Mean Square Error and CDE values for all prediction horizons (30 min in red points, 60 min in green points, 90 min in blue points and 120 min in purple points) for patients 2, 3, 4, and 6 (rows one, two and three) in the *What-if* scenario for GE and MO-GE (left and right column) and both historical values.

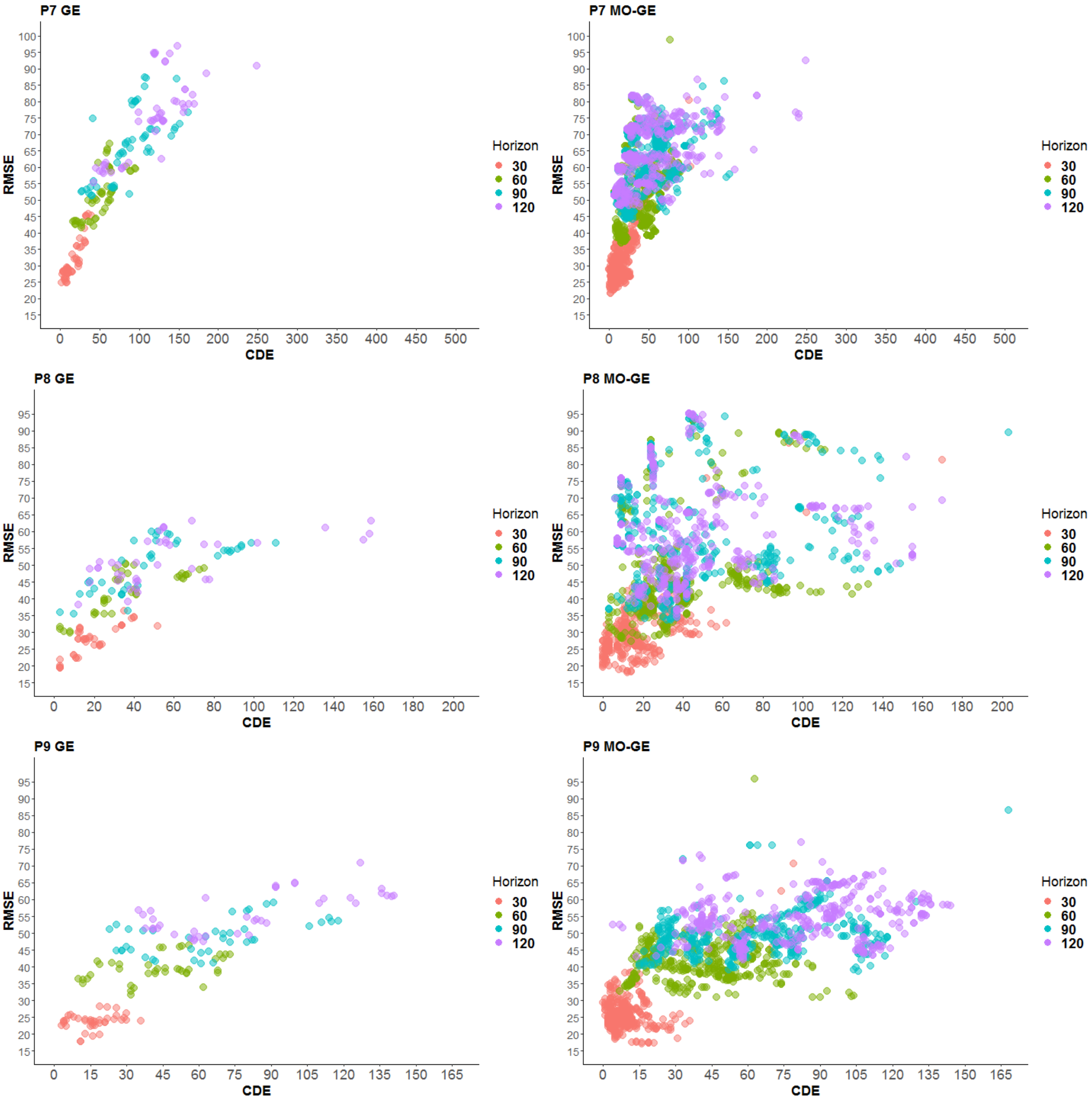


Figure C.6: Root Mean Square Error and CDE values for all prediction horizons (30 min in red points, 60 min in green points, 90 min in blue points and 120 min in purple points) for patients 7, 8 and 9 (rows one, two and three) in the *What-if* scenario for GE and MO-GE (left and right column) and both historical values.

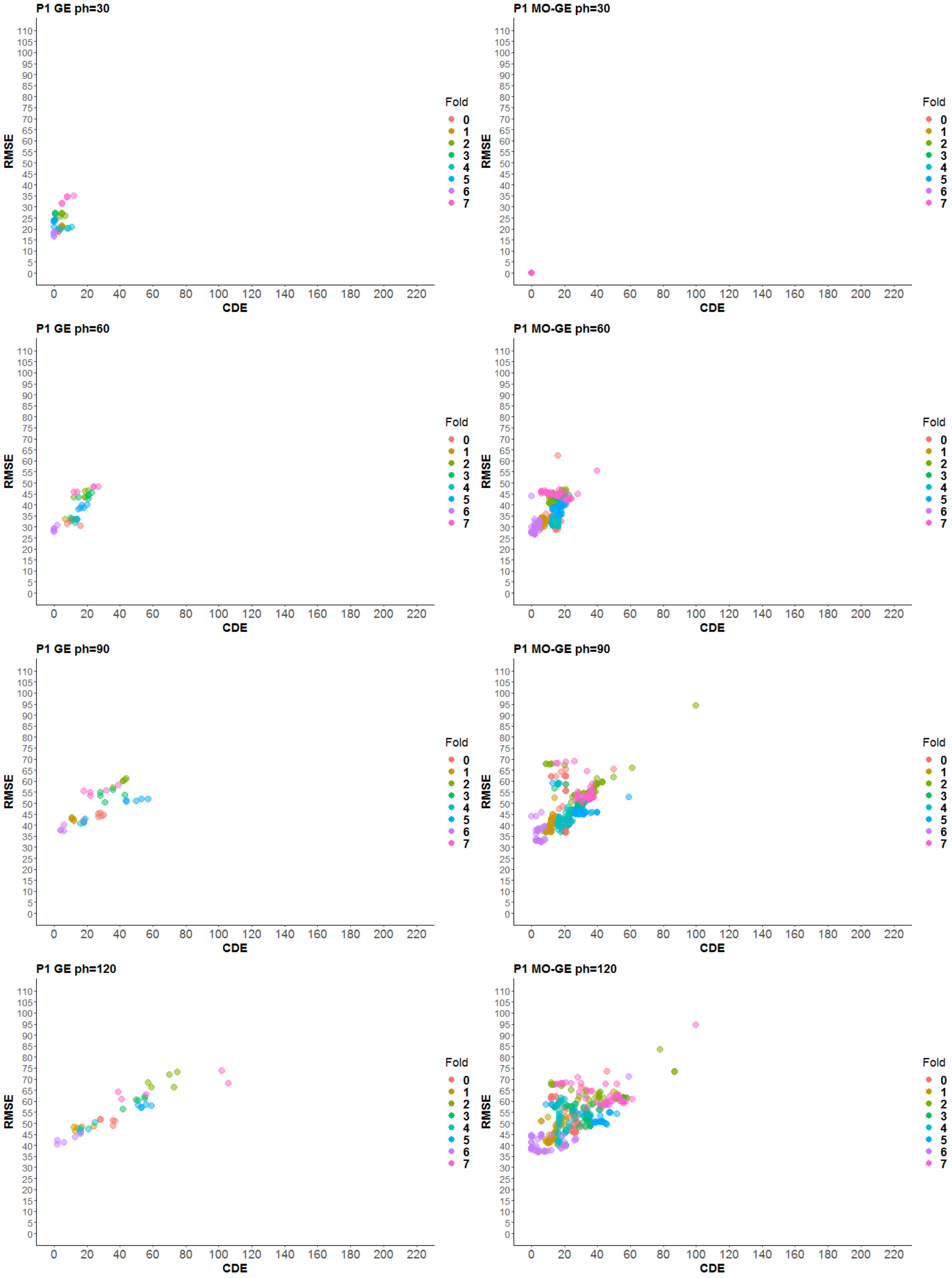


Figure C.7: All folds (each color represents a different fold) for patient 1 in the *What-if* scenario for GE and MO-GE (left and right column) and all prediction horizons (row one for 30 min, row two for 60 min, row three for 90 min and row four for 120 min) and both historical values.

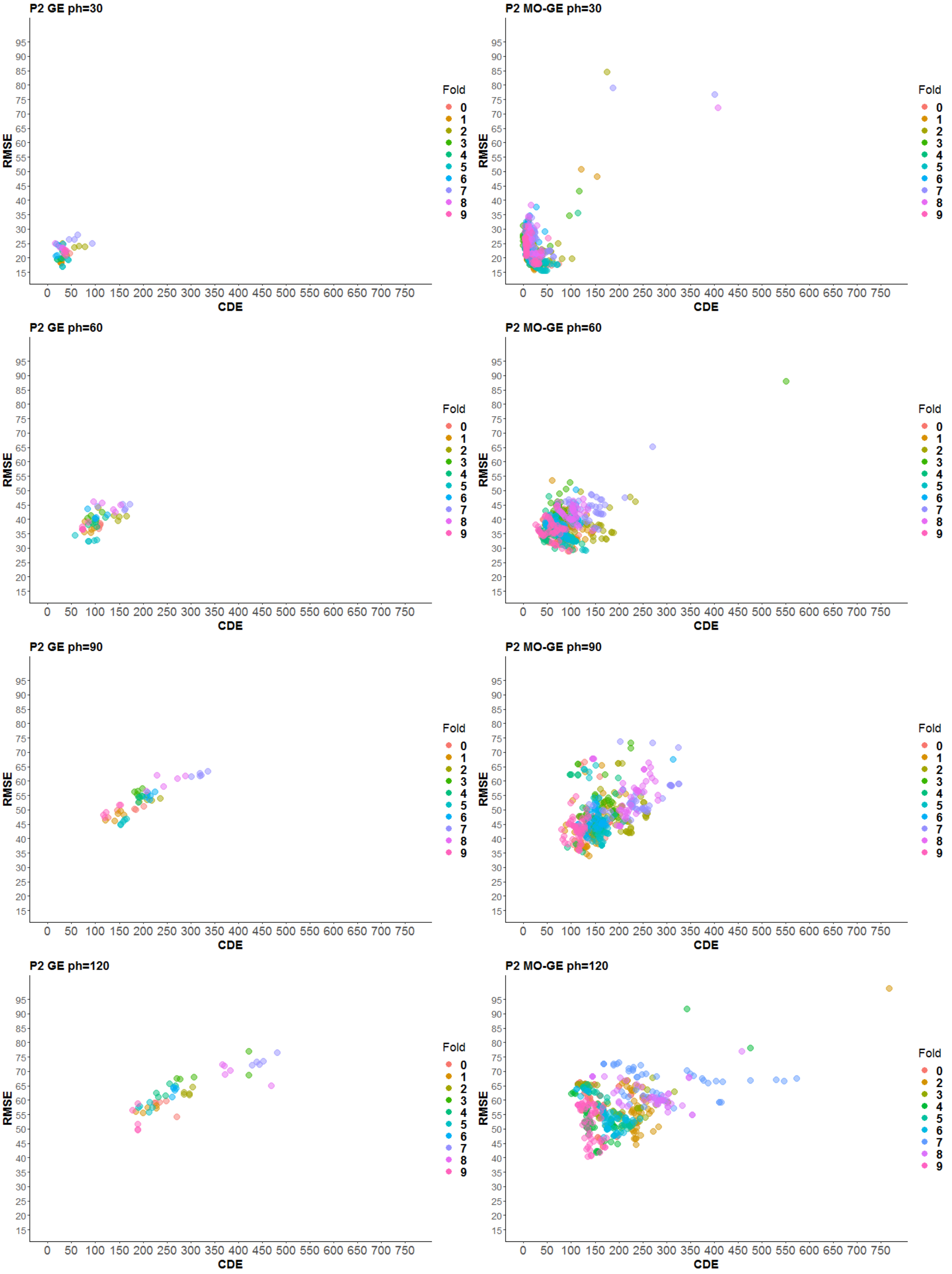


Figure C.8: All folds (each color represents a different fold) for patient 2 in the *What-if* scenario for GE and MO-GE (left and right column) and all prediction horizons (row one for 30 min, row two for 60 min, row three for 90 min and row four for 120 min) and both historical values.

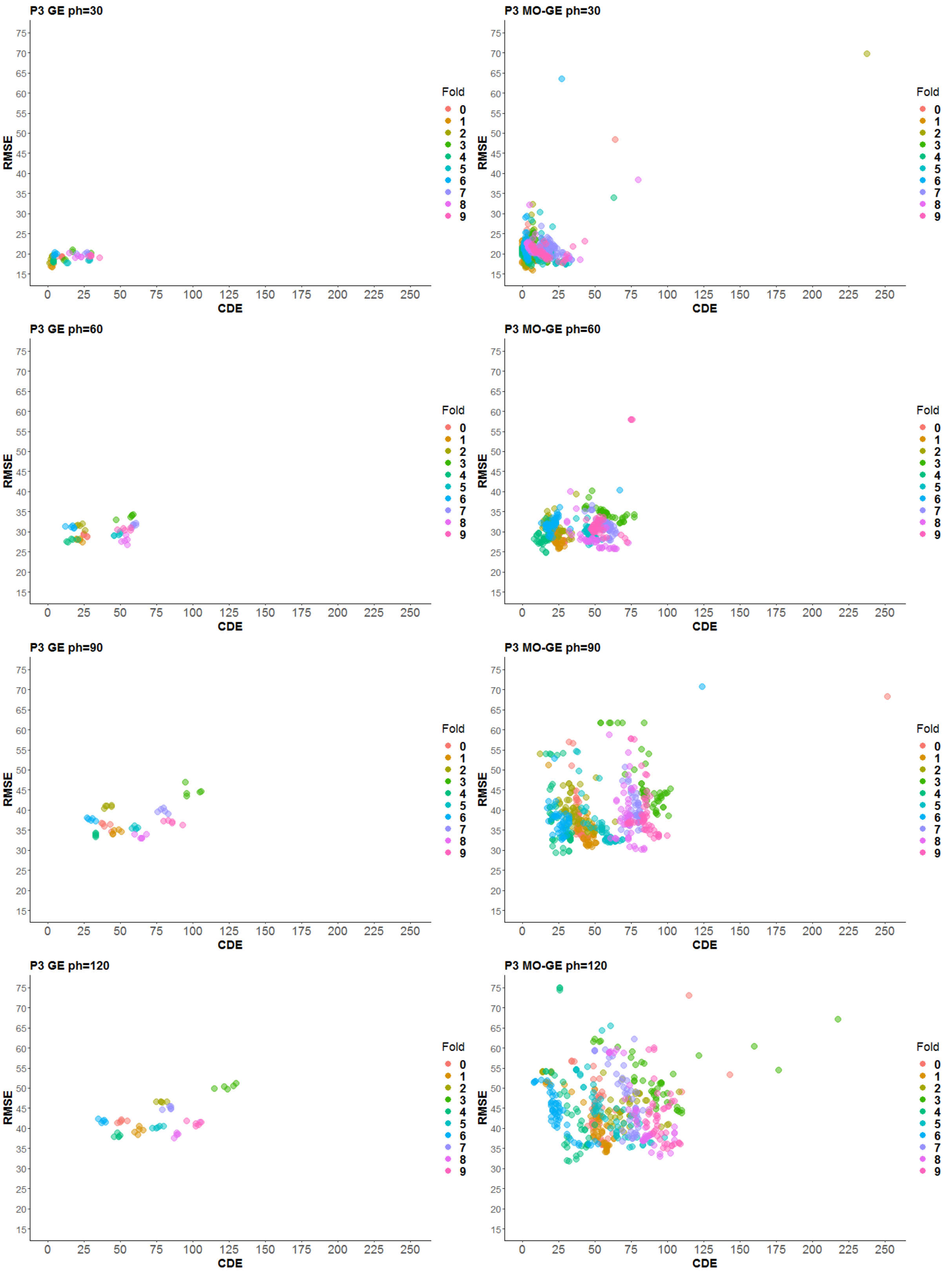


Figure C.9: All folds (each color represents a different fold) for patient 3 in the *What-if* scenario for GE and MO-GE (left and right column) and all prediction horizons (row one for 30 min, row two for 60 min, row three for 90 min and row four for 120 min) and both historical values.

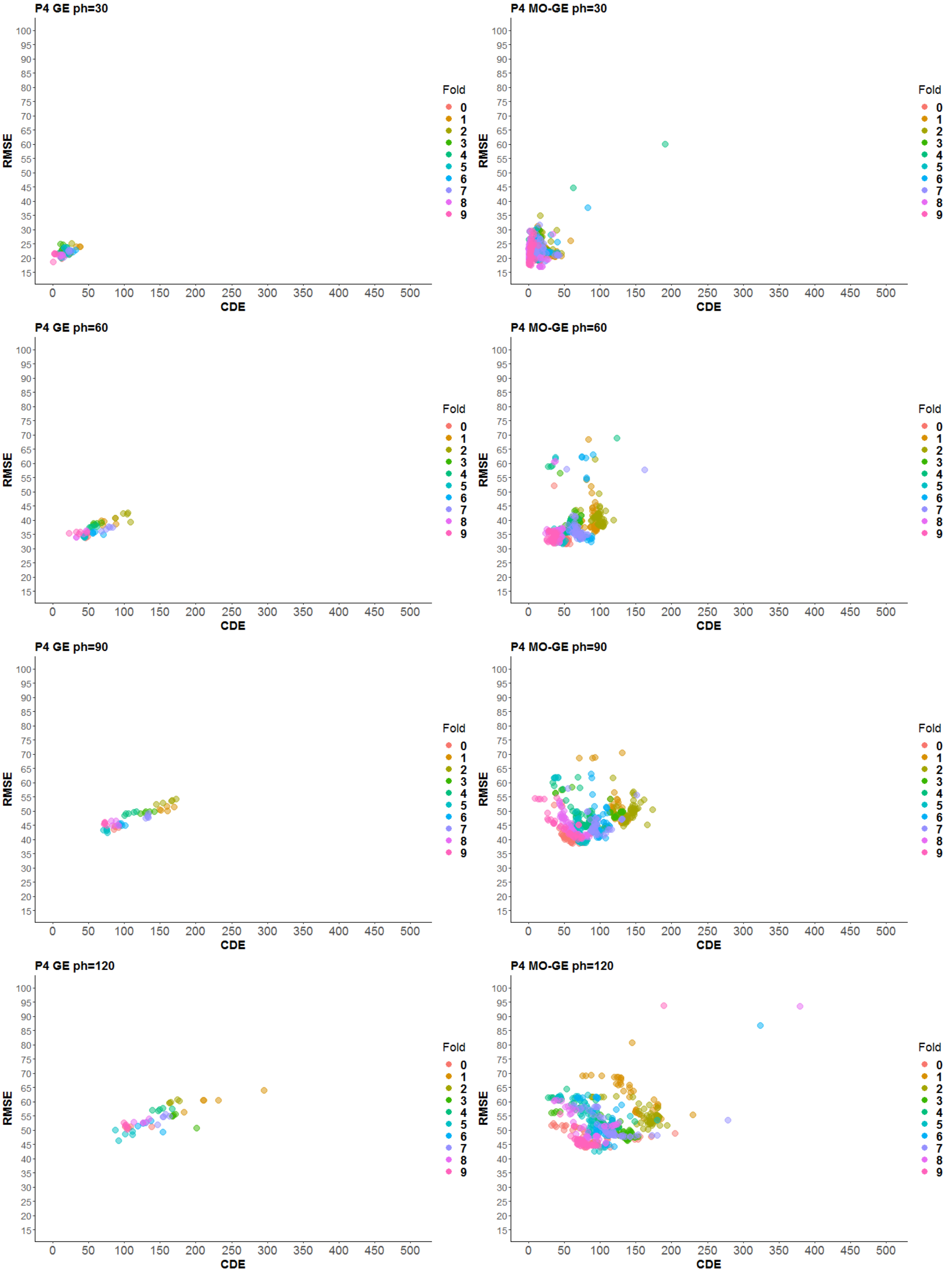


Figure C.10: All folds (each color represents a different fold) for patient 4 in the *What-if* scenario for GE and MO-GE (left and right column) and all prediction horizons (row one for 30 min, row two for 60 min, row three for 90 min and row four for 120 min) and both historical values.

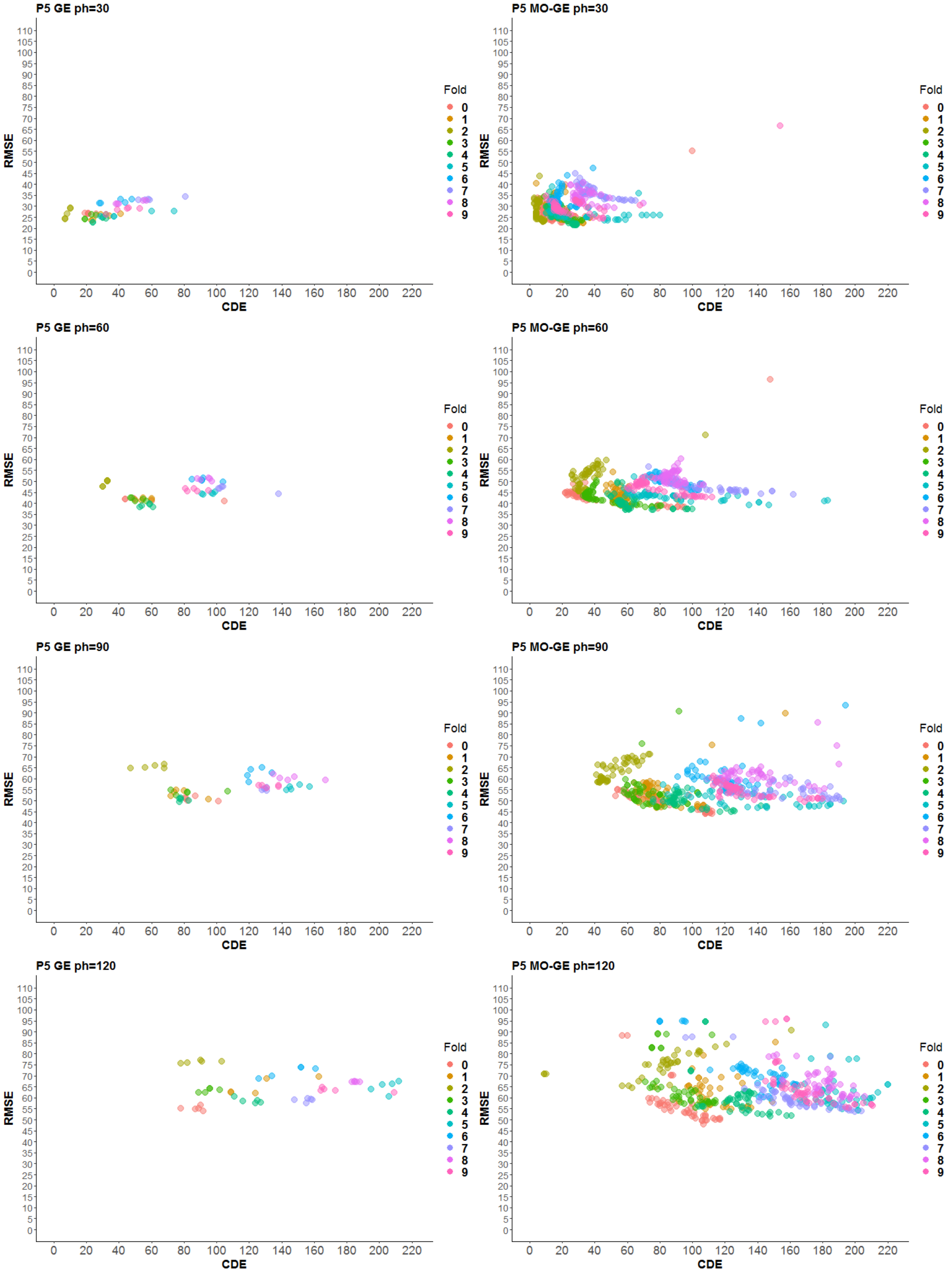


Figure C.11: All folds (each color represents a different fold) for patient 5 in the *What-if* scenario for GE and MO-GE (left and right column) and all prediction horizons (row one for 30 min, row two for 60 min, row three for 90 min and row four for 120 min) and both historical values.

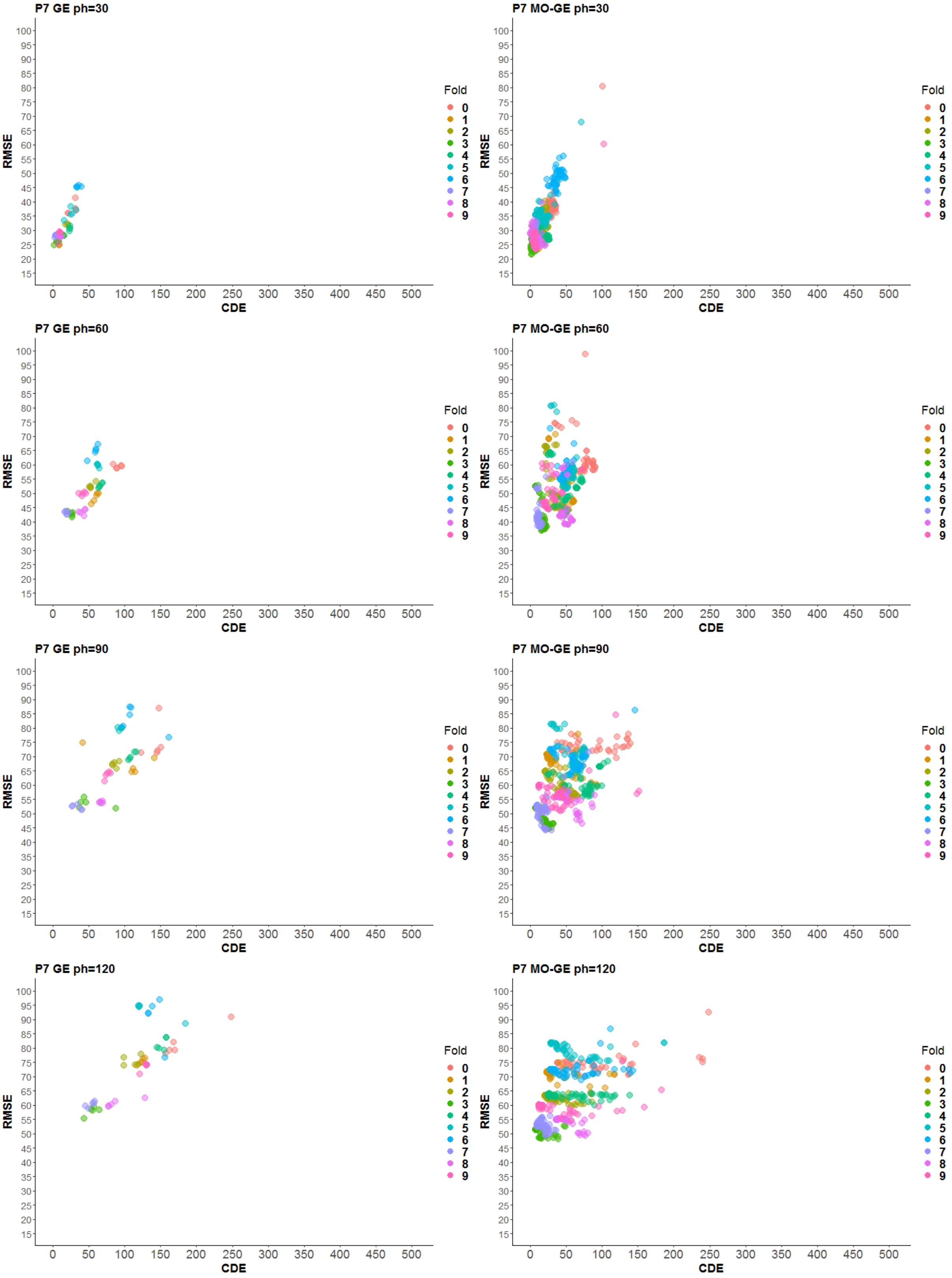


Figure C.12: All folds (each color represents a different fold) for patient 7 in the *What-if* scenario for GE and MO-GE (left and right column) and all prediction horizons (row one for 30 min, row two for 60 min, row three for 90 min and row four for 120 min) and both historical values.

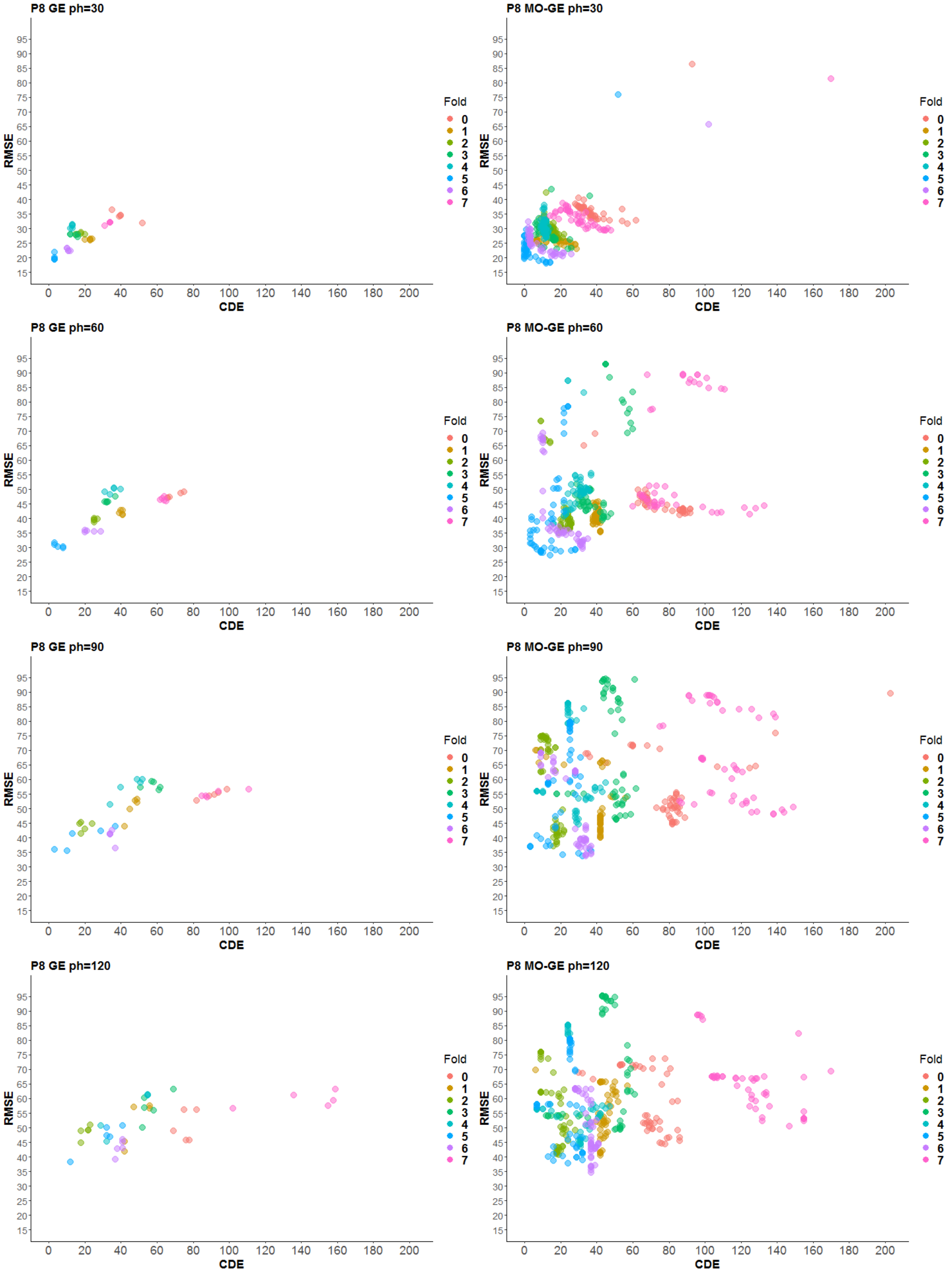


Figure C.13: All folds (each color represents a different fold) for patient 8 in the *What-if* scenario for GE and MO-GE (left and right column) and all prediction horizons (row one for 30 min, row two for 60 min, row three for 90 min and row four for 120 min) and both historical values.

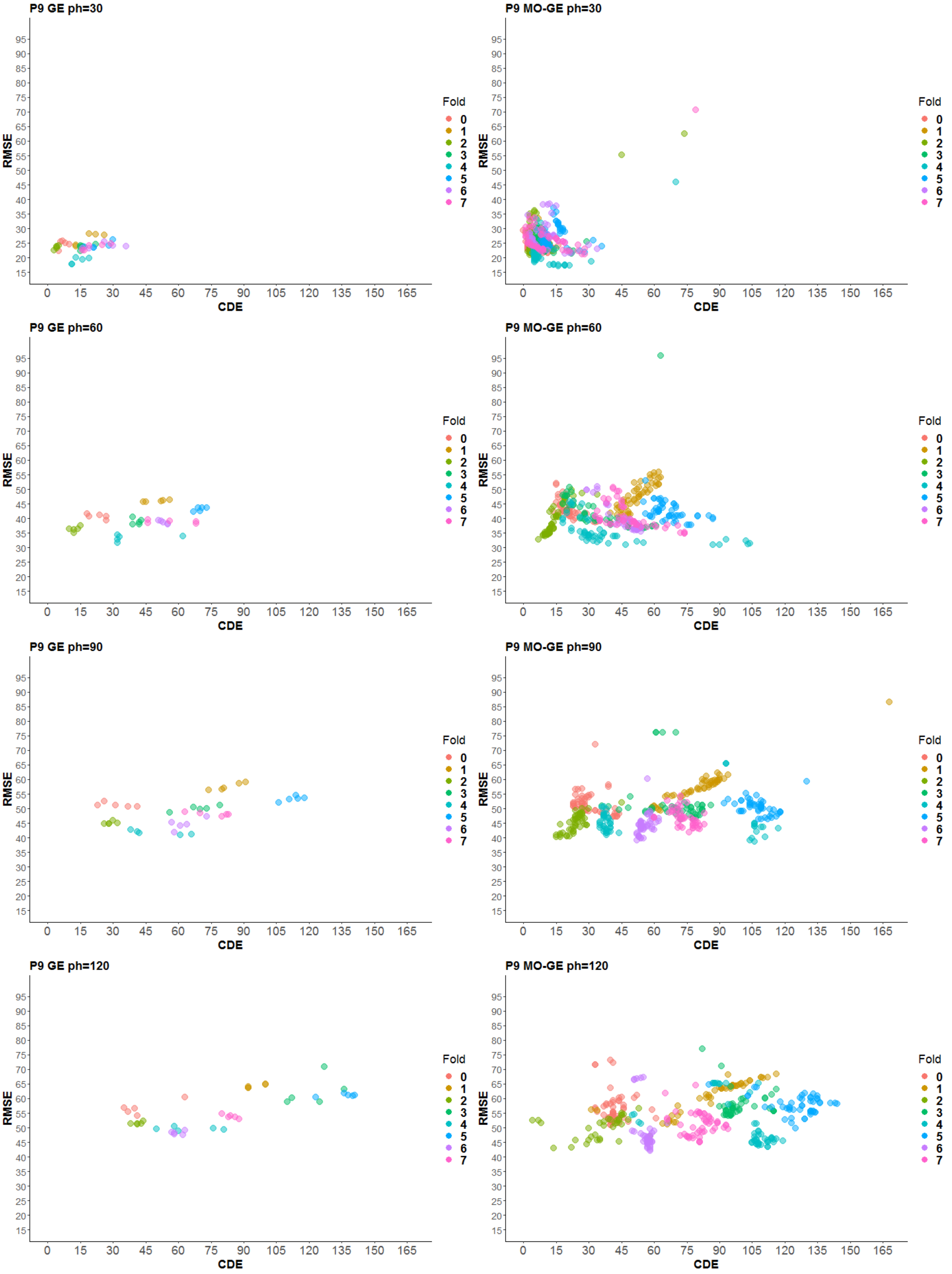


Figure C.14: All folds (each color represents a different fold) for patient 9 in the *What-if* scenario for GE and MO-GE (left and right column) and all prediction horizons (row one for 30 min, row two for 60 min, row three for 90 min and row four for 120 min) and both historical values.

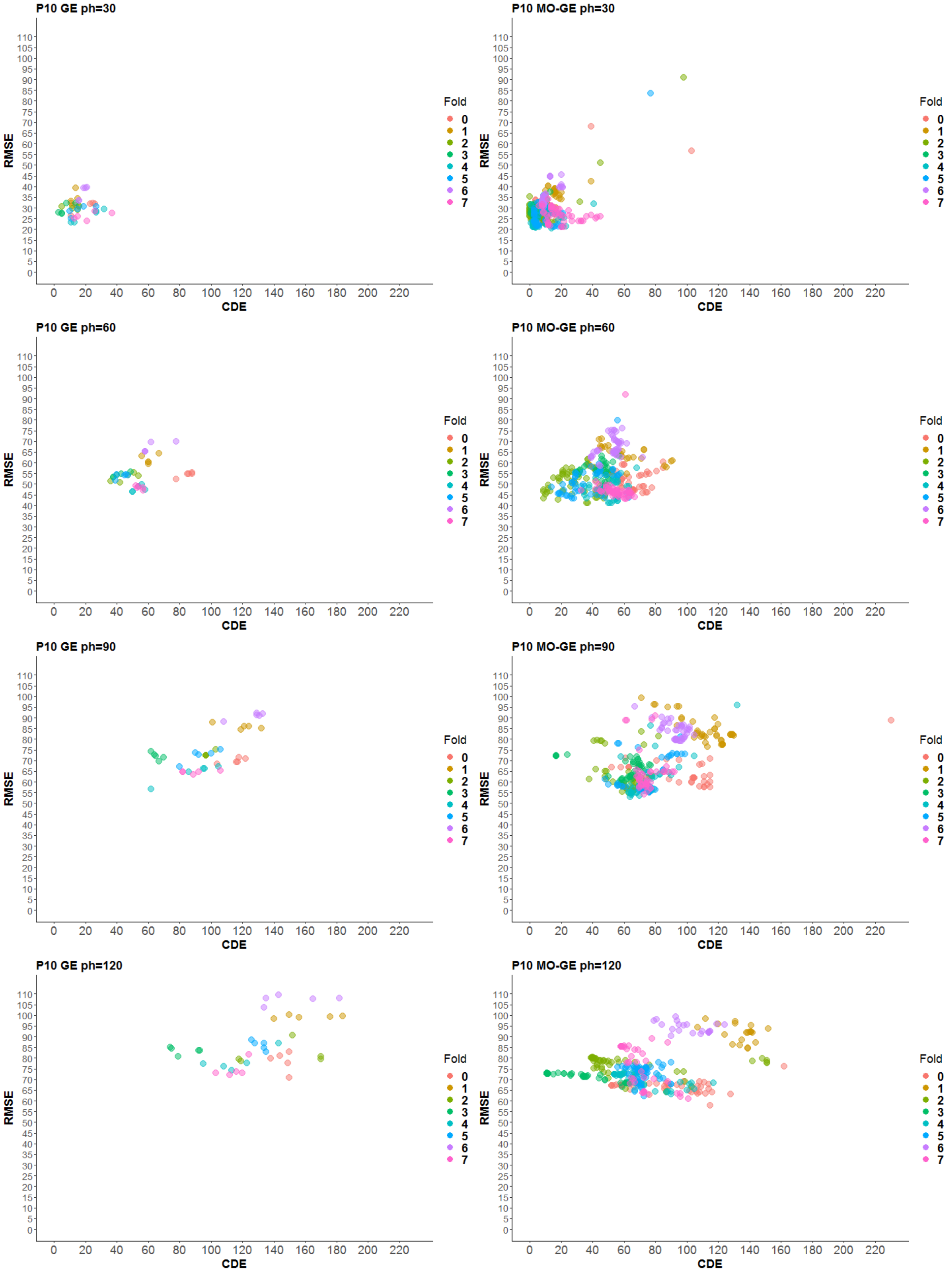


Figure C.15: All folds (each color represents a different fold) for patient 10 in the *What-if* scenario for GE and MO-GE (left and right column) and all prediction horizons (row one for 30 min, row two for 60 min, row three for 90 min and row four for 120 min) and both historical values.

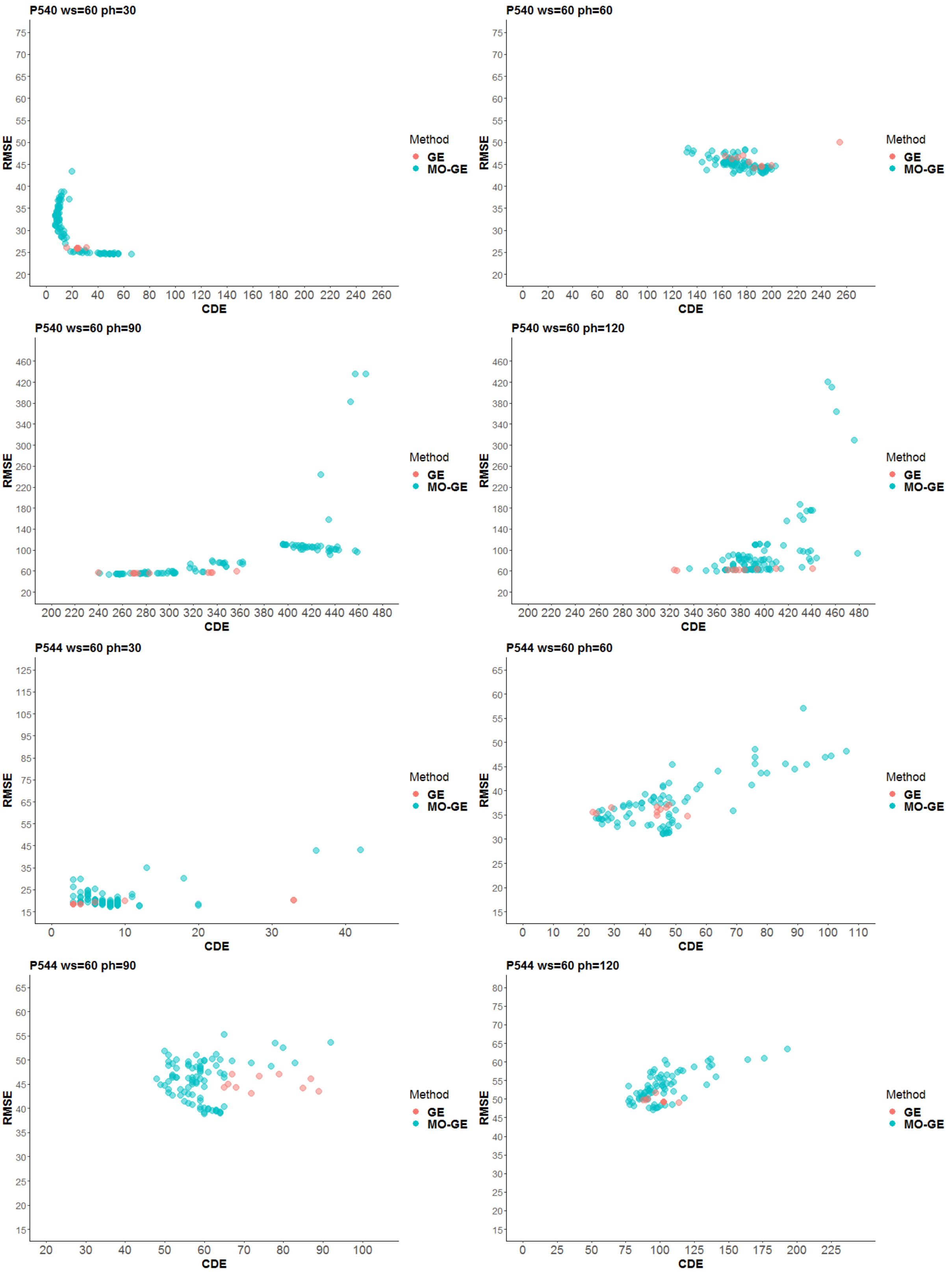


Figure C.16: GE vs. MO-GE (red and green dots) for patients 540 and 544 (rows 1,2 and 3,4) in the *Agnostic* scenario for all prediction horizons (30 and 90 min in left column, and 60 and 120 min in right column) and historical values of 60 min.

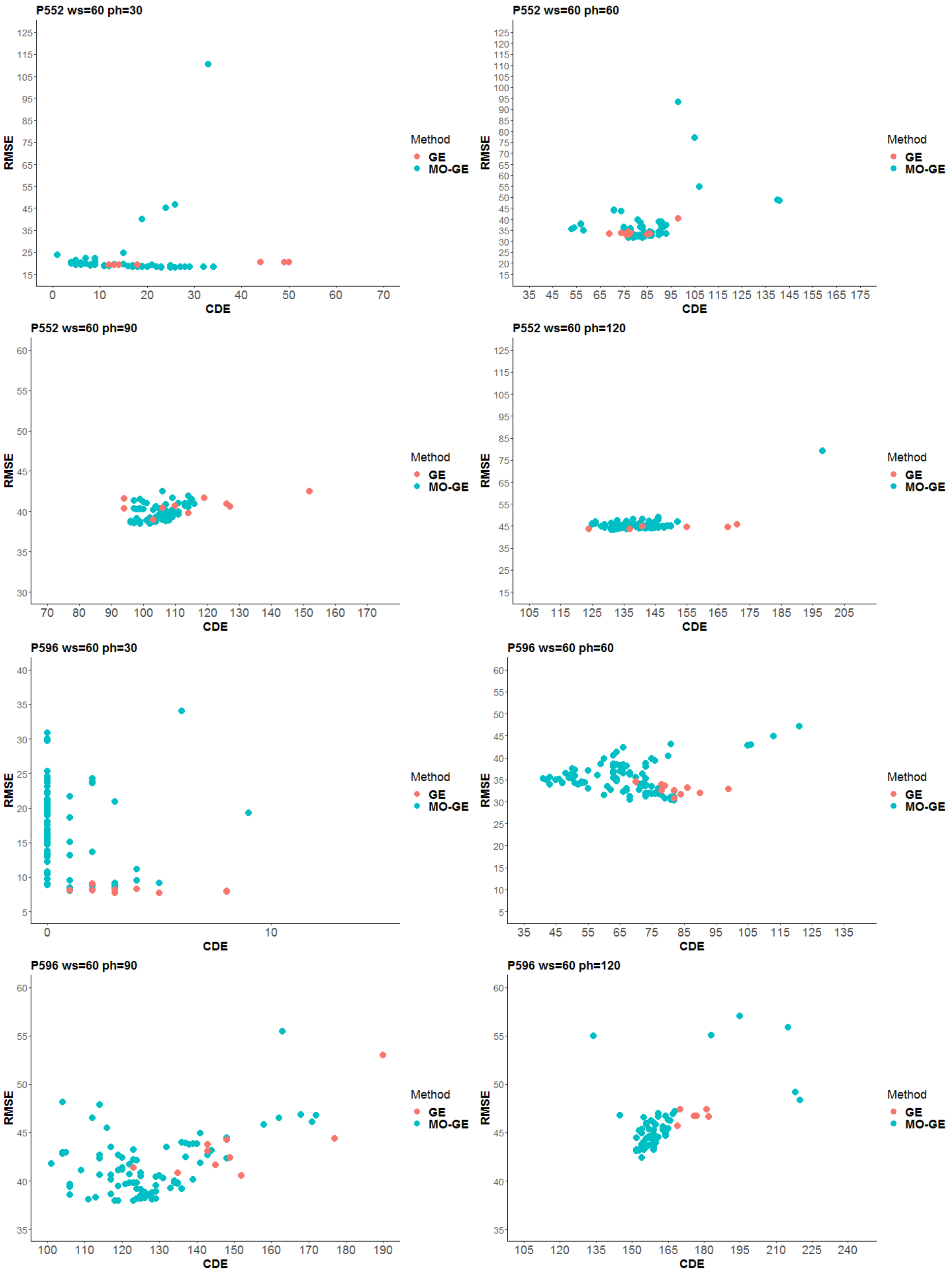


Figure C.17: GE vs. MO-GE (red and green dots) for patients 552 and 596 (rows 1,2 and 3,4) in the *Agnostic* scenario for all prediction horizons (30 and 90 min in left column, and 60 and 120 min in right column) and historical values of 60 min.

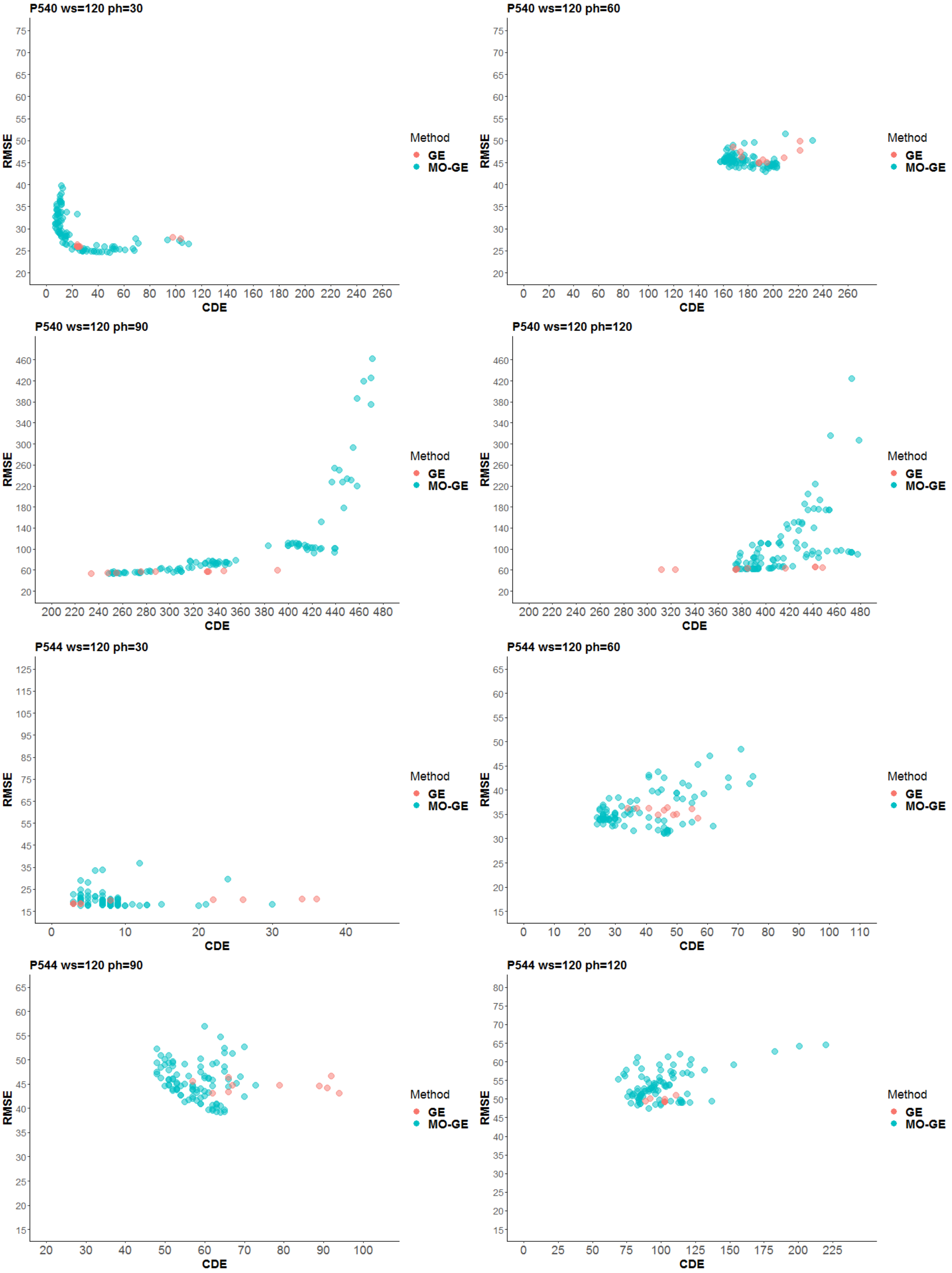


Figure C.18: GE vs. MO-GE (red and green dots) for patients 540 and 544 (rows 1,2 and 3,4) in the *Agnostic* scenario for all prediction horizons (30 and 90 min in left column, and 60 and 120 min in right column) and historical values of 120 min.

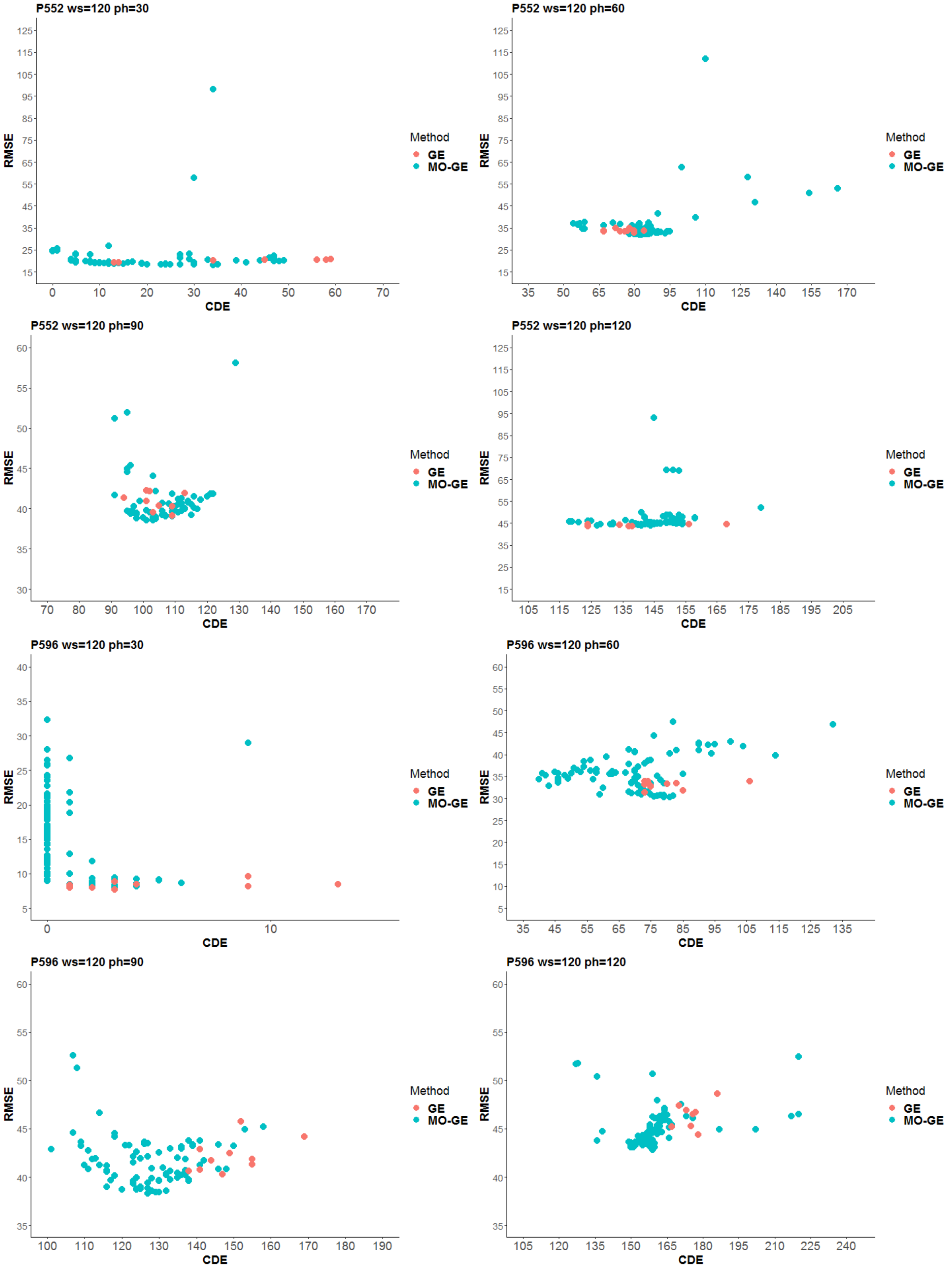


Figure C.19: GE vs. MO-GE (red and green dots) for patients 552 and 596 (rows 1,2 and 3,4) in the *Agnostic* scenario for all prediction horizons (30 and 90 min in left column, and 60 and 120 min in right column) and historical values of 120 min.

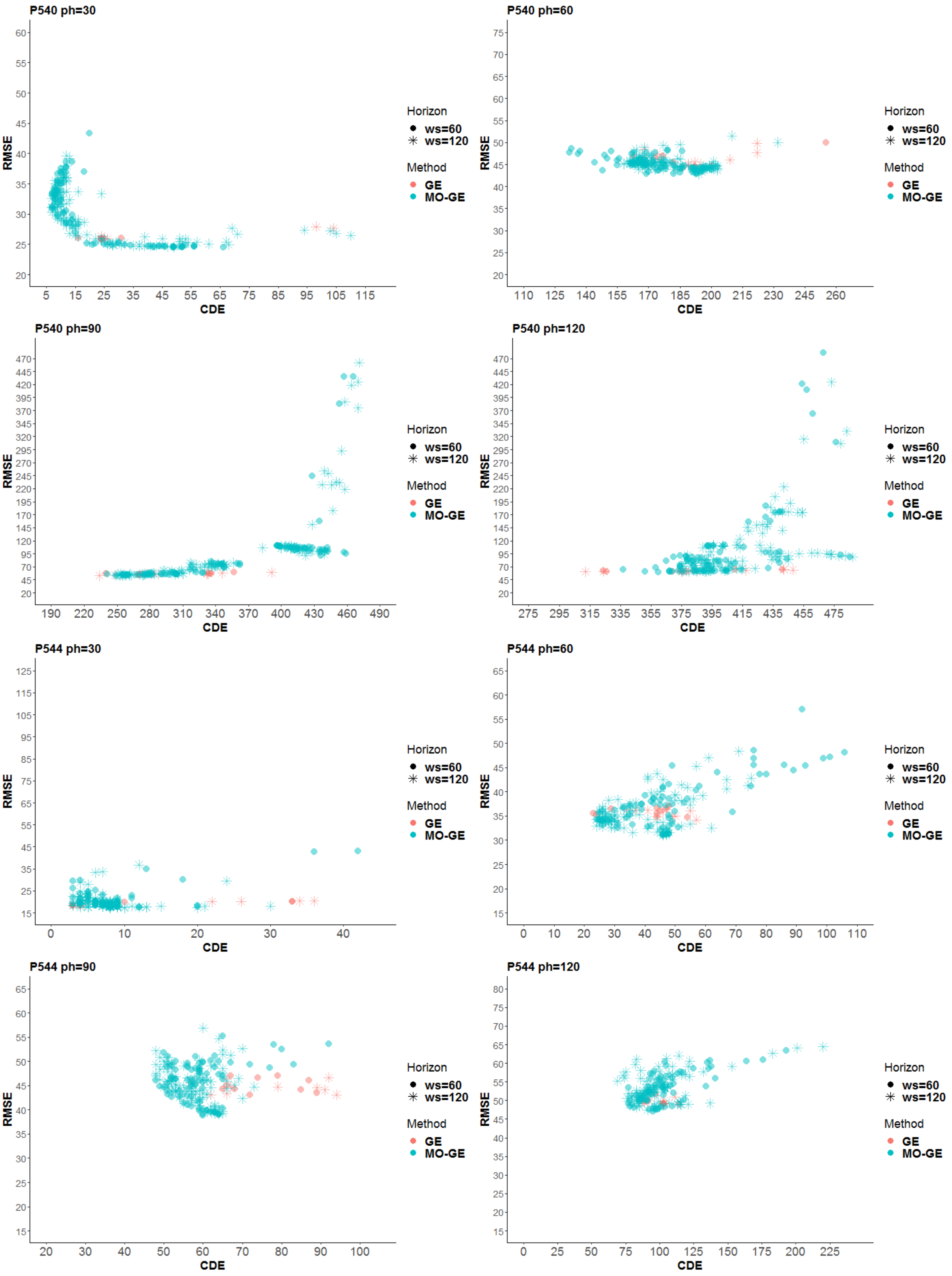


Figure C.20: Solutions coming from historical values (60 min in dot shape and 120 min in star shape) for patients 540 and 544 in the *Agnostic* scenario for all prediction horizons (30 and 90 min in left column, and 60 and 120 min in right column) and both methods GE and MO-GE (red and green dots).

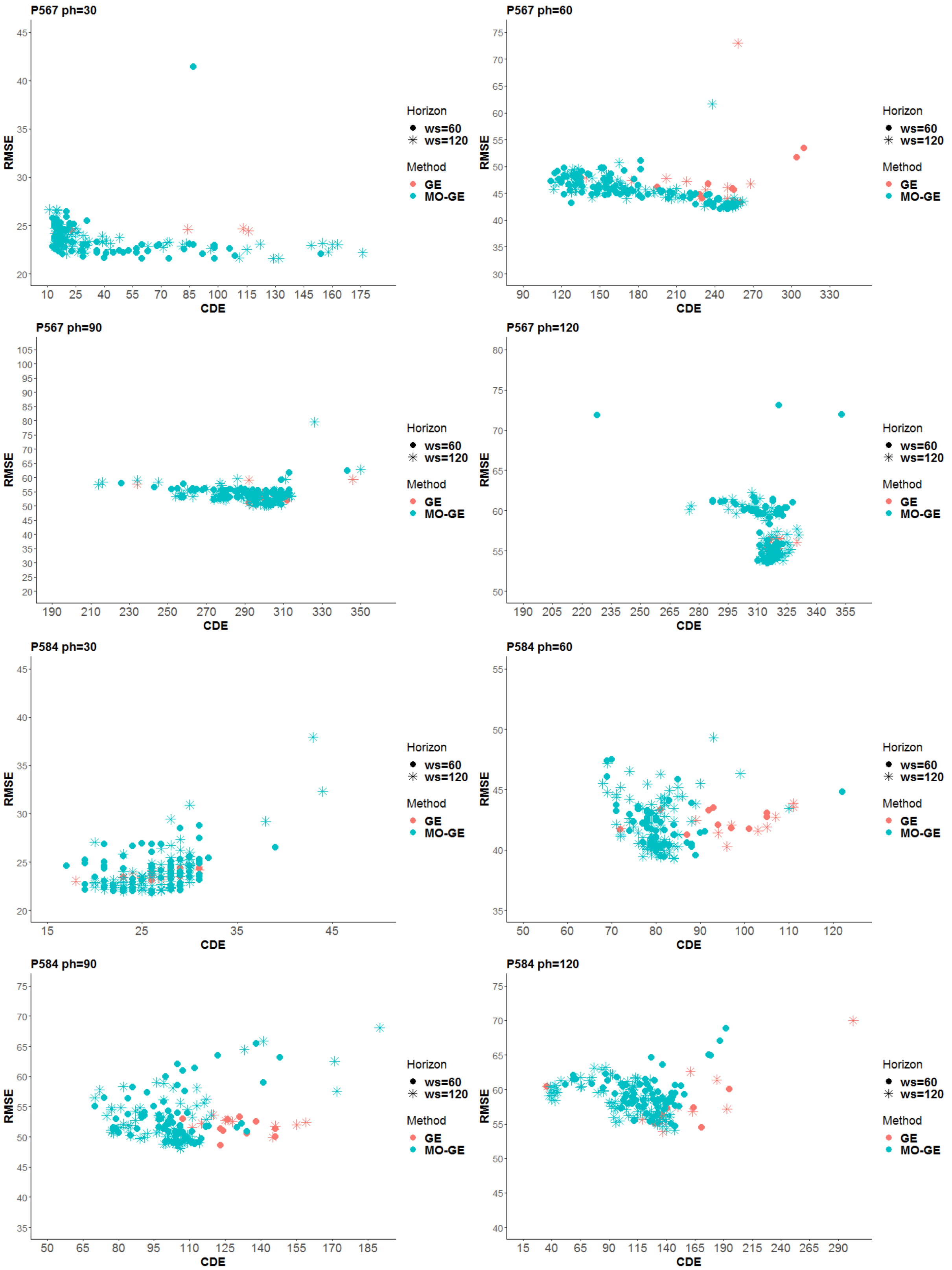


Figure C.21: Solutions coming from historical values (60 min in dot shape and 120 min in star shape) for patients 567 and 584 in the *Agnostic* scenario for all prediction horizons (30 and 90 min in left column, and 60 and 120 min in right column) and both methods GE and MO-GE (red and green dots).

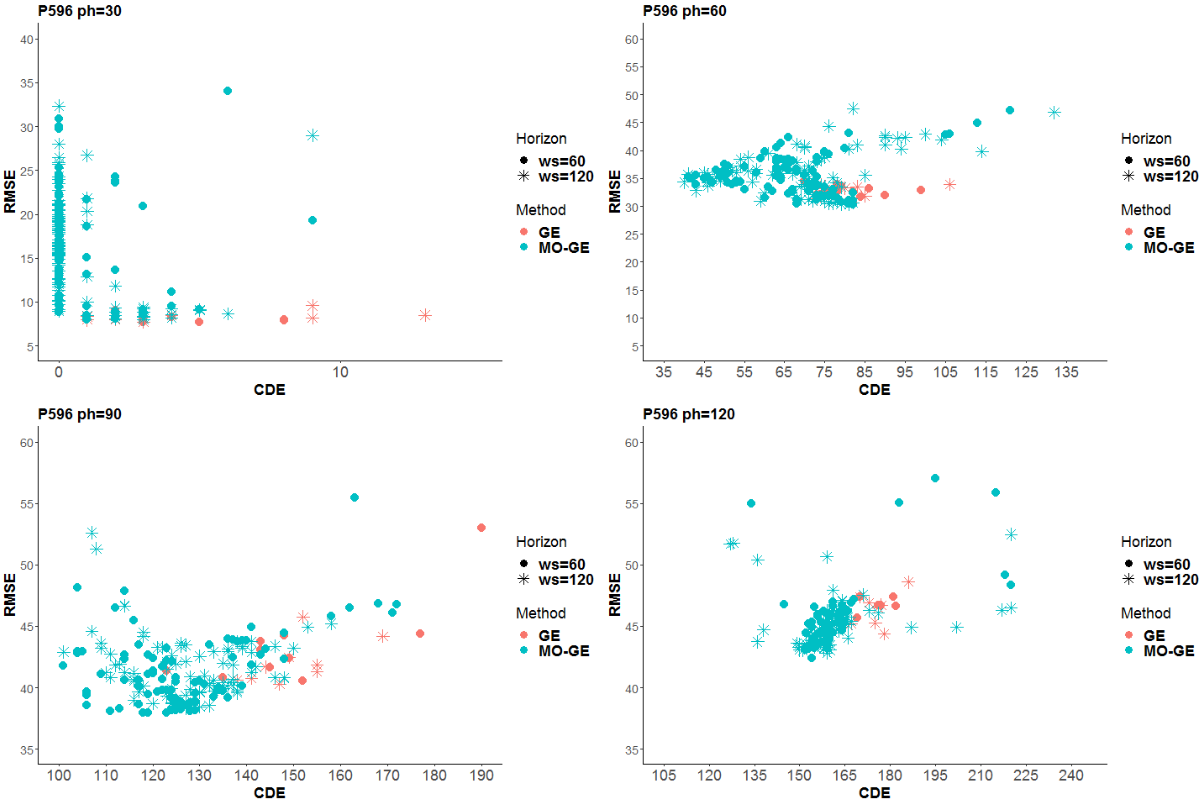


Figure C.22: Solutions coming from historical values (60 min in dot shape and 120 min in star shape) for patient 596 in the *Agnostic* scenario for all prediction horizons (30 and 90 min in left column, and 60 and 120 min in right column) and both methods GE and MO-GE (red and green dots).

Bibliography

[1] F. J. Lozano, J. I. Hidalgo, M. Botella, S. Contador, J. Lanchares, J. M. Velasco, and O. Garnica, "Identification of blood glucose patterns through continuous glucose monitoring sensors and decision trees," *medRxiv*, 2020. 23, 52

[2] B. Calvo and G. Santafé Rodrigo, "Scmamp: statistical comparison of multiple algorithms in multiple problems," *The R Journal*, vol. 8/1, Aug. 2016. 24, 85, 86, 107

[3] J. Demvsar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, Jan. 2006. 24, 79, 85, 86

[4] B. Calvo, O. M. Shir, J. Ceberio, C. Doerr, H. Wang, T. Back, and J. A. Lozano, "Bayesian performance analysis for black-box optimization benchmarking," in *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 1789–1797, 2019. 24, 25, 79, 87, 107, 108

[5] G. Kaplan, D. E Goldberg, S. Everson-Rose, R. D Cohen, R. Salonen, J. Tuomilehto, and J. Salonen, "Perceived health status and morbidity and mortality: evidence from the kuopio ischaemic heart disease risk factor study," *Journal of Epidemiology*, vol. 25, 5 1996. 35

[6] WHO, "Global report on diabetes," *World Health Organization*, 2016. 35

[7] A. D. Association, "Intensive diabetes treatment and cardiovascular outcomes in type 1 diabetes: the dcct/edic study 30-year follow-up," *Diabetes Care*, 2016. 35

[8] Y.-X. Tao, *Glucose homeostatis and the pathogenesis of diabetes mellitus.* Progress in Molecular Biology and Translational Science 121, Academic Press, 1 ed., 2014. 35

[9] G. Sparacino, F. Zanderigo, S. Corazza, A. Maran, A. Facchinetti, and C. Cobelli, "Glucose concentration can be predicted ahead in time from continuous glucose monitoring sensor time-series," *Transactions on Biomedical Engineering*, vol. 54, pp. 931–937, May. 2007. 36, 43

[10] K. R Feingold, B. Anawalt, A. Boyce, and et al., *Endotext: comprehensive free online endocrinology book.* Endotext, 2000. 39

[11] S. Suh and J. H. Kim, "Glycemic variability: how do we measure it and why is it important?," *Diabetes and Metabolism Journal*, vol. 39, no. 4, pp. 273–282, 2015. 40

[12] S. Sevimer Tuncan, M. Uzunlulu, O. telci caklili, H. Huseyin Mutlu, and A. Oguz, "Evaluation of the glycemic fluctuation as defined as the mean amplitude of glycemic excursion in hospitalized patients with type 2 diabetes," *Cyprus Journal of Medical Science*, vol. 1, 11 2016. 40, 41

[13] S. V. S. Krishna, S. K. Kota, and K. D. Modi, "Glycemic variability: clinical implications," *Indian Journal of Endocrinology and Metabolism*, vol. 17, no. 4, p. 611, 2013. 40

[14] L. Nalysnyk, M. Hernandez-Medina, and G. Krishnarajah, "Glycaemic variability and complications in patients with diabetes mellitus: evidence from a systematic review of the literature," *Diabetes, Obesity and Metabolism*, vol. 12, no. 4, pp. 288–298, 2010. 40

[15] S. Frontoni, P. Di Bartolo, A. Avogaro, E. Bosi, G. Paolisso, and A. Ceriello, "Glucose variability: an emerging target for the treatment of diabetes mellitus," *Diabetes Research and Clinical Practice*, vol. 102, no. 2, pp. 86–95, 2013. 40

[16] J. I. Hidalgo, J. M. Colmenar, G. Kronberger, S. M. Winkler, O. Garnica, and J. Lanchares, "Data based prediction of blood glucose concentrations using evolutionary methods," *Journal of Medical Systems*, vol. 41, no. 9, p. 142, 2017. 40, 45, 91

[17] B. Hansen and I. Matytsina, "Insulin administration: selecting the appropriate needle and individualizing the injection technique," *Expert Opinion on Drug Delivery*, vol. 8, no. 10, pp. 1395–1406, 2011. 40

[18] J. Weissberg-Benchell, J. Antisdel-Lomaglio, and R. Seshadri, "Insulin pump therapy," *Diabetes Care*, vol. 26, no. 4, pp. 1079–1087, 2003. 40

[19] P. A. Bakhtiani, L. M. Zhao, J. El Youssef, J. R. Castle, and W. K. Ward, "A review of artificial pancreas technologies with an emphasis on bi-hormonal therapy," *Diabetes, Obesity and Metabolism*, vol. 15, no. 12, pp. 1065–1070, 2013. 40

[20] G. D. Nicolao, L. Magni, C. D. Man, and C. Cobelli, "Modeling and control of diabetes: towards the artificial pancreas," *International Federation of Automatic Control Proceedings Volumes*, vol. 44, no. 1, pp. 7092–7101, 2011. International Federation of Automatic Control World Congress. 41

[21] E. P. Córcoles and M. G. Boutelle, *Biosensors and invasive monitoring in clinical applications*. SpringerBriefs in Applied Sciences and Technology, 2013. 41

[22] E.-H. Yoo and S.-Y. Lee, "Glucose biosensors: an overview of use in clinical practice," *Sensors*, vol. 10, no. 5, pp. 4558–4576, 2010. 41

[23] A. H. Hansen, A. K. Duun-Henriksen, R. Juhl, S. Schmidt, K. Nørgaard, J. B. Jørgensen, and H. Madsen, "Predicting plasma glucose from interstitial glucose observations using bayesian methods," *Journal of Diabetes Science and Technology*, vol. 8, no. 2, pp. 321–330, 2014. 41

[24] P. Aaby Svendsen, T. Lauritzen, U. Sóegaard, and J. Nerup, "Glycosylated haemoglobin and steady-state mean blood glucose concentration in type 1 (insulin-dependent) diabetes," *Diabetologia*, vol. 23, pp. 403–405, Nov. 1982. 41

[25] R. D. Lasker, "The diabetes control and complications trial: implications for policy and practice," *New England Journal of Medicine*, vol. 329, no. 14, pp. 1035–1036, 1993. 41

[26] T. D. Control and C. T. R. Group, "The relationship of glycemic exposure (hba1c) to the risk of development and progression of retinopathy in the diabetes control and complications trial," *Diabetes*, vol. 44, no. 8, pp. 968–983, 1995. 41

[27] B. P. Kovatchev, D. J. Cox, A. Kumar, L. Gonder-Frederick, and W. L. Clarke, "Algorithmic evaluation of metabolic control and risk of severe hypoglycemia in type 1 and type 2 diabetes using self-monitoring blood glucose data," *Diabetes Technology and Therapeutics*, vol. 5, no. 5, pp. 817–828, 2003. 41

[28] G. Marics, Z. Lendvai, C. Lodi, L. Koncz, D. Zakarias, G. Schuster, B. Mikos, C. Hermann, A. J. Szabo, and P. Toth-Heyn, "Evaluation of an open access software for calculating glucose variability parameters of a continuous glucose monitoring system applied at pediatric intensive care unit," *BioMedical Engineering OnLine*, vol. 14, p. 37, Apr. 2015. 41

[29] D. Rodbard, "Interpretation of continuous glucose monitoring data: glycemic variability and quality of glycemic control," *Diabetes Technology and Therapeutics*, vol. 11, no. s1, pp. S–55–S–67, 2009. 41

[30] C. McDonnell, S. Donath, S. Vidmar, G. Werther, and F. Cameron, "A novel approach to continuous glucose analysis utilizing glycemic variation," *Diabetes Technology and Therapeutics*, vol. 7, no. 2, pp. 253–263, 2005. 41

[31] E. A. Ryan, T. Shandro, K. Green, B. W. Paty, P. A. Senior, D. Bigam, A. J. Shapiro, and M.-C. Vantyghem, "Assessment of the severity of hypoglycemia and glycemic lability in type 1 diabetic subjects undergoing islet transplantation," *Diabetes*, vol. 53, no. 4, pp. 955–962, 2004. 41

[32] G. D. Molnar, W. F. Taylor, and M. M. Ho, "Day-to-day variation of continuously monitored glycaemia: a further measure of diabetic instability," *Diabetologia*, vol. 8, pp. 342–348, Nov. 1972. 41

[33] S. E. Siegelaar, F. Holleman, J. B. L. Hoekstra, and J. H. DeVries, "Glucose variability; does it matter?," *Endocrine Reviews*, vol. 31, no. 2, pp. 171–182, 2010. 41, 76

[34] B. P. Kovatchev, E. Otto, D. Cox, L. Gonder-Frederick, and W. Clarke, "Evaluation of a new measure of blood glucose variability in diabetes," *Diabetes Care*, vol. 29, no. 11, pp. 2433–2438, 2006. 41

[35] N. R. Hill, P. C. Hindmarsh, R. J. Stevens, I. M. Stratton, J. C. Levy, and D. R. Matthews, "A method for assessing quality of control from glucose profiles," *Diabetic Medicine*, vol. 24, no. 7, pp. 753–758, 2007. 41

[36] C. S. Derdemezis and J. A. Lovegrove, "Glycemic index, glycemic control and beyond," *Current Pharmaceutical Design*, vol. 20, no. 22, pp. 3620–3630, 2014. 41

[37] J. Geddes, R. J. Wright, N. N. Zammitt, I. J. Deary, and B. M. Frier, "An evaluation of methods of assessing impaired awareness of hypoglycemia in type 1 diabetes," *Diabetes Care*, vol. 30, no. 7, pp. 1868–1870, 2007. 41

[38] D. Rodbard, "The challenges of measuring glycemic variability," *Journal of Diabetes Science and Technology*, vol. 6, no. 3, pp. 712–715, 2012. 41, 76

[39] S. R. Patton and M. A. Clements, "Average daily risk range as a measure for clinical research and routine care," *Journal of Diabetes Science and Technology*, vol. 7, no. 5, pp. 1370–1375, 2013. 41

[40] S. Mirshekarian, H. Shen, R. Bunescu, and C. Marling, "Lstms and neural attention models for blood glucose prediction: comparative experiments on real and synthetic data," in *Annual International Conference in Medicine and Biology*, pp. 706–712, Institute of Electrical and Electronics Engineers, 2019. 42

[41] M. I. Schmidt, A. Hadji-Georgopoulos, M. Rendell, S. Margolis, and A. Kowarski, "The dawn phenomenon, an early morning glucose rise: implications for diabetic intraday blood glucose variation," *Diabetes Care*, vol. 4, no. 6, pp. 579–585, 1981. 42

[42] S. Oviedo, J. Vehí, R. Calm, and J. Armengol, "A review of personalized blood glucose prediction strategies for t1dm patients," *Journal for Numerical Methods in Biomedical Engineering*, vol. 33, no. 6, p. 2833, 2016. 42

[43] M. Messori, C. Toffanin, S. D. Favero, G. D. Nicolao, C. Cobelli, and L. Magni, "Model individualization for artificial pancreas," *Journal of Computer Methods and Programs in Biomedicine*, vol. 171, pp. 133–140, 2016. 43

[44] C. D. Man, F. Micheletto, D. Lv, M. Breton, B. Kovatchev, and C. Cobelli, "The uva/padova type 1 diabetes simulator: new features," *Journal of Diabetes Science and Technology*, vol. 8, no. 1, pp. 26–34, 2014. 43, 45, 47, 48

[45] L. Magni, M. Forgione, C. Toffanin, C. D. Man, B. Kovatchev, G. D. Nicolao, and C. Cobelli, "Run-to-run tuning of model predictive control for type 1 diabetes subjects: in silico trial," *Journal of Diabetes Science and Technology*, vol. 3, no. 5, pp. 1091–1098, 2009. 43

[46] B. Kovatchev, C. Cobelli, E. Renard, S. Anderson, M. Breton, S. Patek, W. Clarke, D. Bruttomesso, A. Maran, S. Costa, A. Avogaro, C. D. Man, A. Facchinetti, L. Magni, G. D. Nicolao, J. Place, and A. Farret, "Multinational study of subcutaneous model-predictive closed loop control in type 1 diabetes mellitus: summary of the results," *Journal of Diabetes Science and Technology*, vol. 4, pp. 1374–1381, 2010. 43

[47] S. Shanthi and Kumar, “A novel approach for the prediction of glucose concentration in type 1 diabetes ahead in time through arima and differential evolution,” *Journal on Computer Science and Engineering*, vol. 38, pp. 4182–4186, 2011. 44

[48] B. Agar, M. Eren, and A. Cinar, “Glucosim: educational software for virtual experiments with patients with type 1 diabetes,” in *Annual International Conference in Medicine and Biology*, pp. 845–848, Institute of Electrical and Electronics Engineering, 2005. 44

[49] J. I. Hidalgo, E. Maqueda, J. L. Risco-Martín, A. Cuesta-Infante, J. M. Colmenar, and J. Nobel, “Glucmodel: a monitoring and modeling system for chronic diseases applied to diabetes,” *Journal of Biomedical Informatics*, vol. 48, pp. 183–192, 2014. 44

[50] J. M. Velasco, O. Garnica, J. Lanchares, M. Botella, and J. I. Hidalgo, “Combining data augmentation, edas and grammatical evolution for blood glucose forecasting,” *Memetic Computing*, Jun. 2018. 44

[51] M. Affenzeller and S. Wagner, “Offspring selection: a new self-adaptive selection scheme for genetic algorithms,” in *Adaptive and Natural Computing Algorithms*, pp. 218–221, Springer, 2005. 44

[52] M. Affenzeller, B. Burlacu, S. Winkler, M. Kommenda, G. Kronberger, and S. Wagner, “Offspring selection genetic algorithm revisited: improvements in efficiency by early stopping criteria in the evaluation of unsuccessful individuals,” in *Computer Aided Systems Theory*, pp. 424–431, Springer, 2018. 44

[53] C. Cervigón, J. I. Hidalgo, M. Botella, and R.-J. Villanueva, “A genetic algorithm approach to customizing a glucose model based on usual therapeutic parameters,” *Progress in Artificial Intelligence*, vol. 6, no. 3, pp. 255–261, 2017. 44

[54] T. Prud’Homme, A. Bock, G. François, and D. Gillet, “Preclinically assessed optimal control of postprandial glucose excursions for type 1 patients with diabetes,” in *Conference on Automation Science and Engineering*, pp. 702–707, Institute of Electrical and Electronics Engineering, 2011. 44

[55] I. De Falco, A. Della Cioppa, T. Koutny, M. Krcma, U. Scafuri, and E. Tarantino, “Genetic programming-based induction of a glucose-dynamics model for telemedicine,” *Journal of Network and Computer Applications*, vol. 119, pp. 1–13, 2018. 44

[56] G. Steil, K. Rebrin, F. Hariri, S. Jinagonda, S. Tadros, C. Darwin, and M. Saad, “Interstitial fluid glucose dynamics during insulin-induced hypoglycaemia,” *Diabetologia*, vol. 48, no. 9, pp. 1833–1840, 2005. 45

[57] W. Clarke, D. Cox, L. Gonder-Frederick, W. Carter, and S. Pohl, “Evaluating clinical accuracy of systems for self-monitoring of blood glucose.,” *Diabetes Care*, vol. 10, pp. 622–628, Sep. 1987. 45, 69, 93

[58] N. Lourenço, J. M. Colmenar, J. I. Hidalgo, and Ó. Garnica, “Structured grammatical evolution for glucose prediction in diabetic patients,” in *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 1250–1257, Association for Computing Machinery, 2019. 45

[59] I. Contreras, S. Oviedo, M. Vettoretti, R. Visentin, and J. Vehí, “Personalized blood glucose prediction: a hybrid approach using grammatical evolution and physiological models,” *PloS one*, vol. 12, no. 11, p. 0187754, 2017. 45

[60] M. H. Zangooei, J. Habibi, and R. Alizadehsani, “Disease diagnosis with a hybrid method svr using nsga-ii,” *Neurocomputing*, vol. 136, pp. 14–29, 2014. 46

[61] K. Deb, *Multi-objective optimization using evolutionary algorithms*, vol. 16. John Wiley and Sons, 2001. 46, 89

[62] D. Dua and C. Graff, “University of california irvine machine learning repository,” 2017. 46

[63] R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: current issues and guidelines," *Journal of Medical Informatics*, vol. 77, no. 2, pp. 81–97, 2008. 46

[64] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann, 2013. 46

[65] D. Vigneswari, N. K. Kumar, V. Ganesh Raj, A. Gugan, and S. R. Vikash, "Machine learning tree classifiers in predicting diabetes mellitus," in *International Conference on Advanced Computing and Communication Systems*, pp. 84–87, 2019. 46

[66] J. Quinlan, "Simplifying decision trees," *Journal of Man-Machine Studies*, vol. 27, no. 3, pp. 221–234, 1987. 46

[67] J. R. Quinlan, *C4.5: programs for machine learning*. Morgan Kaufmann, 1993. 46

[68] T. K. Ho, "Random decision forests," in *Proceedings of International Conference on Document Analysis and Recognition*, vol. 1, pp. 278–282, 1995. 46

[69] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," *Machine Learning*, vol. 59, pp. 161–205, Feb. 2005. 46

[70] L. Rokach and O. Maimon, *Data mining with decision trees*. World Scientific, 2nd ed., 2014. 46

[71] K. C. Howlader, M. S. Satu, A. Barua, and M. A. Moni, "Mining significant features of diabetes mellitus applying decision trees: a case study in bangladesh," *bioRxiv*, 2018. 46

[72] I.-W. Wu, Z.-Y. Chen, J.-J. Shann, and C.-P. Chung, "Instruction set extension exploration in multiple-issue architecture," in *Proceedings of the Conference on Design, Automation and Test in Europe*, pp. 764–769, Association for Computing Machinery, 2008. 46

[73] P. Pumpuang, A. Srivihok, and P. Praneetpolgrang, "Comparisons of classifier algorithms: bayesian network, c4.5, decision forest and nbtree for course registration planning model of undergraduate students," in *International Conference on Systems, Man and Cybernetics*, pp. 3647–3651, Institute of Electrical and Electronics Engineering, 11 2008. 46

[74] T. Patil and S. Sherekar, "Performance analysis of naive bayes and j48 classification algorithm for data classification," *Journal of Computer Science and Applications*, vol. 6, pp. 256–261, 01 2013. 46

[75] S. Oviedo, I. Contreras, C. Quirós, M. Giménez, I. Conget, and J. Vehi, "Risk-based postprandial hypoglycemia forecasting using supervised learning," *Journal of Medical Informatics*, vol. 126, pp. 1–8, 2019. 46

[76] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay, "Scikit-learn: machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011. 46

[77] Q. Sun, M. Jankovic, J. Budzinski, B. Moore, P. Diem, C. Stettler, and S. G. Mougiakakou, "A dual mode adaptive basal-bolus advisor based on reinforcement learning," *Journal of Biomedical and Health Informatics*, 2018. 46

[78] Q. Sun, M. V. Jankovic, L. Bally, and S. G. Mougiakakou, "Predicting blood glucose with an lstm and bi-lstm based deep neural network," in *Symposium on Neural Networks and Applications*, pp. 1–5, Institute of Electrical and Electronics Engineering, 2018. 47

[79] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, 12 1997. 47

[80] E. M. Aiello, G. Lisanti, L. Magni, M. Musci, and C. Toffanin, "Therapy-driven deep glucose forecasting," *Engineering Applications of Artificial Intelligence*, vol. 87, p. 103255, 2020. 47

[81] C. Palerm, H. Zisser, L. Jovanovič, and F. Doyle, "A run-to-run control strategy to adjust basal insulin infusion rates in type 1 diabetes," *Journal of Process Control*, vol. 18, pp. 258–265, Feb. 2008. 47

[82] J. Martinsson, A. Schliep, B. Eliasson, C. Meijner, S. Persson, and O. Mogren, "Automatic blood glucose prediction with confidence using recurrent neural networks," in *International Workshop on Knowledge Discovery in Healthcare Data*, pp. 64–68, Jul. 2018. 47

[83] C. Marling and R. C. Bunescu, "The ohiot1dm dataset for blood glucose level prediction," *CEUR Workshop Proceedings*, vol. 2675, pp. 71–74, 2018. 48

[84] K. Li, J. Daniels, C. Liu, P. Herrero, and P. Georgiou, "Convolutional recurrent neural networks for glucose prediction," *Journal of Biomedical and Health Informatics*, vol. 24, no. 2, pp. 603–613, 2019. 48

[85] C. Zhao, E. Dassau, L. Jovanovič, H. C. Zisser, F. J. Doyle III, and D. E. Seborg, "Predicting subcutaneous glucose concentration using a latent-variable-based statistical method for type 1 diabetes mellitus," *Journal of Diabetes Science and Technology*, vol. 6, no. 3, pp. 617–633, 2012. 48

[86] D. A. Finan, C. C. Palerm, F. J. Doyle, D. E. Seborg, H. Zisser, W. C. Bevier, and L. Jovanovic, "Effect of input excitation on the quality of empirical dynamic models for type 1 diabetes," *AIChE Journal*, vol. 55, pp. 1135–1146, Mar. 2009. 48

[87] M. Reddy, P. Pesl, M. Xenou, C. Toumazou, D. Johnston, P. Georgiou, P. Herrero, and N. Oliver, "Clinical safety and feasibility of the advanced bolus calculator for type 1 diabetes based on case-based reasoning: a 6-week nonrandomized single-arm pilot study," *Diabetes Technology and Therapeutics*, vol. 18, no. 8, pp. 487–493, 2016. 48

[88] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," *Artificial Intelligence in Medicine*, vol. 34, no. 2, pp. 113–127, 2005. 51

[89] H.-Y. Chen, C.-H. Chuang, Y.-J. Yang, and T.-P. Wu, "Exploring the risk factors of preterm birth using data mining," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5384–5387, 2011. 51

[90] C.-D. Chang, C.-C. Wang, and B. C. Jiang, "Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5507–5513, 2011. 51

[91] X.-H. Meng, Y.-X. Huang, D.-P. Rao, Q. Zhang, and Q. Liu, "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors," *The Kaohsiung Journal of Medical Sciences*, vol. 29, no. 2, pp. 93–99, 2013. 51

[92] J. L. Breault, C. R. Goodall, and P. J. Fos, "Data mining: a diabetic data warehouse," *Artificial Intelligence in Medicine*, vol. 26, no. 1, pp. 37–54, 2002. Medical Data Mining and Knowledge Discovery. 51

[93] A. Ramezankhani, O. Pournik, J. Shahrabi, D. Khalili, F. Azizi, and F. Hadaegh, "Applying decision tree for identification of a low risk population for type 2 diabetes. tehran lipid and glucose study," *Diabetes Research and Clinical Practice*, vol. 105, no. 3, pp. 391–398, 2014. 51

[94] G. V. Kass, "An exploratory technique for investigating large quantities of categorical data," *Journal of the Royal Statistical Society*, vol. 29, no. 2, pp. 119–127, 1980. 52, 66

[95] A. Field, *Discovering statistics using IBM SPSS statistics*. Sage, 2013. 53

[96] Z. Gu, L. Gu, R. Eils, M. Schlesner, and B. Brors, "Circlize implements and enhances circular visualization in r," *Bioinformatics*, vol. 30, pp. 2811–2812, 2014. 55

[97] R. C. Team, *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, 2016. 55, 66

[98] S. Fonda, D. Lewis, and R. Vigersky, “Minding the gaps in continuous glucose monitoring: a method to repair gaps to achieve more accurate glucometrics,” *Journal of Diabetes, Science and Technology*, vol. 7, pp. 88–92, 2 2013. 66

[99] V. H. Bhat, P. G. Rao, P. D. Shenoy, K. R. Venugopal, and L. M. Patnaik, “An efficient prediction model for diabetic database using soft computing techniques,” in *Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*, pp. 328–335, Springer, 2009. 66

[100] R. Dougherty, A. Edelman, and J. Hyman, “Nonnegativity, monotonicity, or convexity-preserving cubic and quintic hermite interpolation,” *Mathematics of Computation*, vol. 52, pp. 471–494, Apr. 1989. 66, 76

[101] M. Berger and D. Rodbard, “Computer simulation of plasma insulin and glucose dynamics after subcutaneous insulin injection,” *Diabetes Care*, vol. 12, no. 10, pp. 725–736, 1989. 66, 76

[102] J. M. Velasco, O. Garnica, S. Contador, J. Lanchares, E. Maqueda, M. Botella, and J. I. Hidalgo, “Data augmentation and evolutionary algorithms to improve the prediction of blood glucose levels in scarcity of training data,” in *Congress on Evolutionary Computation*, pp. 2193–2200, Institute of Electrical and Electronics Engineering, Jun. 2017. 67, 77

[103] S. Wagner and M. Affenzeller, *HeuristicLab: a generic and extensible optimization environment.* Springer, 2005. 67, 77

[104] J. R. Koza, *Genetic programming.* The Massachusetts Institute of Technology Press, 1992. 67, 70

[105] H. Akaike, *Information theory and an extension of the maximum likelihood principle*, pp. 199–213. Springer, 1998. 68, 79

[106] J. Parkes, S. Slatin, S. Pardo, and B. Ginsberg, “A new consensus error grid to evaluate the clinical significance of inaccuracies in the measurement of blood glucose.,” *Diabetes Care*, vol. 23, no. 8, pp. 1143–1148., 2000. 69, 79, 80

[107] S. Luke, “Two fast tree-creation algorithms for genetic programming,” *Transactions on Evolutionary Computation*, vol. 4, no. 3, pp. 274–283, 2000. 70

[108] M. Keijzer, *Improving symbolic regression with interval arithmetic and linear scaling.* Springer, 2003. 70

[109] S. Frontoni, P. Bartolo, A. Avogaro, E. Bosi, G. Paolisso, and A. Ceriello, “Glucose variability: an emerging target for the treatment of diabetes mellitus,” *Diabetes Research and Clinical Practice*, vol. 102, pp. 86–95, 09 2013. 76

[110] S. Contador, J. M. Velasco, O. Garnica, and J. I. Hidalgo, “Profiled glucose forecasting using genetic programming and clustering,” in *Proceedings of the Annual Symposium on Applied Computing*, pp. 529–536, Association for Computing Machinery, 2020. 79

[111] B. Calvo, J. Ceberio, and J. A. Lozano, “Bayesian inference for algorithm ranking analysis,” in *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 324–325, 2018. 87, 107

[112] S. Contador, J. M. Colmenar, O. Garnica, and J. I. Hidalgo, “Short and medium term blood glucose prediction using multi-objective grammatical evolution,” in *International Conference on the Applications of Evolutionary Computation*, pp. 494–509, Springer, 2020. 89, 109

[113] B. S. Larsen KH, “Generation of dose calculation data tables using cubic spline interpolation,” *Medical Dosimetry*, vol. 16, 1991. 91

[114] G. Wahba, “Spline models for observational data,” *Regional Conference Series in Applied Mathematics*, vol. 59, 1990. 91

[115] C. Marling and R. Bunescu, “The ohiot1dm dataset for blood glucose level prediction: update 2020,” *CEUR Workshop Proceedings*, 2020. 92, 96, 109

[116] S. R. Colberg, R. J. Sigal, J. E. Yardley, M. C. Riddell, D. W. Dunstan, P. C. Dempsey, E. S. Horton, K. Castorino, and D. F. Tate, "Physical activity/exercise and diabetes: a position statement of the american diabetes association," *Diabetes Care*, vol. 39, no. 11, pp. 2065–2079, 2016. 92

[117] G. Kenny, R. Sigal, and R. McGinn, "Body temperature regulation in diabetes," *Temperature*, vol. 3, pp. 119–145, 2016. 92

[118] S. Umapathy, T. Rajalakshmi, C. Sri, G. Balachander, and K. Shankar, "Non-invasive blood glucose analysis based on galvanic skin response for diabetic patients," *Biomedical Engineering: Applications, Basis and Communications*, vol. 30, p. 1850009, 2018. 92

[119] J. I. Hidalgo, M. Botella, J. M. Velasco, O. Garnica, C. Cervigón, R. Martínez, A. Aramendi, E. Maqueda, and J. Lanchares, "Glucose forecasting combining markov chain based enrichment of data, random grammatical evolution and bagging," *Applied Soft Computing*, vol. 88, p. 105923, 2020. 93

[120] C. Ryan, J. Collins, and M. Neill, "Grammatical evolution: evolving programs for an arbitrary language," in *Genetic Programming*, vol. 1391 of *Lecture Notes in Computer Science*, pp. 83–96, Springer, 1998. 94

[121] J. M. Velasco, O. Garnica, S. Contador, J. M. Colmenar, E. Maqueda, M. Botella, J. Lanchares, and J. I. Hidalgo, "Enhancing grammatical evolution through data augmentation: application to blood glucose forecasting," in *International Conference on the Applications of Evolutionary Computation*, pp. 142–157, Springer, 2017. 94

[122] E. Hemberg, L. Ho, M. O'Neill, and H. Claussen, "A comparison of grammatical genetic programming grammars for controlling femtocell network coverage," *Genetic Programming and Evolvable Machines*, vol. 14, pp. 65–93, Mar. 2013. 94

[123] D. Moreno-Salinas, E. Besada-Portas, J. López-Orozco, D. Chaos, J. de la Cruz, and J. Aranda, "Symbolic regression for marine vehicles identification," *International Federation of Automatic Control PapersOnLine*, vol. 48, no. 16, pp. 210–216, 2015. 94

[124] Y. Bengio and Y. Grandvalet, "No unbiased estimator of the variance of k-fold cross-validation," *Journal of Machine Learning Research*, vol. 5, pp. 1089—1105, 2004. 95

[125] I. Capel, M. Rigla, G. García-Sáez, A. Rodríguez-Herrero, B. Pons, D. Subías, F. García-García, M. Gallach, M. Aguilar, C. Pérez-Gandía, E. Gómez Aguilera, A. Caixàs, and M. E. Hernando, "Artificial pancreas using a personalized rule-based controller achieves overnight normoglycemia in patients with type 1 diabetes," *Diabetes Technology and Therapeutics*, vol. 16, 2013. 95

[126] E. Donga, M. Dijk, J. van Dijk, N. Biermasz, G.-J. Lammers, K. Kralingen, R. Hoogma, E. Corssmit, and J. Romijn, "Partial sleep restriction decreases insulin sensitivity in type 1 diabetes," *Diabetes Care*, vol. 33, pp. 1573–1577, 2010. 95

[127] C. Marling, J. Shubrook, S. Vernier, M. Wiley, and F. Schwartz, "Characterizing blood glucose variability using new metrics with continuous glucose monitoring data," *Journal of Diabetes Science and Technology*, vol. 5, pp. 871–878, 2011. 95

[128] M. Wilinska, L. Chassin, H. Schaller, L. Schaupp, T. Pieber, and R. Hovorka, "Insulin kinetics in type-1 diabetes: continuous and bolus delivery of rapid acting insulin," *Transactions on Biomedical Engineering*, vol. 52, pp. 3–12, 2005. 95

[129] J. I. Hidalgo, J. M. Colmenar, J. M. Velasco, G. Kronberger, S. M. Winkler, O. Garnica, and J. Lanchares, "Identification of models for glucose blood values in diabetics by grammatical evolution," in *Handbook of Grammatical Evolution*, pp. 367–393, Springer, 2018. 97

[130] S. Contador, J. I. Hidalgo, O. Garnica, J. M. Velasco, and J. Lanchares, "Can clustering improve glucose forecasting with genetic programming models?," in *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 1829–1836, Association for Computing Machinery, 2019. 102