mathematics

MDPI

Article

Predicting Glycemic Control in a Small Cohort of Children with Type 1 Diabetes Using Machine Learning Algorithms

Bogdan Neamtu [1,2,3], Mihai Octavian Negrea [2,4,*] and Iuliana Neagu [2]

[1] Clinical Research Department, Pediatric Clinical Hospital Sibiu, 550166 Sibiu, Romania; bogdan.neamtu@ulbsibiu.ro or bmneam01@louisville.edu
[2] Faculty of Medicine, Lucian Blaga University, 550024 Sibiu, Romania; tarceaiulianamihaela@gmail.com
[3] Bioinformatics and Biostatistics Department, University of Louisville, Louisville, KY 40202, USA
[4] County Clinical Emergency Hospital of Sibiu, 2-4 Corneliu Coposu Str., 550245 Sibiu, Romania
* Correspondence: dr.mihai.negrea@gmail.com or mihaioctavian.negrea@ulbsibiu.ro

Abstract: Type 1 diabetes, a chronic condition characterized by insulin deficiency, is associated with various complications and reduced life expectancy and is increasing in global prevalence. Maintaining glycaemic control in children with type 1 diabetes, as reflected by glycated hemoglobin levels (A1C), is a challenging task. The American Association of Diabetes (ADA), the Pediatric Endocrine Society, and the International Diabetes Federation (ISPAD) recommend the adoption of a harmonized A1C of <7.5% across all pediatric groups. Our retrospective study included 79 children with type 1 diabetes and aimed to identify determinants pivotal to forecasting glycemic control, focusing on a single A1C cut-off value and exploring how machine learning algorithms can enhance clinical understanding, particularly with smaller sample sizes. Bivariate analysis identified correlations between glycemic control and disease duration, body mass index (BMI) Z-score at onset, A1C at onset above 7.5 g/dL, family income, living environment, maternal education level, episodes of ketoacidosis, and elevated cholesterol or triglyceride. Binary logistic regression stressed the association of ketoacidosis episodes ($\beta = 21.1$, $p < 0.01$) and elevated A1C levels at onset ($\beta = 3.12$, $p < 0.01$) and yielded an area under the receiver operating characteristic curve (AUROC) of 0.916. Two-step clustering emphasized socioeconomic factors, as well as disease complications and comorbidities, and delineated clusters based on these traits. The classification and regression tree (CART) yielded an AUROC of 0.954, slightly outperforming binary regression, providing a comprehensive view of interactions between disease characteristics, comorbidities, and socioeconomic status. Common to all methods were predictors regarding ketoacidosis episodes, the onset of A1C levels, and family income, signifying their overarching importance in glycaemic control. While logistic regression quantified risk, CART visually elucidated complex interactions and two-step clustering exposed patient subgroups that might require different intervention strategies, highlighting how the complementary nature of these analytical methods can enrich clinical interpretation.

Keywords: pediatric type 1 diabetes; glycemic control; binary regression; machine learning; two-step cluster analysis; CART decision trees

MSC: 92C50; 62H30

Citation: Neamtu, B.; Negrea, M.O.; Neagu, I. Predicting Glycemic Control in a Small Cohort of Children with Type 1 Diabetes Using Machine Learning Algorithms. *Mathematics* **2023**, *11*, 4388. https://doi.org/10.3390/math11204388

Academic Editor: Oluwagbohunmi A. Awosoga

Received: 28 September 2023
Revised: 18 October 2023
Accepted: 20 October 2023
Published: 22 October 2023

1. Introduction

Type 1 diabetes, a chronic condition characterized by insulin deficiency due to autoimmune β-cell depletion, is associated with various complications and reduced life expectancy [1–4]. The global prevalence was 8.4% in 2017 and is projected to reach 9.9% by 2045, with an estimated 425 million cases worldwide, including 58 million in Europe [5]. Multiple factors, such as demographics, biology, and socioeconomic status, influence glycemic control, which is measured using glycated hemoglobin (A1C) in pediatric patients [6–17]. A detailed comprehension of such factors can provide much-needed aid in

understanding and limiting the complications that arise during the progression of type 1 diabetes and in an attempt to improve the quality of life and lifespan of these patients [6–14]. Maintaining glycemic control in type 1 diabetes children is a challenging task in clinical practice. These levels are tailored to age group categories. The current consensus advocates maintaining an A1C goal of under 8.5% for youths under the age of 6 years, 8% for those 6–12 years old, and 7.5% for those aged 13–19 years. Nevertheless, the American Diabetes Association (ADA) proposed the harmonization of its glycemic goals with those of the Pediatric Endocrine Society and the International Diabetes Federation (ISPAD) using a single A1C of <7.5% across all pediatric age groups [15,16].

Our hypothesis centers on leveraging the potential of both supervised(CART and logistic regression) and unsupervised machine learning methods(two-step clustering analysis) to demonstrate the feasibility of a single A1C cut-off <7.5% for the outcome variable (good glycemic control) across all pediatric age groups, even in a small cohort. This approach allows for a thorough analysis and emphasis on the impact of covariates, such as disease duration, body mass index (BMI) Z-score at onset, an A1C at onset above 7.5 g/dL, family income, living environment, maternal education level, episodes of ketoacidosis, and elevated cholesterol or triglycerides.

To achieve this goal, our retrospective study sought to identify the risk factors impacting glycemic control through bivariate analysis first and then explored their influence as predictors using advanced statistical methods, including binary regression, two-step clustering, and the classification and regression tree (CART) algorithm focusing on a single cut-off value consistent across age groups.

Bivariate analysis offers the advantage of being easy to implement and has a relatively intuitive interpretability. It is frequently employed in cross-sectional retrospective studies examining a particular dichotomous outcome. It poses, however, certain key disadvantages due to data collection bias when it comes to small sample sizes or when exposed to multicollinearity and confounding factors. While it is essential to identify potential risk factors for the targeted outcome variable, bivariate analysis fails to explore complex interrelations between the studied variables, which can affect the outcome. Moreover, in addition to these aforementioned limitations, retrospective studies are confronted with different levels of missing data for the independent variables in the dataset. This limitation restrains sample sizes, especially when dealing with linear models.

To address these issues, we adopted advanced algorithms for a better interpretation of our results. We selected binary logistic regression for a deeper insight into the hierarchy of the observed risk factors. This enabled a direct comparison of coefficients, highlighting each predictor's significance. Bivariate analysis assesses the effect sizes for similar variables, but binary regression concurrently manages both categorical and continuous ones, allowing their collective expression of effect. This yields a linear prediction model grounded in each risk factor's weight. Conversely, the CART decision tree, a nonlinear supervised learning method, segments predictor space using optimal cut-off points. This captures intricate data relationships, producing distinct subgroups based on cumulative decisions and the proportion of the outcome—glycemic control in our study.

Clinical interpretation often relies on grouping populations by risk profiles for tailored interventions. The intricate relationship between risk factors is not merely linear; it often necessitates sophisticated methods like CART decision trees to discern how these factors shape groups and outcomes. Two-step cluster analysis, an unsupervised learning method, achieves similar goals using a distinct mathematical approach. It differentiates groups by optimizing Euclidean distances between variable entries and their cluster centroids. Thus, our hypothesis is grounded on the potential complementarity of methods in creating specific risk profile groups, leading to more nuanced interpretations given their flexibility to handle both categorical and continuous variables. Eventually, we intend to compare the effectiveness of CART decision trees, cluster analysis, and logistic regression to enhance the clinical interpretation of results and reduce the impact of confounding variables among a limited group of children with type 1 diabetes.

2. Materials and Methods

2.1. Setting and Time Frame

We performed a retrospective analysis of relevant data extracted from the discharge letters or outpatient files of patients admitted to the Pediatric Clinical Hospital Sibiu between January 2010 and August 2023. All unique health records related to the diagnosis of type 1 diabetes (E.10.11–E.10.90 according to the International Classification of Disease-ICD-10) and meeting the inclusion criteria were retrieved from the Pediatric Clinical Hospital Sibiu's database. Of the 506 distinct cases identified, only 79 had complete data records for each variable. A filtering approach was employed based on the following sequence of commands in SPSS v.21: [select cases → if condition(list of variables) → not missing].

2.2. Inclusion and Exclusion Criteria

Patients aged between 0 and 18 years old diagnosed with type 1 diabetes were enrolled in the study. The diagnosis of type 1 diabetes was established according to the latest recommendations of the American Diabetes Association as follows [4]: FPG $\geq$ 126 mg/dL or 2 h PG $\geq$ 200 mg/dL during OGTT or A1C $\geq$ 6.5% or a random plasma glucose $\geq$ 200 mg/dL in a patient with classic symptoms of hyperglycemia or hyperglycemic crisis (FPG—fasting plasma glucose where fasting is defined as no caloric intake for at least 8 h; 2h-PG—2 h plasma glucose; OGTT—oral glucose tolerance test as described by the World Health Organisation (WHO) using a glucose load containing the equivalent of 75 g anhydrous glucose dissolved in water [18]). Patients were excluded from analysis if data were missing in any of the studied variables or if they were documented as at risk for blood disorders affecting the HbA1c measurement, pancreatitis with a personal or familial history of MEN types 2A or B, thyroid issues, medullary thyroid cancer or taking steroids in the past 2 months prior to admittance. Out of the 427 patients excluded based on incomplete health records, none of them met the medical exclusion criteria.

2.3. Data Collection and Structure

We recorded characteristics describing the patients' demographic, socioeconomic, anthropometric, laboratory, and disease-related factors in a similar approach to Niba et al. [14] and Urbach et al. [19]. Then, we checked for duplicate detection and made use of descriptive statistics(mean, median, mode, standard deviation, minimum, maximum) along with histograms and box plots to identify outliers or values that did not make sense. For categorical variables, cross-tabulations were used to understand how different categories intersected.

Table 1 provides an overview of the data collected, how they were processed, and relevant references in the literature with similar, validated approaches where appropriate. The level of education was measured according to ISCED levels (International Standard Classification of Education [20]). The insulin regimen in all patients fell into one of two categories: regimen 1, with 3 rapid-acting insulin administrations and one long-acting insulin administration, and "Regimen 2", with 3 short-acting insulin administrations and one slow insulin administration. In addition, the mean A1C was calculated as the average, taking into account all the available documented values for this parameter measured during follow-up within the above-specified period of a patient. This approach is similar to Beck et al. [20]. A dichotomous approach was then implemented using the cut-off value of mean AIC = 7.5 g/dL. Values larger than or equal to the cut-off were considered as poor overall glycemic control according to the aforementioned guidelines [15,16].

Consequently, the outcome variable in our study was "Glycemic control", which divided our patients into two groups—one with adequate overall glycemic control and one without—based on an A1C cutoff of 7.5 g/dL. For bivariate analysis, primary hypotheses were then generated based on the tested variable type. For categorical variables, the null hypotheses were phrased as follows: "there is no significant association between glycemic control and [categorical variable]". For continuous variables, namely the body

mass index (BMI) Z-score and disease duration, the null hypotheses were phrased as "The means of [continuous variable] do not differ significantly between patients with or without adequate overall glycemic control". Variables that were found to significantly correlate with glycemic control were essentially highlighted as risk factors in this regard and subsequently considered for inclusion in further advanced statistical methods or machine-learning algorithms.

Table 1. Collected data and references with similar approaches to data curation for each variable.

Variable	Raw Data	Processed Variable Type (Explanation)	Ref.
Demographic characteristics			
Gender	Dichotomous	Dichotomous (Male/Female)	[14,19]
Paternal level of education	Ordinal (ISCED 0–7)	Dichotomous (ISCED > 4 vs. ISCED $\leq$ 4) Higher education vs. no higher education	[13,14,21–27]
Maternal level of education			
Family income	Continuous	Categorial (High/average/low compared to the average income in Romania for 2018)	
Living conditions	Dichotomous	Dichotomous (Rural/Urban)	
Disease characteristics			
Family history of diabetes	Dichotomous	Dichotomous (Yes/No)	[14,21,28]
Type of onset	Nominal	Categorial (Asymptomatic or symptomatic with no ketoacidosis/ Mild or moderate ketoacidosis/ severe ketoacidosis)	[28,29]
A1C at onset	Continuous	Dihotomous (<7.5 g/dL/$\geq$7.5 g/dL)	[22,30–33]
Age of onset	Continuous	Categorial (<5 years/5–10 years/>10 years)	[13,22,28]
BMI at onset (Z-score, according to WHO 2007 growth reference data [34])	Continuous	Continuous, Categorical (underweight/normal weight/overweight or obese) according to WHO Reference values for BMI Z-score (<2; 2–1; >1)	[13,14,35]
Disease duration	Continuous	Continuous	[24,36,37]
Identified associated autoimmune diseases	Nominal	Dichotomous (Yes/No)	[38]
Episodes of hypoglicemia	Discrete	Dichotomous (Yes/No)	[21,39]
Episodes of ketoacidosis (excluding onset [39])	Discrete	Dichotomous (Yes/No)	[39,40]
Episodes of viral infections	Discrete	Dichotomous (Yes/No)	[41]
Episodes of bacterial infections	Discrete	Dichotomous (Yes/No)	
Episodes of microalbuminuria	Discrete	Dichotomous (Yes/No)	[21,32,42]
Serum Cholesterol (last documented value)	Continuous	Dihotomous ($\geq$200 mg/dL/<200 mg/dL)	[13,35]
Serum Triglycerides (last documented value)	Continuous	Dihotomous ($\geq$150 mg/dL/<150 mg/dL)	
Insulin regimen	Nominal	Dichotomous (Regimen 1/Regimen 2—explained in text)	[13,14,43]
Neutrophil to lymphocyte ratio	Continuous	Continuous	[44–46]
Mean platelet volume	Continuous	Continuous	[47–49]

Specifically, we adjusted disease duration and the BMI Z-score by centering them around the sample's mean for use in binary logistic regression. This process involved subtracting the sample mean from each data point within these variables, a step undertaken to diminish the numerical instability that can arise in regression models due to large variable values.

For categorical variables like family income, we adopted a dummy variable approach. Here, income levels were categorized as low, average, or high, with each category represented as a separate, dichotomous variable, assuming values of 0 or 1. This approach simplifies the interpretation of regression outcomes, particularly for variables like "high family income." To further enhance interpretability, we reverse-coded this variable as "low or average income", converting the 0s to 1s and vice versa. This strategy was intended to yield positive regression coefficients in binary regression, streamlining the data interpretation process by aligning coding with expected outcome directions.

2.4. Data Analysis

2.4.1. Bivariate Analysis

Categorical variable analyses and results were presented as frequencies and percentages, while continuous variables were presented as means, standard deviations, minimum and maximum values, the interquartile range, and 95% confidence intervals for means. Continuous variables were tested for normality using the Kolmogorov–Smirnov or Shapiro–Wilk tests where appropriate. Groups were compared using the Chi-square or Fischer's exact test for categorical variables and the independent *t*-test or Mann–Whitney U-test for continuous variables. An α-level of 0.05 was considered statistically significant for bivariate analysis and potential inclusion in logistic regression models as well as the two-step clustering and CART algorithms.

2.4.2. Logistic Regression

Variables found to significantly correlate with glycemic control in bivariate analysis were included in the binary logistic regression models. Binary logistic regression is used to predict the outcome of a dichotomous dependent variable based on the input predictor variables. Inputs can be mapped using the sigmoid function, which turns them into a value between 0 and 1, enabling them to be treated as probabilities [50]:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{1}$$

where

- $\sigma(z)$ is the output value (constrained between 0 and 1).
- z is the input to the function, calculated as the weighted sum of inputs.

The sigmoid function is the inverse of the logit function, which transforms a probability value back to a real-value number, which can then be used as the target variable in linear regression models [50,51].

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) \tag{2}$$

To predict the outcome targeted by the model, probability estimation is employed [50]:

$$p = \sigma(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k) \tag{3}$$

where

- p is the probability that the dependent event occurred based on the linear combination of independent variables.
- β_i is the regression coefficient for variable x_i
- β_0 is the intercept value.

Consequently,

$$\text{logit}(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{4}$$

In order to measure the performance of the classification model where the prediction input is a probability value between 0 and 1, the Log Loss function [50] can be used.

$$\text{Log Loss} = -(y\log(p) + (1-y)\log(1-p)) \tag{5}$$

where

- y is the true class label (0 or 1 in binary regression);
- p is the predicted probability of the instance belonging to class 1.

Smaller values of Log Loss indicate the better performance of the classification model.

In order to estimate the distribution of regression coefficients, resampling methods such as bootstrapping can be implemented, which repeatedly samples replacements from the data. The bias-corrected and accelerated (BCa) method is an advanced bootstrapping technique that adjusts for both bias and skewness in bootstrap distribution.

Bias correction [52] comes from the proportion of bootstrap resamples that are less than the original statistic if

- θ^* denotes the statistic from a bootstrap sample;
- θ is the original statistic, and the bias correction is given as follows:

$$z_0 = \Phi^{-1}\left(\frac{1}{B}\sum_{i=1}^{B} I(\theta_i^* < \theta)\right) \tag{6}$$

where

- Φ^{-1} is the quantile function of the standard normal distribution,
- B is the number of bootstrap resamples,
- I is the indicator function.

Bias correction accounts for any systematic overestimation or underestimation using the bootstrap samples compared to the original sample estimate.

Acceleration accounts for the skewness of the bootstrap distribution and is calculated from jackknife estimates. The jackknife method involves systematically leaving out one observation at a time from the sample set and calculating the estimate over $n-1$ observations [53].

$$\alpha = \frac{\sum_{i=1}^{n}\left(\overline{\theta} - \hat{\theta}^{(-i)}\right)^3}{6\left[\sum_{i=1}^{n}\left(\overline{\theta} - \hat{\theta}^{(-i)}\right)^2\right]^{3/2}} \tag{7}$$

where

- $\hat{\theta}^{(-i)}$ is the statistic with the ith observation removed.
- $\overline{\theta}$ is the average of jackknife estimates.

The acceleration adjustment corrects for the skewness in the bootstrap distribution, ensuring that the confidence intervals are symmetric around the sample estimate. This method provides more accurate confidence intervals, especially in situations where data may not be perfectly symmetrical or when the sample size is small.

With the bias correction z_0 and acceleration α calculated, confidence intervals can be computed as follows:

$$CI = \left(\theta^*_{\alpha_1}, \theta^*_{\alpha_2}\right) \tag{8}$$

where

$$\alpha_1 = \Phi\left(z_0 + \frac{z_0 + z_{\alpha/2}}{1 - \alpha(z_0 + z_{\alpha/2})}\right) \tag{9}$$

$$\alpha_2 = \Phi\left(z_0 + \frac{z_0 + z_{1-\alpha/2}}{1 - \alpha(z_0 + z_{1-\alpha/2})}\right) \quad (10)$$

- $z_{1-\alpha/2}$ is the $\alpha/2$ quantile of the standard normal distribution.

In our study, the successive exhaustive addition or removal of variables was employed in order to find the optimal regression model. Continuous variables were centered around their mean to avoid multicollinearity, and categorical variables were recoded into dummy variables. Bootstrapping with 10,000 samples and calculated 95% confidence intervals for the regression coefficients were performed using the bias-corrected and accelerated method. Variables with a significant contribution to predict glycemic control ($p < 0.05$ and both 95% CI limits for coefficients above 0) were kept in the model.

2.4.3. Two-Step Cluster Algorithm

Variables found to significantly correlate with glycemic control were further fed to a two-step cluster algorithm. The functioning principle behind two-step clustering algorithms is based on combining a pre-clustering step using the k-means algorithm and hierarchical agglomerative clustering in sequence in order to classify cases based on similar characteristics, both regarding categorical and continuous variables. In the first step, the algorithm scans through the dataset to create many small sub-clusters. This is performed by measuring the distance (or similarity) between data points [54]. The centroid of a cluster resulting after the initial pre-cluster step serves as the 'average' for all points in a cluster, thereby summarizing the features of this cluster. The centroid of a cluster [55] is calculated by the following formula:

$$\text{Centroid} = \frac{1}{|C|}\sum_{x \in C} x \quad (11)$$

where

- $|C|$ is the number of data points in the cluster.
- x corresponds to individual data points in the cluster.

The centroid function's essence is to calculate the average of all data points within a cluster. By doing so, the algorithm can efficiently group similar data points together and ensure that the clusters are compact and distinct from each other. K-means clustering aims to find cluster centroids, defining clusters that are as cohesive as possible. To measure this aspect, the following objective function (or the Within-Cluster Sum of Squares (WCSS) using distance as the norm) [56] is implemented.

$$\text{Objective function} = \sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij} \| x_i - c_j \|^2 \quad (12)$$

where

- n is the total number of data points.
- k is the number of clusters.
- w_{ij} is a weight (defined as one if the datapoint I belongs to cluster j and 0 otherwise);
- x_i is the ith data point;
- c_j is the centroid of the jth cluster;
- $\| x_i - c_j \|^2$ is the squared Euclidean distance, which measures how far a point is from the centroid of its cluster.

K-means clustering aims to minimize WCSS, meaning that data points are as close as possible to the centroids of their respective clusters, leading to tight, distinct clusters.

Hierarchical clustering then merges the closest individual points in order to produce increasingly large clusters. The final number of clusters and the most accurate model can

be selected by calculating the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) [50]:

$$\text{AIC} = -2 \times \text{Log-likelihood} + 2 \times \text{k} \tag{13}$$

$$\text{BIC} = -2 \times \text{Log-likelihood} + \log(\text{N}) \times \text{k} \tag{14}$$

where

- k is the number of parameters in the model;
- N is the number of data points in the set.

BIC penalizes models with more parameters. The Log-likelihood distance measure gauges how well the model explains observed data, and a higher value indicates a better fit:

$$\text{Log-likelihood} = \sum \log p(x_i|\theta) \tag{15}$$

where

- θ represents the parameters of the model;
- $p(x_i|\theta)$ is the probability of the data point x_i being observed given the parameters θ.

In the context of the two-step clustering method, we can refine this to the following:

$$\text{Log} - \text{likelihood} = -N_s \left(\sum_{k=1}^{K^A} \frac{1}{2} \log\left(\sigma_k^2 + \sigma_{sk}^2\right) + \sum_{k=1}^{K^B} E_{sk} \right) \tag{16}$$

where

- N_s is the total number of data records in cluster s.
- K^A is the total number of continuous variables.
- K^B is the total number of categorical variables.
- σ_k^2 is the estimated variance of the continuous variable k for the entire dataset.
- σ_{sk}^2 is the estimated variance of the continuous variable k in cluster s.
- E_{sk} represents the entropy-based metric for categorical variables k in cluster s.

$$E_{sk} = -\sum_{l=1}^{L_k} \frac{N_{skl}}{N_s} \log \frac{N_{skl}}{N_s} \tag{17}$$

where

- L_k is the number of categories for the k-th categorical variable [54].
- N_{skl} is the number of observations in cluster s where the categorical variable k takes the value of l.
- N_s is the total number of observations where the categorical variable k takes the value of l.

Consequently, smaller values for AIC and BIC indicate a better model. Comparing these parameters across various cluster models, which differ in their number of clusters, can yield the most accurate representation.

Finally, to measure how similar an object is to its own cluster compared to other clusters, the Silhouette Score can be used [57].

$$\text{Silhouette score} = \frac{b - a}{\max(a, b)} \tag{18}$$

where

- a is the average distance from the sample to the other points in the same cluster.
- b is the smallest average distance from the sample to the points in other clusters and minimized over clusters.

A high Silhouette Score indicates that the object is adequately matched to its own cluster and poorly matched to neighboring clusters, showcasing good clustering.

In our study, Akaike's Information Criterion was implemented for optimal cluster number selection. The successive exhaustive addition or removal of variables was employed until a good average silhouette of cohesion separation (>0.5) was obtained. Variables with a predictor importance of at least 0.5 ($+/-0.01$) were kept in the model. Bivariate analysis was then applied in order to explore correlations between the resulting cluster variable and glycemic control.

2.4.4. CART Algorithm

The CART algorithm was further applied to explore a prediction model. Decision trees are supervised machine-learning algorithms that classify data, revealing patterns tied to user-defined outcomes and providing a visual representation of the model. The construction of this model initiates from the principal root and progresses through branching until no further divisions are feasible, correlating all predictor variables with the anticipated outcome. These bifurcations arise from conditions (internal nodes) set on predictor variables, culminating in further subdivisions and resolutions. Child nodes or "leaves" situated at the end of a branch represent the final resolutions given by the algorithm [58].

The metrics governing tree expansion encompass several measures to quantify and guide the bifurcations. One of the core measures used is Entropy: a metric that gauges disorder or impurity in a dataset. Mathematically [59], it is defined as:

$$\text{Entropy}(t) = -\sum_{i=1}^{k} p_i \log_2(p_i) \tag{19}$$

where p_i represents the proportion (or probability) of samples that belong to class i at node t. It is calculated by dividing the number of samples of class i by the total number of samples at that node. The whole formula captures the unpredictability or randomness of the data. When Entropy is low (close to 0), it means the data are very predictable (or pure). Conversely, data are very unpredictable (or impure) when Entropy is high. In the context of decision trees, splits that result in child nodes with low Entropy are generally preferred, as they lead to clearer distinctions between classes.

After a potential split, considering the different sizes and heterogeneities of the resulting child nodes, weighted Entropy can be calculated. It gives the average Entropy across all child nodes (or subsets) that result from the split, weighted by the size of each child subset:

$$\text{Weighted Entropy}(\text{children}) = \sum_{c \in \text{Children}} \frac{|c|}{|t|} \text{Entropy}(c) \tag{20}$$

where

- $|c|$ represents the number of samples in child node c;
- $|t|$ is the total number of samples in the current node (parent node) before the split.

As such, the division between these two parameters yields a weight for each child node based on its size relative to the parent node. It ensures that larger child nodes have a more significant impact on the overall weighted Entropy than smaller child nodes. By weighing the entropies based on the relative sizes of the child nodes, one can obtain an appropriately balanced view of the overall disorder or impurity resulting from the split.

A crucial decision criterion, called "Information Gain", is then computed as the difference between Entropy and weighted Entropy to measure how well a feature separates the dataset as follows:

$$\text{Information Gain} = -\sum_{i=1}^{k} p_i \log_2(p_i) - \sum_{c \in \text{Children}} \frac{|c|}{|t|} \text{Entropy}(c) \tag{21}$$

However, to ensure that a potential split does not merely add complexity without substantive Information Gain, "Split Information" is then considered:

$$\text{Split Information} = -\sum_{c \in \text{Children}} \frac{|c|}{|t|} \log_2\left(\frac{|c|}{|t|}\right) \tag{22}$$

This parameter gauges the potential for a feature to overfit based on the number and size of the child nodes it creates. To balance the trade-off between Information Gain and Split Information, the "Gain Ratio" is computed:

$$\text{Gain Ratio} = \frac{\text{Information Gain}}{\text{Split Information}} = \frac{-\sum_{i=1}^{k} p_i \log_2(p_i) - \sum_{c \in \text{Children}} \frac{|c|}{|t|} \text{Entropy}(c)}{-\sum_{c \in \text{Children}} \frac{|c|}{|t|} \log_2\left(\frac{|c|}{|t|}\right)} \tag{23}$$

The Gain Ratio normalizes Information Gain using Split Information, thereby reducing the potential for overfitting.

The Gini Index is a further measure of impurity, like Entropy. A smaller Gini Index value indicates a purer node.

$$\text{Gini Index} = 1 - \sum_{i=1}^{k} p_i^2 \tag{24}$$

where p_i is the proportion of samples that belong to class i for a particular node t. The Gini Index is 0 when all data in a node belong to a single class (pure node), and it is the maximum when data are evenly distributed across multiple classes. The Gini Index is vital not only in the initial splitting decisions of the tree but also in the pruning process. When pruning the tree, branches that result in nodes with high Gini Index values (indicating impurity or reduced predictive power) are primary candidates for removal. By focusing on nodes that do not significantly decrease the Gini Index, or in other words, do not add substantial clarity or predictive power, the pruning process ensures a more optimal and interpretable tree structure.

CART refers to binary decision trees that leverage the abovementioned metrics to partition data based on predictor variables and node homogeneity. The optimal division is chosen from all possible bifurcation trajectories while progressing from the parent to the child node. This methodology iterates until halting conditions are met, and no further decrease in node heterogeneity is possible. After reaching its maximal depth, the decision tree undergoes pruning, which removes nodes with minimal informational values. This pruning process is mathematically represented as the Complexity Cost:

$$\text{Cost}(T) = \sum_{\text{leaves} \in T} w_{\text{leaf}} \times \text{error}_{\text{leaf}} + \alpha \times (\text{number of terminal nodes}) \tag{25}$$

where

- w_{leaf} is the proportion of samples reaching the leaf;
- $\text{error}_{\text{leaf}}$ is the error at that leaf;
- α controls the trade-off between the complexity and fit of the tree.

Pruning aims to minimize this cost to avoid overfitting by removing sections of the tree that provide little predictive power, effectively reducing its size.

Pruned CART models deploy cross-validation through cost-complexity techniques for refinement, aiming to reduce average mean square prediction errors and enhance model stability. The CART decision tree's adaptability, resilience to outliers, and proficient management of absent data underscore its versatility in clinical data analysis.

We employed CART in the pruning mode to avoid overfitting, which internally uses cross-validation to select the best-pruned tree. This algorithm was fed with the variables that correlated with glycemic control, while those that did not contribute to the prediction

were removed to attain the optimal model. We allowed for the automatic selection of maximum growing levels (i.e., 5), with 5 as the minimum number of cases for parent nodes and 3 for child nodes. Regarding the Gini impurity measure, a minimum change in the improvement of 0.0001 was set (the maximum difference of risk in standard errors was 0).

2.4.5. Method Comparison and Agreement

The area under the receiver operating characteristic curve (AUROC) was used to evaluate the performance of our predictive models. Specifically, it revealed the discriminatory abilities of the CART and binary regression techniques for predicting inadequate glycemic control.

The receiver operating characteristic (ROC) curve is a graphical plot illustrating the diagnostic capacity of a binary classifier system as its discrimination threshold varies. The ROC curve is created by plotting the true positive rate (sensitivity) against the false positive rate (1 specificity) at various threshold settings. The true positive rate denotes the proportion of actual positives correctly identified as such, while the false positive rate represents the proportion of negatives incorrectly classified as positives.

The area under the ROC curve (AUROC) is a crucial metric derived from the ROC curve that quantifies the overall ability of the model to discriminate between the two categories under consideration. The AUROC value ranges from 0 to 1, where a value of 0.5 suggests no discrimination (akin to random guessing), and a value of 1 indicates perfect discrimination. Essentially, a higher AUROC value signifies that the model has a higher likelihood of classifying outcomes correctly. In the context of binary logistics regression, the AUROC interprets the probability estimates that each case belongs to a particular category (e.g., good versus poor glycemic control) based on predictor variables. A higher area under the curve indicates that this model has a good measure of separability, successfully distinguishing between patients with different outcomes based on the input variables. When it comes to CART algorithms, which are typically more complex due to their hierarchical structure, the AUROC is calculated using predicted probabilities derived from the classification tree. These probabilities, representing the likelihood of each case, belong to a specific category and are assessed at each terminal node of the tree. As with logistics regression, a higher AUROC value indicates a more robust model that is adept at segregating outcomes based on the decision rules generated during the analysis.

We operated the Kappa coefficient to measure consistency between risk classes derived from two-step clustering and CART decision tree analysis. The Kappa coefficient measures the agreement between two categorical methods, correcting for agreement via chance. It is calculated using a contingency table, where the observed agreement (P_o) and the probability of expected agreement (P_e) are assessed. The Kappa value is derived using the following formula:

$$\text{Kappa} = \frac{P_o - P_e}{1 - P_e} \tag{26}$$

The kappa coefficient ranges from -1 (total disagreement) to +1 (perfect agreement), and a score of 0 indicates an agreement equivalent to chance. Guidelines for interpreting the strength of the kappa agreement vary, but one common interpretation by Landis and Koch (1977) [60] is as follows: <0—Poor agreement, 0 to 0.20—Slight agreement, 0.21 to 0.40—Fair agreement, 0.41 to 0.60—Moderate agreement, 0.61 to 0.80—Substantial agreement, 0.81 to 1.00—Almost perfect agreement. In our research, we categorized the results of CART and two-step cluster analysis into three risk classes to enable method agreement measurements.

2.4.6. Handling Outliers

The outlier's problem for continuous variables (4 cases for disease duration) was handled by running the prediction models, including and excluding the respective records to check the stability of our results.

3. Results

3.1. Bivariate Analysis

3.1.1. Demographic and Socioeconomic Characteristics

A total of 79 patients were included in the study. In total, 46 patients (58.2%) had a mean A1C $\geq$ 7.5 g/dL, corresponding to poor glycemic control.

Table 2 shows the results for each demographic and socioeconomic variable studied and the associated *p*-value for appropriate statistical tests and correlation with glycemic control.

Table 2. Demographic and socioeconomic characteristics.

Variable	Categories	Frequency of Good Glycemic Control (% of Category)	Frequency of Poor Glycemic Control (% of Category)	Total (% of Grand Total)	*p*-Value
Gender	Male Female	22 (48.9%) 11 (32.4%)	23 (51.1%) 23 (67.6%)	45 (57%) 34 (43%)	0.14
Family income	High Average Low	8 (80%) 18 (45%) 7 (24.1%)	2 (20%) 22 (55%) 22 (75.9%)	10 (12.7%) 40 (50.6%) 29 (36.7%)	<0.01
Living Environment	Urban Rural	26 (50%) 7 (25.9%)	26 (50%) 20 (74.1%)	52 (65.8%) 27 (34.2%)	0.04
Maternal education	$\leq$ISCED 4 >ISCED 4	24 (36.4%) 9 (69.2%)	42 (63.6%) 4 (30.8%)	66 (83.5%) 13 (16.5%)	0.028
Paternal education	$\leq$ISCED 4 >ISCED 4	27 (39.1%) 6 (60%)	42 (60.9%) 4 (40%)	69 (87.3%) 10 (12.7%)	0.305

3.1.2. Disease Characteristics

Tables 3 and 4 show the results for each variable studied and the associated *p*-value for appropriate statistical tests (Chi-square or Fisher exact) for correlation with the glycemic control.

Table 3. Disease characteristics—categorical variables.

Variable	Categories	Frequency of Good Glycemic Control (% of Category)	Frequency of Poor Glycemic Control (% of Category)	Total (% of Grand Total)	*p*-Value
Family history of diabetes	Yes No	7 (38.9%) 26 (42.6%)	11 (61.1%) 35 (57.4%)	18 (22.8%) 61 (77.2%)	0.778
Type of onset (Ketoacidosis severity)	None Mild/moderate Severe	5 (62.5%) 23 (39.7%) 5 (38.5%)	3 (37.5%) 35 (60.3%) 8 (61.5%)	8 (10.1%) 58 (73.4%) 13 (16.5%)	0.384
A1C at onset	<7.5 g/dL $\geq$7.5 g/dL	8 (80%) 25 (36.2%)	2 (20%) 44 (63.8%)	10 (12.7%) 69 (87.3%)	0.011
Age of onset	<5 years 5–10 years >10 years	6 (68.4%) 10 (34.5%) 15 (51.7%)	13 (31.6%) 19 (65.5%) 14 (48.3%)	19 (24.7%) 29 (37.7%) 29 (37.7%)	0.275
Nutritional status at onset	Underweight Normal weight Overweight/obese	11 (64.7%) 18 (37.5%) 4 (28.6%)	6 (35.3%) 30 (62.5%) 10 (71.4%)	17 (21.5%) 48 (60.8%) 14 (17.7%)	0.08
Autoimmune diseases	Yes No	2 (33.3%) 31 (42.5%)	4 (66.7%) 42 (57.5%)	6 (7.6%) 73 (92.4%)	1.0
Hypoglicemia episodes	Yes No	4 (36.4%) 29 (42.6%)	7 (63.6%) 39 (57.4%)	11 (13.9%) 68 (86.1%)	0.754

Table 3. *Cont.*

Variable	Categories	Frequency of Good Glycemic Control (% of Category)	Frequency of Poor Glycemic Control (% of Category)	Total (% of Grand Total)	*p*-Value
Ketoacidosis episodes	Yes No	0 (0%) 33 (51.6%)	15 (100%) 31 (48.4%)	15 (19%) 64 (81%)	<0.01
Viral infections	Yes No	11 (30.6%) 22 (51.2%)	25 (69.4%) 21 (48.8%)	36 (45.6%) 43 (54.4%)	0.064
Bacterial infections	Yes No	11 (32. 4%) 22 (48.9%)	23 (67.6%) 23 (51.1%)	34 (43%) 45 (57%)	0.140
Microalbuminuria	Yes No	8 (32%) 25 (46.3%)	17 (68%) 29 (53.7%)	25 (31.6%) 54 (68.4%)	0.231
Serum Cholesterol	≥200 mg/dL <200 mg/dL	3 (12%) 30 (55.6%)	22 (88%) 24 (44.4%)	25 (31.6%) 54 (68.4%)	<0.01
Serum Triglycerides	≥150 mg/dL <150 mg/dL	2 (15.4%) 31 (47%)	11 (84.6%) 35 (53%)	13 (16.5%) 66 (83.5%)	0.035
Insulin Regimen	1 2	12 (34.3%) 21 (47.7%)	23 (65.7%) 23 (52.3%)	35 (44.3%) 44 (55.7%)	0.229

Table 4. Disease characteristics—continuous variables.

Variable	Descriptive Parameter	Glycemic Control		*p*-Value
		Good	Poor	
Disease duration (months)	Mean	68.09	90.65	0.04
	StdDev	57.78	58.27	
	IQR	64	87	
	MIN	6	7	
	MAX	252	252	
	95%CI	47.57–88.55	73.35–107.96	
BMI Z-score	Mean	−1.32	−0.32	0.01
	StdDev	1.55	1.72	
	IQR	2.47	2.52	
	MIN	−3.91	−4.61	
	MAX	1.59	2.93	
	95%CI	−1.87–−0.77	−0.83–0.19	
NLR	Mean	1.91	1.73	0.477
	StdDev	1.28	1	
	IQR	0.97	1.02	
	MIN	0.55	0.52	
	MAX	7.15	5.09	
	95%CI	1.46–2.37	1.43–2.03	
MPV (fL)	Mean	10.55	10.67	0.62
	StdDev	0.9	1.24	
	IQR	1	1.35	
	MIN	8.7	7.9	
	MAX	12.8	13.6	
	95%CI	10.22–10.87	10.3–11.04	

StdDev—standard deviation, IQR—interquartile range; MIN—minimum observed value; MAX—maximum observed value; 95%CI—95% confidence interval for the mean; BMI—body mass index; NLR—Neutrophil-to-lymphocyte ratio; MPV—mean platelet volume.

3.2. Binary Logistic Regression

The variables that correlated with glycemic control were the category of A1C at onset, with at least one episode of ketoacidosis (apart from disease onset), the BMI z-score (centered around the mean), low or average family income (dummy variable derived from

the initial categorical variable) and the last-documented value for total cholesterol above 200 mg/dL; this provided an adequate binary logistic regression model for predicting glycemic control. The model had an overall efficiency of 83.5% (87.9% for predicting adequate glycemic control and 80.4% for predicting poor glycemic control) and satisfactory goodness of fit (Hosmer–Lemeshow p-value = 0.655). Table 5 provides details regarding the statistical significance of the included parameters as well as the 95% confidence intervals for the regression coefficients, as calculated via the BCa method.

Table 5. Binary regression model.

Variable	β	*p*	BCa 95% CI for β	
			Lower	Higher
Ketoacidosis episodes (X1)	21.1	<0.01	17.62	39.2
A1C at onset ≥ 7.5 g/dL (X2)	3.12	<0.01	0.38	30.25
Family income low or average (X3)	2.73	0.018	0.32	57.02
Cholesterol ≥ 200 mg/dL (X4)	2.43	0.029	0.15	38.74
BMI Z-score (mean-centered) (X5)	0.58	<0.01	0.1	1.94
Constant	−5.57	<0.01	−24.84	−5.08

β—regression coefficient; BCa 95% CI for β—95% confidence interval for β calculated using the bias-corrected accelerated method; BMI (mean-centered)—body mass index centered around the mean of the sample.

Consequently, according to (1)–(4):

$$\text{logit}(\text{p}) = \ln\left(\frac{\text{p}}{1-\text{p}}\right) = -5.57 + 21.1\text{X}_1 + 3.12\text{X}_2 + 2.73\text{X}_3 + 2.43\text{X}_4 + 0.58\text{X}_5$$

This equation provides the log odds of the probability of poor glycemic control given the predictor values. The odds can then be converted to the probability using the following equation:

$$\text{p} = \frac{e^{\text{logit}(\text{p})}}{1 + e^{\text{logit}(\text{p})}}$$

3.3. Two-Step Clustering Algorithm

We implemented a two-step cluster algorithm using Akaike's Information Criterion. The variables included were maternal education, living environment, family income, ketoacidosis episodes, and the presence of elevated triglycerides. The resulting model defined four clusters with an average silhouette of cohesion separation at 0.6, indicative of good model quality. An overview of the model and its resulting clusters is presented in Table 6.

Table 6. Two-step cluster analysis overview.

Variable	Category	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Predictor Importance
Count	-	19 (24.1%)	19 (24.1%)	28 (35.4%)	13 (16.5%)	-
Higher maternal education	No Yes	19 (100%) 0 (0%)	19 (100%) 0 (0%)	28 (100%) 0 (0%)	0 (0%) 13 (100%)	1
Living environment	Rural Urban	8 (42.1%) 11 (57.9%)	19 (100%) 0 (0%)	0 (0%) 28 (100%)	0 (0%) 13 (100%)	0.73
Family income	Low Average High	9 (47.4%) 10 (52.6$) 0 (0%)	11 (57.9%) 8 (42.1%) 0 (0%)	9 (31.1%) 19 (67.9%) 0 (0%)	0 (0%) 3 (23.1%) 10 (76.9%)	0.67
Ketoacidosis episodes	Yes No	15 (78.9%) 4 (21.1%)	0 (0%) 19 (100%)	0 (0%) 28 (100%)	0 (0%) 13 (100%)	0.73
Elevated triglycerides	Yes No	12 (63.2%) 7 (36.8%)	0 19 (100%)	0 28 (100%)	1 (7.7%) 12 (92.3%)	0.49

A visual representation of the cluster characteristics is further provided in Figure 1.

Figure 1. Cluster comparison.

The frequency of poor glycemic control across clusters is presented in Figure 2. The differences observed are statistically significant ($p < 0.01$).

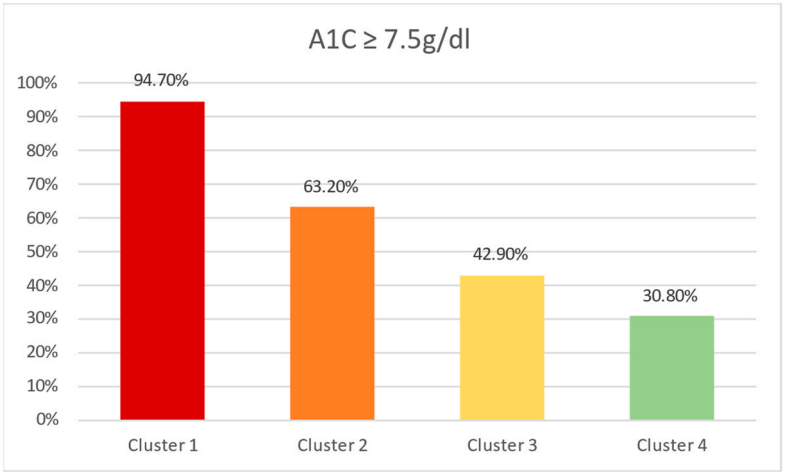

Figure 2. Frequency of poor glycemic control across clusters.

3.4. CART Decision Tree

The CART decision tree model was generated based on the following variables: disease duration (continuous), low family income (dichotomous), living conditions (dichotomous), A1C at onset (dichotomous), episodes of ketoacidosis (dichotomous), and elevated cholesterol (dichotomous). The resultant model is presented in Figure 3.

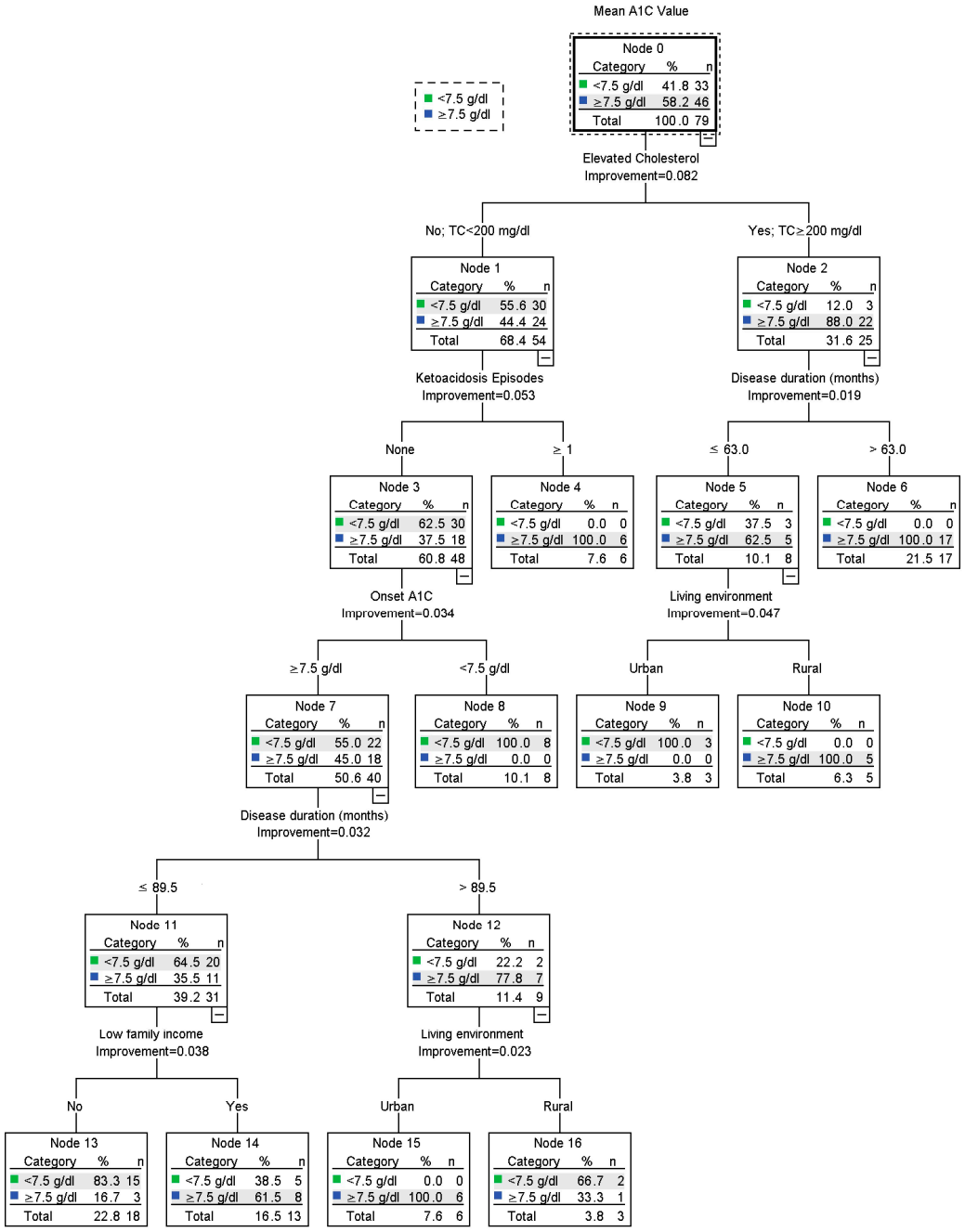

Figure 3. CART decision tree.

The CART decision tree showed an 88.6% overall accuracy (91.3% for predicting poor glycemic control and 84.8% for predicting adequate glycemic control). The decision paths indicated by our algorithm were distinguished between separate risk categories defined by a series of particular traits. Consequently, three risk levels were identified. Table 7 presents an overview of the identified risk categories, the characteristics of patients within each category, and the percentage of inadequate glycemic control observed in each analyzed subgroup.

Table 7. CART risk categories and subgroup characteristics composed.

Risk Category	Poor Glycemic Control (%)	Subgroup Characteristics
High	100%	22 patients with high cholesterol, 17 with a disease duration above 63 months, and 6 with a disease duration under 63 months but living in a rural environment.
		12 patients with normal cholesterol, 6 of which presented at least one episode of ketoacidosis (excluding onset) and 6 of which did not present any episodes of ketoacidosis but had an onset A1C above 7.5 g/dL, a disease duration above 89.5 months and lived in an urban environment.
Moderate	61.5%	13 patients with normal cholesterol levels, no ketoacidosis episodes, an A1C onset above 7.5 g/dL, a disease duration under 89.5 months, and a low family income.
	33.3%	3 patients with normal cholesterol levels, no ketoacidosis episodes, an onset of A1C above 7.5 g/dL, and disease duration above 89.5 months, originating from a rural environment.
Low	16.7%	18 patients with normal cholesterol levels, average or high family income, a disease duration of 89.5 months or under, an onset A1C of 7.5 g/dL or under, and no ketoacidosis episodes.
	0%	8 patients with normal cholesterol levels, no ketoacidosis episodes, and an onset A1C of 7.5 g/dL or under.
	0%	3 patients with high cholesterol levels had a disease duration of 63 months or less and lived in an urban environment.

3.5. Comparison and Agreement between Methods

3.5.1. CART vs. Regression

Figure 4 depicts the receiver operating characteristic (ROC) curve, which maps the predicted probabilities of inadequate glycemic control as determined by both the CART and binary regression methods. The AUROC for CART was 0.954, while the value for binary regression was 0.916.

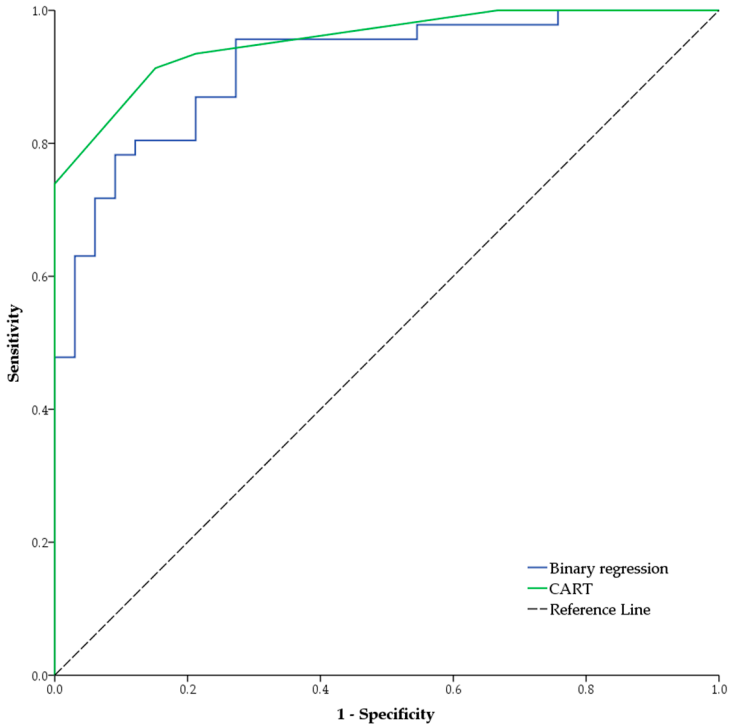

Figure 4. ROC for binary regression and CART.

3.5.2. Two-Step-Clustering vs. CART

Comparing the results from the CART decision tree and two-step clustering analysis, we evaluated the distribution of terminal nodes and their corresponding risk classes within the generated clusters. Risk classes were defined according to the proportion of patients exhibiting inadequate glycemic control within the node or the cluster. Accordingly, high risk was defined by a proportion of 66.6–100% (cluster 1 and nodes 4, 6, 10, 15), moderate risk as 33.3–66.5% (clusters 2 and 3 and nodes 14, 16, 13) and low risk as 0–33.2% (clusters 4 and nodes 8, 9).

A visual representation of patient distributions within the terminal nodes and across clusters is given in Table 8.

Table 8. Patient distribution across clusters and terminal nodes with corresponding risk class (% across column categories).

		High Risk			Medium Risk				Low Risk	
		Node 4	**Node 6**	**Node 10**	**Node 15**	**Node 14**	**Node 16**	**Node 13**	**Node 8**	**Node 9**
High risk	**Cluster 1**	6 (100%)	9 (52.9%)	2 (40%)	1 (16.7%)	1 (7.7%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Moderate risk	**Cluster 2**	0 (0%)	4 (23.5%)	3 (60%)	0 (0%)	5 (38.5%)	3 (100%)	2 (11.1%)	2 (25%)	0 (0%)
	Cluster 3	0 (0%)	2 (11.8%)	0 (0%)	4 (66.7%)	7 (53.8%)	0 (0%)	9 (50%)	5 (62.5%)	1 (33%)
Low risk	**Cluster 4**	0 (0%)	0 (0%)	0 (0%)	1 (16.7%)	0 (0%)	0 (0%)	7 (38.9%)	1 (12.5%)	2 (66%)

The kappa coefficient of agreement between the methods was 0.363, highlighting their capacity to achieve a 'fair agreement' when categorizing patients into the three risk classes ($p < 0.01$).

3.6. Handling Outliers

We excluded the four outliers in our data regarding disease duration (204 months or above). Nevertheless, cluster analysis revealed similar patterns regardless of the missing patients, while binary logistics regression had a marginally reduced overall performance (82.7% overall accuracy; 87.1% for predicting good glycemic control; 79.5% for predicting poor glycemic control; Hosmer–Lemeshow p-value = 0.308; Au-ROC = 0.908). Including the disease duration variable in the binary regression model with the outlier patients filtered out did not significantly contribute to its prediction capacities ($p = 0.208$).

When it came to the CART algorithm, removing outliers yielded a similar tree. Nevertheless, in the selection process, two of these predictors were swapped (elevated cholesterol and the presence of ketoacidosis). Notably, the cut-off values for disease duration in subsequent nodes remained unchanged, and the overall performance of the tree had a marginal reduction (overall accuracy: 88%, 83.9% for predicting good glycemic control, 90.9% for predicting poor glycemic control, AUROC-0.949).

4. Discussion

Our study included 79 pediatric patients with type 1 diabetes, for which we attempted to analyze potential predictors for adequate long-term glycemic control. This approach and sample size were successfully implemented before [61]. Our main findings, as revealed by bivariate analysis with references highlighting similar results, are showcased in Appendix A. Essentially, the bivariate analysis found correlations between poor glycemic control (A1C > 7.5 g/dL) and longer disease duration, a higher BMI Z-score, A1C at onset > 7.5 g/dL, lower family income, rural living environment, lower maternal education level, experiencing ketoacidosis episodes beyond the onset of disease, total cholesterol levels > 200 mg/dL and triglyceride levels > 150 m/dL. We further leveraged the potential of three distinct analytical methods—binary logistic regression, two-step clustering, and CART decision trees—to uncover predictors of long-term glycemic control. Each method offered unique advantages, revealing various layers of understanding while confirming the

influence of clinical and socioeconomic variables. Binary logistic regression found correlations between poor glycemic control and ketoacidosis episodes, A1C at onset > 7.5 g/dL, lower family income, total cholesterol levels >200 mg/dL, and BMI Z-score. Meanwhile, cluster analysis grouped patients with poor glycemic control together if they experienced ketoacidosis episodes, mothers lacking higher education, rural background, and triglyceride levels >150 m/dL. The CART decision tree further highlighted a higher risk for poor glycemic control with increasing disease duration, total cholesterol levels >200 mg/dL, lower family income, and rural living conditions.

Binary logistics regression quantified with an accuracy rate of 83.5% the significance of each predictor, notably identifying the strong association between ketoacidosis episodes ($\beta = 21.1$, $p < 0.01$) and uncontrolled A1C levels at onset ($\beta = 3.12$, $p < 0.01$) with poor glycemic control. Family income, elevated cholesterol, and BMI Z-score were also significant, showcasing the complex interplay between socioeconomic status, body weight, and glucose control. On a similar sample size but with arbitrary cut-off points based on the A1C values distribution documented in the study population., Niba et al. [14] showed a statistically independent association between having a mother as the primary caregiver ($\beta = -3.436$, $p < 0.01$) and good glycemic control.

Two-step clustering blended the strengths of both hierarchical and partitioning methods, offering an advanced technique for data categorization. By initially employing the k-means algorithm to form numerous small sub-clusters and subsequently using hierarchical agglomerative clustering, the algorithm ensured precise and interpretable groupings. With a good average silhouette of cohesion separation (>0.5), we are in agreement with Dalmaijer ES et al. [62], suggesting that sufficient statistical power can be achieved with relatively small samples. Our approach highlights the importance of socioeconomic factors and a nuanced interplay between socioeconomic factors and indicators in relation to disease progression and associated risk, which appear to synergistically influence each other. Cluster 1, marked by statistically significant poor glycemic control (94%), mainly comprises patients with at least one ketoacidosis episode (beyond onset), elevated triglycerides, and those who come from families with a low-to-average income lacking maternal higher education. Conversely, in Cluster 4, only 30.8% exhibited inadequate glycemic control. This group had protective factors like maternal tertiary education, urban living, and a high family income. None experienced ketoacidosis, and over 90% had normal triglyceride levels. Clusters 2 and 3 presented intermediate traits compared to Clusters 1 and 4. Both featured the absence of maternal higher education and had normal triglyceride levels. Cluster 2, however, demonstrated a greater prevalence of poor glycemic control and was entirely rural. Cluster 3 only included urban patients. Neither had high-income families. Nevertheless, Cluster 3 leaned toward an average family income, while Cluster 2 had more low-income families. This two-step clustering algorithm emphasized the profound interplay between socioeconomic and clinical factors in influencing glycemic control. The stark disparities between clusters underscore the imperative to understand and address both socioeconomic and clinical factors collectively in the pursuit of improving patient outcomes in diabetes care. Rohan et al. [63] sufficiently documented the importance of two-step cluster analysis in profiling diabetic children in terms of different self-management groups based on youth, maternal, and paternal reports. An analysis of variance indicated that the pattern of less optimal diabetes and self-management was associated with worse glycemic control, suggesting interventions based on these specific patterns of self-management to improve the management of diabetes.

CART decision trees provided a holistic, visual framework to identify complex interactions between predictors and outcomes. With an AUROC of 0.954, it slightly outperformed logistic regression's 0.916. CART, which pinpointed risk groups based on disease duration, cholesterol levels, living conditions, and family income, offered an 88.6% accuracy rate in predicting glycemic outcomes. CART decision trees are robust at handling small samples and yielding stable predictive performances above 10 entries per variable, as shown

in previous implementations studying various clinical dichotomous outcomes in small datasets [64–66].

The comparison of CART and two-step clustering revealed nuances in patient stratification. Some nodes and clusters showed strong agreement, while others demonstrated considerable variation. This variation suggests that while both methods can be used to stratify patients, they might capture different aspects or nuances of data, which can be beneficial depending on the clinical question or operational need.

The Kappa coefficient pointed out a fair agreement of 0.363 between the two methods, suggesting they capture different but complementary aspects of the data. This fair agreement might indicate the utility of employing both methods for a more comprehensive risk assessment, depending on the clinical question at hand or the research context.

Table 9 provides an overview of how each of the advanced methods implemented in our study impacted the interpretation of variables that correlated with glycemic control via bivariate analysis.

Table 9. Impact of advanced data processing methods on clinical interpretation.

Variable	Method	Key Findings	Clinical Implications
Ketoacidosis episodes	Binary logistic regression	Highest impact variable in binary logistics regression. Essential predictor is prevalent across all methods. Steep modulatory impact across the conglomeration of all other predictor categories in CART and two-step cluster analysis.	High impact predictor. Requires attentive management and assertive prevention.
	Two-step cluster analysis		
	CART		
Family Income	Binary logistic regression	Essential predictor is prevalent across all methods.	An integrative approach is necessary to handle the environmental conditions of pediatric patients with type 1 diabetes. In particular, interventions should be adapted to caretaker comprehension levels and available resources with tailored information and management strategies for each category.
	Two-step cluster analysis		
	CART		
Living environment	Two-step cluster analysis	Important predictor, prevalent across methods that employ patient categorization	
	CART		
Higher maternal education	Two-step cluster analysis	Important modulatory effect when taken in conjunction with other essential socioeconomic and disease-related characteristics	
A1C at onset	Binary logistics regression	Prevalent across prediction algorithms	Disease onset can predict future outcomes. More aggressive screening strategies may be warranted.
	CART		
Elevated Cholesterol	Binary logistics regression	Prevalent across prediction algorithms	An integrated approach to lifestyle management and risk factor mitigation is essential. While some predictors may have a smaller impact on glycemic control, their modifiable nature presents unique therapeutic opportunities for an overall increased effect on clinical outcomes.
	CART		
Elevated Triglycerides	Two-step cluster analysis	Modulatory effect in conjunction with other disease characteristics and socioeconomic status	
BMI Z-score	Binary logistics regression	Low impact predictor, but present nonetheless	

Table 9. *Cont.*

Variable	Method	Key Findings	Clinical Implications
Disease duration	CART	Modulatory impact on disease profiles	Dynamic changes within the disease patterns require attentive monitoring of pediatric patients with type 1 diabetes. Treatment up-scaling may take this aspect into consideration.

Notably, all three methods underscored the influence of socioeconomic factors on glycemic outcomes. However, their emphasis was varied and nuanced; they played a pivotal role in two-step clustering, which included all three of the variables related to socioeconomic status. Family income was consistently present in all three techniques while living conditions featured in only two of the methods employed. In addition, the recurrent appearance of ketoacidosis across all methods underscored its severity in disease manifestation within our studied group. CART analysis uniquely emphasized disease duration, suggesting stratified exposures to determinants and hinting at a more heterogeneous patient profile. While the BMI Z-score was confined to regression, the inclusion of triglycerides in clustering underscored socioeconomic impacts on health. Elevated cholesterol was a generic predictor in regression, but in CART, it either intensified vulnerabilities or became less prominent against stronger determinants. Suboptimal A1C at onset played a significant role in regression, but in CART, it relied on interactions like disease duration and environmental factors.

Strengths and Limitations

The main limitation of our study is related to the small sample size. Nevertheless, our results align with other reports regarding performance and adaptability to small sample sizes [64–67]. We overcame this drawback by leveraging the capacity of machine-learning algorithms to generate patterns tailored to unique clinical interpretations. However, another issue relates to the retrospective nature of this study.

Retrospective studies are inherently susceptible to certain biases. Recall bias occurs when collected data are dependent upon the memory of the participants involved. This may impact variables such as family income or the education levels of parents. Nevertheless, these records were carefully checked for consistency with the general practitioners' databases if proof of income or education level was missing at admittance. Future research in prospective studies could mitigate these aspects by asking participants to provide proof of income or education level. Recording and selection bias are also capable of influencing data. Recording bias refers to the fact that the methods of recording data were not intended for the research question addressed by our study but rather for the clinical interpretation of each individual case. The studied variables, however, employed standardized quantification methods used in the Pediatric Clinical Hospital of Sibiu, in particular referring to laboratory and anthropometric measurements. Selection bias may have arisen due to the inclusion only of complete datasets, as previously described. This may impact the generalizability of our findings. However, two key points should be considered. Firstly, our results are in agreement with previous findings on larger datasets concerning the correlations found between glycemic control and its predictors. Secondly, the main purpose of this study was to showcase the potential role of advanced statistical methods and machine-learning algorithms to enhance data interpretation. While our findings require larger datasets from multiple centers to provide a broader, more generalizable overview of glycemic control predictors in pediatric type 1 diabetic patients, they still provide valuable input pertaining to the particular population in Sibiu County. More importantly, our design offers a solid framework for approaching similar research questions in different datasets—both in small samples whereby the results may influence local practices or larger ones where the purpose is the generalization of a conclusion.

Another point in terms of strengths and limitations relates to the intrinsic discrimination capacity of each method, given their distinct analytical perspectives. Binary logistics regression is notable for its ability to evaluate multiple predictors concurrently, offering a holistic predictive model. For instance, in our sample, the occurrence of ketoacidosis episodes markedly overshadowed other factors like the onset of A1C, family income, cholesterol levels, or BMI Z-score in terms of influence. However, the comprehensive insight provided by binary regression remains invaluable for each variable considered. While ketoacidosis is a significant complication in diabetes and a critical indicator of glycemic control, its prevention is already a key management target [38,39]. Other predictors like the onset of A1C and family income, despite their substantial impact, offer limited intervention opportunities. Arguably, more diligent screening programs could exert some influence on the values of A1C onset; still, such initiatives are notoriously cumbersome. Cholesterol levels and weight status, despite having modest effects, offer promising paths for therapeutic intervention to potentially improve glycemic control outcomes. Importantly, while binary logistic regression can be influenced by outliers, consistent results upon outlier exclusion in our analysis addressed this concern. By contrast, the CART algorithm has a sequential, hierarchical strategy to craft detailed patient subgroups, which are crucial for personalized clinical interventions. Our algorithm highlights the importance of factors akin to those in binary regression yet facilitates a deeper patient segmentation in relation to the living environment and disease duration. Notably, the inclusion of disease duration was insightful, especially given its outlier values, exemplifying CART's robustness against such extremes. This resilience is rooted in CART's methodology, prioritizing node purity enhancement with each division based on the decision variable's optimal cut-off values. However, the risk of overfitting, especially in smaller samples, remains an acknowledged limitation of CART. To alleviate this, we employed CART in the pruning mode. In parallel, two-step clustering analysis effectively identifies patient subgroups but can be skewed by outliers due to its reliance on centroid calculations. Nevertheless, it did not select the variable with outliers(disease duration) prior to or after extreme values have been removed. Two-step clustering analysis revealed interplays between disease characteristics and socio-economic factors, solidifying ketoacidosis episodes as a primary predictor and highlighting unique variables like maternal education level and triglyceride levels. This insight is invaluable for tailoring interventions, considering lifestyle factors and caregiver comprehension capacities. When viewed collectively, the complementary nature of these methods is clear. Each has its vulnerabilities, but their combined use in our study helped mitigate these issues, reinforcing our conclusions' reliability.

An integrated approach in clinical decision-making allows clinicians to provide tailored care for specific patient subgroups. In our study, each method brought a unique perspective on the clinical data. Two-step clustering-segmented patients based on common traits emphasized the socioeconomic and clinical influences on glycemic control. Binary logistic regression predicted the likelihood of a poor control using specific variables, quantifying each predictor's impact. The CART decision tree blended both segmentation and prediction, visually representing patient profiles and showing how combined factors influence outcomes step by step.

All three methods offer insights into patient profiles, each adding its unique layer to data interpretation and aiding clinicians in personalized care for better diabetes outcomes. Specific populations may respond better to tailored interventions, as described, for example, in patients with type 2 diabetes, whereby low-income patients show improvements in glycemic control upon their inclusion in chronic care management programs, which include patient visits and education [68].

Our results highlight the diverse factors influencing glycemic control, from clinical to socioeconomic aspects. Future research should investigate why methods differ in patient groups and consider including more variables or qualitative data for deeper understanding.

5. Conclusions

Exploring pediatric patients with type 1 diabetes requires a holistic understanding of the determinants of glycemic control that have emerged based on advanced statistical methods. Binary logistics regression quantified relationships, the two-step clustering algorithm underscored the importance of socioeconomic and clinical attributes, while the CART decision tree revealed the interplay between intricate variables. Together, these methodologies emphasize the concept of complementarity, given that no single approach can capture by itself a complete view of such complex interactions. The variations observed in stratifications between clustering and CART analysis highlight the richness of clinical data and underscore the need for an integrated analytical framework. Future research should consider combining these techniques, delving deeper into these nuances, and seeking to merge qualitative and quantitative insights. Such an integrated approach holds promise for a more tailored and evidence-based care paradigm in pediatric type 1 diabetes management.

Author Contributions: Conceptualization, B.N. and M.O.N.; methodology, B.N. and M.O.N.; software, B.N. and M.O.N.; validation, B.N.; formal analysis, B.N.; investigation, B.N. and I.N.; resources, B.N.; data curation, B.N. and M.O.N.; writing—original draft preparation, B.N.; writing—review and editing, B.N., M.O.N. and I.N.; visualization, B.N. and M.O.N.; supervision, B.N.; project administration, B.N.; funding acquisition, B.N. All authors have read and agreed to the published version of the manuscript.

Funding: Project financed from Lucian Blaga University of Sibiu research grants LBUS-IRG-2018-04.

Data Availability Statement: The data presented in this study are available upon reasonable request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A. Results Obtained through Bivariate Analysis

In our study, we found no differences between genders concerning glycemic control. Mohammad et al. [13] found similar results, although when stratifying patients by age and sex, girls above the age of 15 had a higher prevalence of inadequate glycemic control when compared to boys in the same age category. Their study included a sample much larger than ours, thus permitting an adequate analysis of such sub-strata of the population. Springer et al. found the female sex to be associated with poor glycemic control [23], while Noorani et al. found no such connection [61], and other studies, such as the one conducted by Niba et al., found girls to have a lower A1C without, however, reaching statistical significance [14]. It is difficult to draw a hard conclusion based on the available inconsistent data. Multiple factors pertaining to, but not limited to, geographical and cultural particularities, as well as a great variability in physiological changes attributable to growth, especially during puberty, can all contribute to conflicting results. Further epidemiological studies may shed light on these aspects.

Socioeconomic factors, including higher family income, urban versus rural environment, and maternal level of education, showed a significant correlation with better glycemic control in our study. Interestingly, the paternal level of education did not show any correlation to glycemic control. The relationship between good glycemic control and maternal involvement in the caregiving of patients with type 1 diabetes has been previously described in the literature [14,61]. The greater influence of maternal education in these patients is, therefore, somewhat to be expected. With regard to socioeconomic status and family income, there is a well-documented correlation to glycemic control in the literature [23–27]. In addition, a correlation between rural environment and poorer glycemic control has also been documented, particularly in patients under the age of 26 [69].

A family history of diabetes showed no correlation to glycemic control in our study. This is a similar finding to Niba et al. [14]. Fredheim et al. found that a positive family history of diabetes was correlated with a lower likelihood of presenting with diabetic ketoacidosis but a higher AIC level over time [28]. Larger samples taking into account more refined characteristics concerning diabetes type, evolution, and control within the family may be warranted in order to clarify these aspects. Disease awareness (rather than mere presence) within the family may play a fundamental role in this regard as a confounding factor.

We stratified our patients according to the type of onset in a similar fashion to Fredheim et al. [28]. In contrast to their study, we found no significant correlation between onset type and glycemic control. This is probably due to the comparatively smaller sample we used, which did not provide enough statistical power to reach the same conclusions as Fredheim et al.

A1C levels at onset were a strong predictor for long-term glycemic control in our study. This result is consistent with other findings in the literature, which show that early glycemic control has a great prognostic value for the subsequent course of type 1 diabetes [22,30–33].

In relation to the age of onset, we stratified our patients in a similar fashion to Mohammad et al. [13]. Their study concluded that patients with an onset of diabetes under 5 years of age were more likely to have better glycemic control. We did not, however, find any significant correlation between the onset age category and glycemic control. On the one hand, inconsistent data might arise due to small sample sizes. This applies both to our study as well as to the one conducted by Mohammad et al., where the group of patients with an onset of T1D under the age of 5 was represented by only 23 patients. On the other hand, there is also a lack of consensus in this regard in the literature. For example, Samuelsson et al. found that patients with a lower age at T1D onset had higher AIC values during follow-up [22], in opposition to the results presented by Mohammad et al., while Svensson et al. had findings similar to ours, with no differences among onset age categories which closely resembled the ones we utilized [33]. Further studies geared at clarifying these aspects could shed more light on existing data, particularly if conducted with the goal of adjusting for various confounding factors that may interfere with the results.

However, weight status at disease onset has been previously linked to glycemic control, with inconsistent results [13,35]. In our study, patients with poor glycemic control had a higher BMI z-score than those with adequate control.

We found increasing disease duration to have a significant correlation with poor glycemic control. McKinney et al. found a similar link with their results indicating increasing mean AIC levels with increasing disease duration [36], similar to Carter et al. [24] and Hiliard et al. [37].

Various autoimmune diseases are frequently associated with T1D [38]. Our study did not, however, find a correlation between glycemic control and the presence of these diseases.

Attempting to obtain glycemic control in T1D patients inherently carries the potential of hypoglycemia episodes. Existing data suggest a possible correlation between lower A1C levels and the number of hypoglycemia events [39]. Our study did not, however, show a correlation between glycemic control and the presence or absence of hypoglycemic events during follow-up. This may be attributed to the small sample size in conjunction with a more cautious approach to insulin therapy. Ketoacidosis episodes, on the other hand, are acute manifestations of improper glycemic control and have been found to correlate with poorer A1C values in T1D patients [39,40]. Our study found similar results.

Microalbuminuria is a potential manifestation of nephropathy that is associated with type 1 diabetes and constitutes part of the spectrum of microvascular complications related to this disease [21]. Virk et al. found A1C variability to correlate with the microvascular complications of T1D [42]. In our study, glycemic control did not correlate with the presence of microalbuminuria. This is most probably due to inconsistent screening protocols and a small sample size.

The systemic dysregulation characteristic of type 1 diabetes often extends beyond the metabolism of glucose. As such, type 1 diabetes may be associated with a complex dysmetabolic aggregate, which may sometimes be associated with elevated circulating lipids. As an image of this spectrum, high-serum cholesterol and triglycerides have been previously shown to correlate with inadequate glycemic control in patients with type 1 diabetes [13,35]. Our results concur with these findings.

An interplay between viral or bacterial infections and type 1 diabetes has been previously described in the literature [41]. Our study did not find a correlation between the presence of one or more viral or bacterial infections requiring hospital admission and glycemic control. While a dysregulation in glucose metabolism attributable to the acute changes characteristic of infectious diseases is to be somewhat expected, multiple factors may come into play when considering their long-term effects on glycemic control in type 1 diabetes patients. The frequency of episodes that require hospitalization within a certain time frame may be more relevant than their dichotomous presence or absence. However, our study did not have a large enough sample to evaluate this aspect. Further studies in this direction may bring a better understanding of the subject.

The neutrophil-to-lymphocyte ratio has been proposed as a predictor for ketoacidosis in T1D patients [44–46]. Our study showed no differences between glycemic control groups with regard to this variable. This may be due to the limits imposed by our small sample size. The same can most probably be stated concerning the mean platelet volume, a marker for which we found no prognostic value in glycemic control, although there are data to suggest a possible connection in this regard [47–49].

Treatment regimens did not, in our study, correlate with glycemic control. Niba et al. found that a treatment regimen based on two administrations a day was associated with better glycemic control compared with multiple (three or more) injections a day [14]. These results may, however, be biased due to the upscaling of treatment targeted toward patients who do not achieve optimal control with only two injections a day. Mohammad et al. found a higher prevalence of adequate glycemic control in patients with one basal dose of insulin and three injections of regular insulin every day when compared to either two injections of premixed intermediate-acting and regular insulin daily or two injections of intermediate-acting insulin associated with one or more injections of regular insulin daily [13]. Alemzadeh et al. found that flexible therapy with multiple insulin administrations improved glycemic control and reduced hypoglycemic episodes [43]. Finally, insulin pumps have been shown to contradict evidence in the literature. Svoren et al., for example, found lower A1C values in patients using these pumps and fewer hypoglycemic events [70], while Holl et al. found no correlation between insulin pump use or multiple insulin injections and A1C values [71]. The lack of a general rule and the conflicting evidence point toward the necessity of relying on clinical judgment in every case to find a tailored approach for each patient. The interplay between the numerous factors that intervene in achieving optimal glycemic control is a key element in guiding treatment decisions.

References

1. DiMeglio, L.A.; Evans-Molina, C.; Oram, R.A. Type 1 Diabetes. *Lancet* **2018**, *391*, 2449–2462. [CrossRef] [PubMed]
2. Huo, L.; Harding, J.L.; Peeters, A.; Shaw, J.E.; Magliano, D.J. Life Expectancy of Type 1 Diabetic Patients during 1997–2010: A National Australian Registry-Based Cohort Study. *Diabetologia* **2016**, *59*, 1177–1185. [CrossRef] [PubMed]
3. Petrie, D.; Lung, T.W.C.; Rawshani, A.; Palmer, A.J.; Svensson, A.-M.; Eliasson, B.; Clarke, P. Recent Trends in Life Expectancy for People with Type 1 Diabetes in Sweden. *Diabetologia* **2016**, *59*, 1167–1176. [CrossRef] [PubMed]
4. American Diabetes Association Professional Practice Committee. 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2020. *Diabetes Care* **2020**, *43* (Suppl. S1), S14–S31. [CrossRef]
5. Cho, N.H.; Shaw, J.E.; Karuranga, S.; Huang, Y.; da Rocha Fernandes, J.D.; Ohlrogge, A.W.; Malanda, B. IDF Diabetes Atlas: Global Estimates of Diabetes Prevalence for 2017 and Projections for 2045. *Diabetes Res Clin Pr.* **2018**, *138*, 271–281. [CrossRef] [PubMed]
6. Heier, M.; Margeirsdottir, H.D.; Brunborg, C.; Hanssen, K.F.; Dahl-Jørgensen, K.; Seljeflot, I. Inflammation in Childhood Type 1 Diabetes; Influence of Glycemic Control. *Atherosclerosis* **2015**, *238*, 33–37. [CrossRef]

7. Gourgari, E.; Dabelea, D.; Rother, K. Modifiable Risk Factors for Cardiovascular Disease in Children with Type 1 Diabetes: Can Early Intervention Prevent Future Cardiovascular Events? *Curr. Diabetes Rep.* **2017**, *17*, 134. [CrossRef]
8. Hafez, M.; Hassan, M.; Musa, N.; Abdel Atty, S.; Azim, S.A. Vitamin D Status in Egyptian Children with Type 1 Diabetes and the Role of Vitamin D Replacement in Glycemic Control. *J. Pediatr. Endocrinol. Metab.* **2017**, *30*, 389–394. [CrossRef]
9. Jaiswal, M.; Divers, J.; Dabelea, D.; Isom, S.; Bell, R.A.; Martin, C.L.; Pettitt, D.J.; Saydah, S.; Pihoker, C.; Standiford, D.A.; et al. Prevalence of and Risk Factors for Diabetic Peripheral Neuropathy in Youth with Type 1 and Type 2 Diabetes: SEARCH for Diabetes in Youth Study. *Diabetes Care* **2017**, *40*, 1226–1232. [CrossRef]
10. Savastio, S.; Cadario, F.; Genoni, G.; Bellomo, G.; Bagnati, M.; Secco, G.; Picchi, R.; Giglione, E.; Bona, G. Vitamin D Deficiency and Glycemic Status in Children and Adolescents with Type 1 Diabetes Mellitus. *PLoS ONE* **2016**, *11*, e0162554. [CrossRef]
11. Ordooei, M.; Shojaoddiny-Ardekani, A.; Hoseinipoor, S.H.; Miroliai, M.; Zare-Zardini, H. Effect of Vitamin D on HbA1c Levels of Children and Adolescents with Diabetes Mellitus Type 1. *Minerva Pediatr.* **2017**, *69*, 391–395. [CrossRef]
12. Rewers, M.; Ludvigsson, J. Environmental Risk Factors for Type 1 Diabetes. *Lancet* **2016**, *387*, 2340–2348. [CrossRef]
13. Mohammad, H.; Farghaly, H.; Metwalley, K.; Monazea, E.; Abd El-Hafeez, H. Predictors of Glycemic Control in Children with Type 1 Diabetes Mellitus in Assiut-Egypt. *Indian. J. Endocrinol. Metab.* **2012**, *16*, 796. [CrossRef]
14. Niba, L.L.; Aulinger, B.; Mbacham, W.F.; Parhofer, K.G. Predictors of Glucose Control in Children and Adolescents with Type 1 Diabetes: Results of a Cross-Sectional Study in Cameroon. *BMC Res. Notes* **2017**, *10*, 207. [CrossRef] [PubMed]
15. Chiang, J.L.; Kirkman, M.S.; Laffel, L.M.B.; Peters, A.L. Type 1 Diabetes Through the Life Span: A Position Statement of the American Diabetes Association. *Diabetes Care* **2014**, *37*, 2034–2054. [CrossRef] [PubMed]
16. DiMeglio, L.A.; Acerini, C.L.; Codner, E.; Craig, M.E.; Hofer, S.E.; Pillay, K.; Maahs, D.M. ISPAD Clinical Practice Consensus Guidelines 2018: Glycemic Control Targets and Glucose Monitoring for Children, Adolescents, and Young Adults with Diabetes. *Pediatr. Diabetes* **2018**, *19*, 105–114. [CrossRef]
17. Cutfield, S.W.; Derraik, J.G.B.; Reed, P.W.; Hofman, P.L.; Jefferies, C.; Cutfield, W.S. Early Markers of Glycaemic Control in Children with Type 1 Diabetes Mellitus. *PLoS ONE* **2011**, *6*, e25251. [CrossRef] [PubMed]
18. World Health Organisation. WHO Expert Committee on Diabetes Mellitus: Second Report. *World Health Organ. Tech. Rep. Ser.* **1980**, *646*, 1–80.
19. Urbach, S.L.; LaFranchi, S.; Lambert, L.; Lapidus, J.A.; Daneman, D.; Becker, T.M. Predictors of Glucose Control in Children and Adolescents with Type 1 Diabetes Mellitus. *Pediatr. Diabetes* **2005**, *6*, 69–74. [CrossRef] [PubMed]
20. Eurostat. International Standard Classification of Education (ISCED). Available online: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=International_Standard_Classification_of_Education_(ISCED) (accessed on 22 September 2023).
21. Beck, R.W.; Tamborlane, W.V.; Bergenstal, R.M.; Miller, K.M.; DuBose, S.N.; Hall, C.A. The T1D Exchange Clinic Registry. *J. Clin. Endocrinol. Metab.* **2012**, *97*, 4383–4389. [CrossRef] [PubMed]
22. Samuelsson, U.; Steineck, I.; Gubbjornsdottir, S. A High Mean-HbA1c Value 3-15 Months after Diagnosis of Type 1 Diabetes in Childhood Is Related to Metabolic Control, Macroalbuminuria, and Retinopathy in Early Adulthood-a Pilot Study Using Two Nation-Wide Population Based Quality Registries. *Pediatr. Diabetes* **2014**, *15*, 229–235. [CrossRef]
23. Springer, D.; Dziura, J.; Tamborlane, W.V.; Steffen, A.T.; Ahern, J.H.; Vincent, M.; Weinzimer, S.A. Optimal Control of Type 1 Diabetes Mellitus in Youth Receiving Intensive Treatment. *J. Pediatr.* **2006**, *149*, 227–232. [CrossRef] [PubMed]
24. Carter, P.J.; Cutfield, W.S.; Hofman, P.L.; Gunn, A.J.; Wilson, D.A.; Reed, P.W.; Jefferies, C. Ethnicity and Social Deprivation Independently Influence Metabolic Control in Children with Type 1 Diabetes. *Diabetologia* **2008**, *51*, 1835–1842. [CrossRef]
25. Gallegos-Macias, A.R.; Macias, S.R.; Kaufman, E.; Skipper, B.; Kalishman, N. Relationship between Glycemic Control, Ethnicity and Socioeconomic Status in Hispanic and White Non-Hispanic Youths with Type 1 Diabetes Mellitus. *Pediatr. Diabetes* **2003**, *4*, 19–23. [CrossRef]
26. Overstreet, S.; Holmes, C.S.; Dunlap, W.P.; Frentz, J. Sociodemographic Risk Factors to Disease Control in Children with Diabetes. *Diabet. Med.* **1997**, *14*, 153–157. [CrossRef]
27. Hassan, K.; Loar, R.; Anderson, B.J.; Heptulla, R.A. The Role of Socioeconomic Status, Depression, Quality of Life, and Glycemic Control in Type 1 Diabetes Mellitus. *J. Pediatr.* **2006**, *149*, 526–531. [CrossRef] [PubMed]
28. Fredheim, S.; Johannesen, J.; Johansen, A.; Lyngsøe, L.; Rida, H.; Andersen, M.L.M.; Lauridsen, M.H.; Hertz, B.; Birkebæk, N.H.; Olsen, B.; et al. Diabetic Ketoacidosis at the Onset of Type 1 Diabetes Is Associated with Future HbA1c Levels. *Diabetologia* **2013**, *56*, 995–1003. [CrossRef]
29. Duca, L.M.; Wang, B.; Rewers, M.; Rewers, A. Diabetic Ketoacidosis at Diagnosis of Type 1 Diabetes Predicts Poor Long-Term Glycemic Control. *Diabetes Care* **2017**, *40*, 1249–1255. [CrossRef] [PubMed]
30. Viswanathan, V.; Sneeringer, M.R.; Miller, A.; Eugster, E.A.; DiMeglio, L.A. The Utility of Hemoglobin A1c at Diagnosis for Prediction of Future Glycemic Control in Children with Type 1 Diabetes. *Diabetes Res. Clin. Pr.* **2011**, *92*, 65–68. [CrossRef]
31. Shalitin, S.; Phillip, M. Which Factors Predict Glycemic Control in Children Diagnosed with Type 1 Diabetes before 6.5 Years of Age? *Acta Diabetol.* **2012**, *49*, 355–362. [CrossRef] [PubMed]
32. Rudberg, S.; Ullman, E.; Dahlquist, G. Relationship between Early Metabolic Control and the Development of Microalbuminuria? A Longitudinal Study in Children with Type 1 (Insulin-Dependent) Diabetes Mellitus. *Diabetologia* **1993**, *36*, 1309–1314. [CrossRef]
33. Svensson, M.; Eriksson, J.W.; Dahlquist, G. Early Glycemic Control, Age at Onset, and Development of Microvascular Complications in Childhood-Onset Type 1 Diabetes. *Diabetes Care* **2004**, *27*, 955–962. [CrossRef]

34. Borghi, E.; de Onis, M.; Garza, C.; Van den Broeck, J.; Frongillo, E.A.; Grummer-Strawn, L.; Van Buuren, S.; Pan, H.; Molinari, L.; Martorell, R.; et al. Construction of the World Health Organization Child Growth Standards: Selection of Methods for Attained Growth Curves. *Stat. Med.* **2006**, *25*, 247–265. [CrossRef] [PubMed]
35. Guy, J.; Ogden, L.; Wadwa, R.P.; Hamman, R.F.; Mayer-Davis, E.J.; Liese, A.D.; D'Agostino, R.; Marcovina, S.; Dabelea, D. Lipid and Lipoprotein Profiles in Youth with and without Type 1 Diabetes. *Diabetes Care* **2009**, *32*, 416–420. [CrossRef] [PubMed]
36. McKinney, P.A.; Feltbower, R.G.; Stephenson, C.R. Children and Young People with Diabetes in Yorkshire: A Population Based Clinical Audit of Patient Data 2005/6. *Diabet. Med.* **2008**, *25*, 1276–1282. [CrossRef] [PubMed]
37. Hilliard, M.E.; Wu, Y.P.; Rausch, J.; Dolan, L.M.; Hood, K.K. Predictors of Deteriorations in Diabetes Management and Control in Adolescents with Type 1 Diabetes. *J. Adolesc. Health* **2013**, *52*, 28–34. [CrossRef]
38. Hughes, J.W.; Riddlesworth, T.D.; DiMeglio, L.A.; Miller, K.M.; Rickels, M.R.; McGill, J.B. Autoimmune Diseases in Children and Adults with Type 1 Diabetes from the T1D Exchange Clinic Registry. *J. Clin. Endocrinol. Metab.* **2016**, *101*, 4931–4937. [CrossRef]
39. Rewers, A. Predictors of Acute Complications in Children with Type 1 Diabetes. *JAMA* **2002**, *287*, 2511. [CrossRef]
40. Fritsch, M.; Rosenbauer, J.; Schober, E.; Neu, A.; Placzek, K.; Holl, R.W. Predictors of Diabetic Ketoacidosis in Children and Adolescents with Type 1 Diabetes. Experience from a Large Multicentre Database. *Pediatr. Diabetes* **2011**, *12 Pt 1*, 307–312. [CrossRef]
41. Guglielmi, C.; Leslie, R.D.; Pozzilli, P. Epidemiology and Risk Factors of Type 1 Diabetes. In *Diabetes. Epidemiology, Genetics, Pathogenesis, Diagnosis, Prevention, and Treatment*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 41–54. [CrossRef]
42. Virk, S.A.; Donaghue, K.C.; Cho, Y.H.; Benitez-Aguirre, P.; Hing, S.; Pryke, A.; Chan, A.; Craig, M.E. Association between HbA 1c Variability and Risk of Microvascular Complications in Adolescents with Type 1 Diabetes. *J. Clin. Endocrinol. Metab.* **2016**, *101*, 3257–3263. [CrossRef]
43. Alemzadeh, R.; Berhe, T.; Wyatt, D.T. Flexible Insulin Therapy with Glargine Insulin Improved Glycemic Control and Reduced Severe Hypoglycemia among Preschool-Aged Children with Type 1 Diabetes Mellitus. *Pediatrics* **2005**, *115*, 1320–1324. [CrossRef]
44. Singh, K.; Martinell, M.; Luo, Z.; Espes, D.; Stålhammar, J.; Sandler, S.; Carlsson, P.-O. Cellular Immunological Changes in Patients with LADA Are a Mixture of Those Seen in Patients with Type 1 and Type 2 Diabetes. *Clin. Exp. Immunol.* **2019**, *197*, 64–73. [CrossRef] [PubMed]
45. Cheng, Y.; Yu, W.; Zhou, Y.; Zhang, T.; Chi, H.; Xu, C. Novel Predictor of the Occurrence of DKA in T1DM Patients without Infection: A Combination of Neutrophil/Lymphocyte Ratio and White Blood Cells. *Open Life Sci.* **2021**, *16*, 1365–1376. [CrossRef] [PubMed]
46. Scutca, A.-C.; Nicoară, D.-M.; Mărăzan, M.; Brad, G.-F.; Mărginean, O. Neutrophil-to-Lymphocyte Ratio Adds Valuable Information Regarding the Presence of DKA in Children with New-Onset T1DM. *J. Clin. Med.* **2022**, *12*, 221. [CrossRef] [PubMed]
47. Baghersalimi, A.; Koohmanaee, S.; Darbandi, B.; Farzamfard, V.; Hassanzadeh Rad, A.; Zare, R.; Tabrizi, M.; Dalili, S. Platelet Indices Alterations in Children with Type 1 Diabetes Mellitus. *J. Pediatr. Hematol. Oncol.* **2019**, *41*, e227–e232. [CrossRef]
48. Söbü, E.; Demir Yenigürbüz, F.; Özçora, G.D.K.; Köle, M.T. Evaluation of the Impact of Glycemic Control on Mean Platelet Volume and Platelet Activation in Children with Type 1 Diabetes. *J. Trop. Pediatr.* **2022**, *68*, fmac063. [CrossRef]
49. Pirgon, O.; Asya Tanju, I.; Alev Erikci, A. Association of Mean Platelet Volume between Glucose Regulation in Children with Type 1 Diabetes. *J. Trop. Pediatr.* **2007**, *55*, 63–64. [CrossRef]
50. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: Stanford, CA, USA, 2008.
51. Lin, Y.-Q.; Zhang, Y.-S.; Tian, G.-L.; Ma, C.-X. Fast QLB Algorithm and Hypothesis Tests in Logistic Model for Ophthalmologic Bilateral Correlated Data. *J. Biopharm. Stat.* **2021**, *31*, 91–107. [CrossRef]
52. Davison, A.C.; Hinkley, D.V. *Bootstrap Methods and Their Application*; Cambridge University Press: Cambridge, UK, 1997. [CrossRef]
53. Carpenter, J.; Bithell, J. Bootstrap Confidence Intervals: When, Which, What? A Practical Guide for Medical Statisticians. *Stat. Med.* **2000**, *19*, 1141–1164. [CrossRef]
54. Șchiopu, D. Applying TwoStep Cluster Analysis for Identifying Bank Customer's Profile. *Econ. Insights—Trends Chall.* **2010**, *62*, 66–75.
55. Tan, P.-N.; Steinbach, M. *Introduction to Data Mining*, 2nd ed.; Pearson: London, UK, 2018.
56. Jain, A.K.; Murty, M.N.; Flynn, P.J. Data Clustering. *ACM Comput. Surv.* **1999**, *31*, 264–323. [CrossRef]
57. Rousseeuw, P.J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
58. Conn, D.; Ramirez, C.M. Random Forests and Fuzzy Forests in Biomedical Research. In *Computational Social Science*; Cambridge University Press: Cambridge, UK, 2016; pp. 168–196. [CrossRef]
59. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning with Applications in R*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2013.
60. Landis, J.R.; Koch, G.G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **1977**, *33*, 159. [CrossRef]
61. Noorani, M.; Ramaiya, K.; Manji, K. Glycaemic Control in Type 1 Diabetes Mellitus among Children and Adolescents in a Resource Limited Setting in Dar Es Salaam—Tanzania. *BMC Endocr. Disord.* **2016**, *16*, 29. [CrossRef] [PubMed]
62. Dalmaijer, E.S.; Nord, C.L.; Astle, D.E. Statistical Power for Cluster Analysis. *BMC Bioinform.* **2022**, *23*, 205. [CrossRef]
63. Rohan, J.M.; Delamater, A.; Pendley, J.S.; Dolan, L.; Reeves, G.; Drotar, D. Identification of Self-Management Patterns in Pediatric Type 1 Diabetes Using Cluster Analysis. *Pediatr. Diabetes* **2011**, *12*, 611–618. [CrossRef] [PubMed]

64. Lu, M.-Y.; Liu, T.-W.; Liang, P.-C.; Huang, C.-I.; Tsai, Y.-S.; Tsai, P.-C.; Ko, Y.-M.; Wang, W.-H.; Lin, C.-C.; Chen, K.-Y.; et al. Decision Tree Algorithm Predicts Hepatocellular Carcinoma among Chronic Hepatitis C Patients Following Viral Eradication. *Am. J. Cancer Res.* **2023**, *13*, 190–203.
65. Machuca, C.; Vettore, M.V.; Krasuska, M.; Baker, S.R.; Robinson, P.G. Using Classification and Regression Tree Modelling to Investigate Response Shift Patterns in Dentine Hypersensitivity. *BMC Med. Res. Methodol.* **2017**, *17*, 120. [CrossRef]
66. Althnian, A.; AlSaeed, D.; Al-Baity, H.; Samha, A.; Dris, A.B.; Alzakari, N.; Abou Elwafa, A.; Kurdi, H. Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain. *Appl. Sci.* **2021**, *11*, 796. [CrossRef]
67. Van der Ploeg, T.; Austin, P.C.; Steyerberg, E.W. Modern Modelling Techniques Are Data Hungry: A Simulation Study for Predicting Dichotomous Endpoints. *BMC Med. Res. Methodol.* **2014**, *14*, 137. [CrossRef]
68. Reininger, B.M.; Lopez, J.; Zolezzi, M.; Lee, M.; Mitchell-Bennett, L.A.; Xu, T.; Park, S.K.; Saldana, M.V.; Perez, L.; Payne, L.Y.; et al. Participant Engagement in a Community Health Worker-Delivered Intervention and Type 2 Diabetes Clinical Outcomes: A Quasiexperimental Study in MexicanAmericans. *BMJ Open* **2022**, *12*, e063521. [CrossRef] [PubMed]
69. Gill, A.; Gothard, M.D.; Briggs Early, K. Glycemic Outcomes among Rural Patients in the Type 1 Diabetes T1D Exchange Registry, January 2016–March 2018: A Cross-Sectional Cohort Study. *BMJ Open Diabetes Res Care* **2022**, *10*, e002564. [CrossRef]
70. Svoren, B.M.; Volkening, L.K.; Butler, D.A.; Moreland, E.C.; Anderson, B.J.; Laffel, L.M.B. Temporal Trends in the Treatment of Pediatric Type 1 Diabetes and Impact on Acute Outcomes. *J. Pediatr.* **2007**, *150*, 279–285. [CrossRef] [PubMed]
71. Holl, R.; Swift, P.; Mortensen, H.; Lynggaard, H.; Hougaard, P.; Aanstoot, H.-J.; Chiarelli, F.; Daneman, D.; Danne, T.; Dorchy, H.; et al. Insulin Injection Regimens and Metabolic Control in an International Survey of Adolescents with Type 1 Diabetes over 3 Years: Results from the Hvidore Study Group. *Eur. J. Pediatr.* **2003**, *162*, 22–29. [CrossRef] [PubMed]