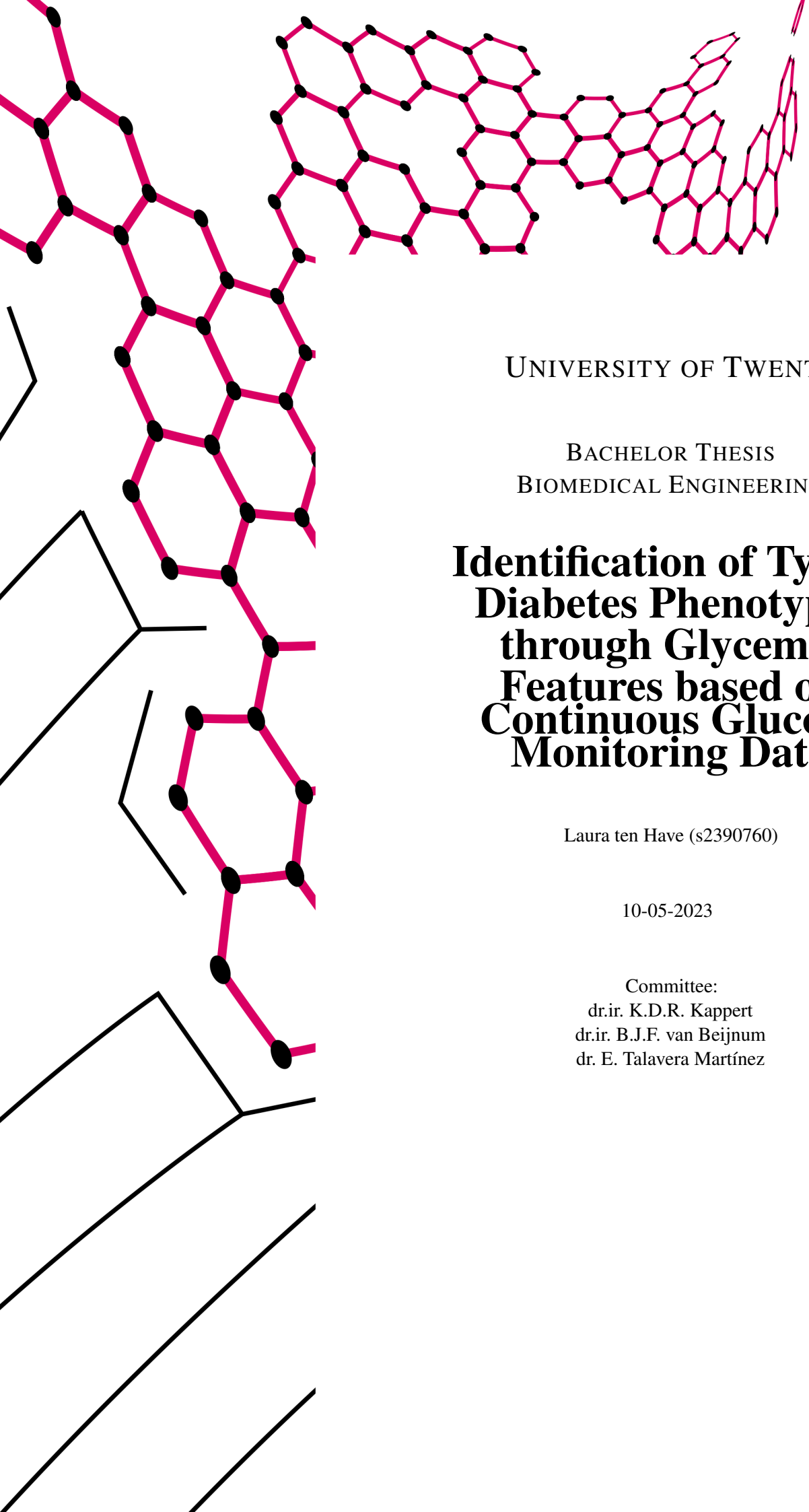


UNIVERSITY OF TWENTE

BACHELOR THESIS
BIOMEDICAL ENGINEERING

Identification of Type 1 Diabetes Phenotypes through Glycemic Features based on Continuous Glucose Monitoring Data

Laura ten Have (s2390760)

10-05-2023

Committee:
dr.ir. K.D.R. Kappert
dr.ir. B.J.F. van Beijnum
dr. E. Talavera Martínez

Abstract

Objective:
No personalized therapy options are yet available for T1D patients. It is believed that CGM data could possibly be the source for distinguishing different phenotypes which could eventually aid developing health care. However no studies were yet able to identify and validate distinct clinically relevant CGM-based phenotypes. Therefore, the aim of the current study is to distinguish different T1D phenotypes based on cluster analyses to ultimately contribute to better T1D health care.

Methods:
The study sample included patients from the DIABASE cohort who were clustered based on 20 selected CGM features measuring hypoglycemia, hyperglycemia and glycemic variability during different times of the day. Agglomerative hierarchical clustering was used as the clustering method in combination with the 'agnes' function in R, with Euclidean distance as the distance metric and Ward's minimum variance method as the linkage method. The optimal number of clusters was determined based on analysis of the elbow plot and the dendrogram. Statistical analysis of the numeric features for overall-tests of difference was carried out using the Kruskal-wallis test. Inter-cluster testing was performed with the Dunn's test. An extra method of validation called the silhouette value was used to analyse the placement of patients in the clusters.

Results:
The study sample included 78 adult patients with T1D (53,56% male, mean age 53 years). Five clusters were identified with overall significant differences ($p<0,05$), inter-cluster differences were not all significant, especially between cluster 1 and 5. Cluster 1 had overall moderate feature values which stayed closed to the average values of all patients. The silhouette plot showed that not every patients was correctly placed in this cluster. Cluster 2 was characterized by the lowest hypoglycemia metrics and severe hyperglycemic incidence. Cluster 3 showed severe hyperglycemic exposure compared to other clusters and had the highest mean glucose value. Cluster 4 was characterized with the lowest hyperglycemia metrics, lowest glycemic variability and the lowest mean glucose value. Cluster 5 showed severe hypoglycemia and glycemic variability and the silhouette width indicated that patients were placed incorrect.

Conclusion:
The current study showed that CGM data analysis can be used to discover different subgroups of T1D, which could eventually result in more personalized health care. Five clusters were identified by using agglomerative hierarchical clustering analysis based on glycemic features. However, to validate if distinct phenotypes of T1D were found, additional research is needed.

Contents

Introduction

Worldwide roughly 537 million adults (20-79 years) are living with Diabetes Mellitus (DM), which is 1 in 10 adults [1]. This number is expected to rise to 643 million in 2030 and to 783 million by 2045 [1]. Every five seconds someone dies due to DM, making it one of the largest health challenges worldwide. Of the total number of adults living with DM, 5-10% are affected by Type 1 Diabetes (T1D). T1D is an auto-immune disease in which the production of insulin is heavily reduced or there is no insulin production at all [2]. Insulin is a hormone secreted by the beta-cells of the pancreas, which ensures that blood glucose is taken up by cells. Without insulin, glucose will remain in the blood leading to high blood glucose values, also known as hyperglycemia [2]. This can lead to severe long-term complications such as kidney, eye and foot problems or nerve damage and heart failure [3]. Maintaining blood glucose levels within a healthy range, will lower the chance at these long-term complications. Treatment of T1D includes administration of insulin, either by pump or by insulin pens. An overdose of insulin leads to low blood glucose values, also known as hypoglycemia, which could lead to severe consequences such as seizures, coma or even death [3].

The diagnosis and prognosis of long-term diabetes is estimated by measuring glycated hemoglobin A1c (HbA1c) levels, which is considered as the gold standard. HbA1c is measured in the hospital through a blood test and indicates the mean glucose values over the past 60-90 days. However, this method lacks information about direct glycemic excursions and fails to identify the magnitude of glucose variations [4]. In addition, other medical conditions, iron deficiency, pregnancy and chronic kidney disease reduce the reliability of the test. Nowadays it is the only evaluated tool for assessing the risk for diabetes complications but this is expected to change soon as new methods are being developed.

Continuous Glucose Monitoring (CGM) is a new method for continuously following glucose levels in the interstitial fluid by placing a sensor on the skin or by implanting the sensor below the skin. CGM prevents the use of a fingerstick, consequently reducing the burden for patients. CGM can replace or enhance the use of HbA1c by informing on direct observations of glycemic excursions which can then lead to immediate therapy decisions or lifestyle modifications [4]. It can also assess glucose variability during different times of the day and identify hypo-and hyperglycemia patterns. Nevertheless only a small segment of the population is able to use CGM devices. This is due to lack of insurance coverage (there is only insurance coverage for patients with T1D), discomfort of wearing a device and perceived sensor inaccuracy [5]. This is expected to improve in the coming years, which stimulates the development of CGM data analysis.

A combination of CGM metrics from the Advanced Technologies and Treatments for Diabetes (ATTD) congress consensus statement is often used for the analysis of CGM data. A panel consisting of physicians, researchers and experts in CGM wanted to provide guidance for utilizing, interpreting, and reporting CGM [6]. As a result, fourteen core metrics for the analysis of CGM data were established, including glycemic variability, mean glucose value, Time in Range (TIR), Time Above Range (TAR), Time Below Range (TBR), number of hyper-hypoglycemic episodes lasting longer than 15 minutes and number of days CGM worn. Kahkosha *et al.* [7] used a combination of eight of these CGM features to identify new phenotypes in T1D.

Previous studies have tried to identify different relevant subgroups sharing the same patterns, called phenotypes [7, 8] to enhance glucose control. Phenotypes are subtypes of diabetes based on phenotypic criteria, such as insulin resistance, body mass index (BMI), age at diabetes diagnosis and HbA1c [9]. Deutsch *et al.* [10] identified five phenotypes of diabetes based on phenotypic similarity, and Oana and colleagues [11] validated the reproducibility of these five phenotypes by using different clustering methods and variables. However, these studies did not use CGM analysis while establishing the phenotypes, which is in contrast to more recent studies that tried to identify reproducible phenotypes based on CGM data [12–14]. For example Szczesniak *et al.* [14] discovered three phenotypes based on distinct longitudinal patterns of glucose, insulin and blood pressure control in pregnant women. Although an increase in research is conducted into this topic, there remains little data on treatment response and specific clinical recommendations.

Cluster analyses has already been used to reveal hidden structures in CGM data and it has been proved that it is a viable method to identify groups of diabetes patients with different features [15]. Studies showed that CGM analyses can improve glucose control and showed that CGM metrics refine the estimation of glucose control given by HbA1c [7, 16]. Hall *et al.* [17] found different fluctuation modes in individuals without diabetes, with diabetes and with pre-diabetes patients. Tao *et al.* [8] clustered based on several standardized metrics including coefficient of variation (CV), mean amplitude of glycemic excursion (MAGE) and mean sensor glucose (MSG) to divide T2D patients

into subgroups. Pollè *et al.* [18] clustered T1D patients who were diagnosed in the last 21 days based on global control measures, hypoglycemia, time in range, hyperglycemia, glucose global, within- and between-day variability, and combination scores. They researched if CGM metrics strongly relate with features of diabetes control related to Beta-cell function. This unravelled clinical milestones of remission status during the first year of T1D. Kietsiriroje *et al.* [19] clustered based on low, intermediate and high glucose variability (GV) to show higher GV is associated with increased thrombotic biomarkers in T1D patients. Lobo *et al.* [20] created an algorithm which made a finite set of representative daily CGM profiles. This was done to match each individual daily CGM profile to a representative daily CGM profile. Mao *et al.* [21] developed a classification algorithm to categorize diabetes patients based on three broad types of features, measures of centrality and spread (max, min, mean, SD, MAGE, TIR, TBR, TAR) measures of glucose excursion (average rise and fall, slopes) and fluctuations in glucose (average CGM deviation from mean).

As previously described, extensive research has been conducted regarding methods on analysing CGM data however, the gold standard to analyse CGM data is not yet known. This will be critical for tailored individual approaches to monitor glucose levels in the diabetes health sector. Not enough validation studies based on practical clinical applications are executed, Thus, additional research is needed to find the best analysis strategy. As mentioned before some research has already been conducted to establish what the best method is to analyse CGM data. The current study is mainly based on two previous studies [7, 18] but differs on various aspects. Kahkoska *et al.* [7] used Self-Organizing Map (SOM) as a clustering algorithm while in this study hierarchical clustering method is used similar to Pollè *et al.* [18]. The optimal recommended data analysis period of 14 days is used [6] compared to the 7 day time period used by Kahkoska *et al.* [7]. Previous research considered daily metrics or night/day metrics but this study will contain CGM metrics which are stratified by 4 periods; night (00:00-6:00) morning (6:00-12:00), afternoon (12:00-18:00) and evening (18:00-00:00).

No personalized therapy options are available for T1D patients while this is desirable for optimal health care. Patients differ in glucose regulation, and therefore different phenotypes could potentially be identified. These can aid in developing personalized therapy options for T1D patients. It is believed that CGM data could possibly be the source for distinguishing different phenotypes. Therefore, the aim of the current study is to distinguish different T1D phenotypes based on cluster analyses. First, the current study aims to identify relevant features for cluster analysis. Second, the aim is to explore methods for the identification of different clusters. Third, clusters will be compared to known clinical parameters of the patients to see how they differ. Finally, the identified clusters will be tested to see if they differ significantly from each other.

Methods

Study Population

The current study uses baseline data from the DIABASE cohort, obtained from T1D patients located at Ziekenhuisgroep Twente (ZGT) hospital in Hengelo and Almelo, in the Netherlands. CGM data was collected from different glucose measurement devices consisting of various versions of the FreeStyle Libre and the Dexcom device. The insulin values were obtained from the Carelink Medtronic or the Diasend insulinpump. To be included in the study the patients had to be 18 years or older, had to be diagnosed with T1D, could not be severely ill and could not have mental disorders. Patients are still included today, expectedly reaching 1000 patients by the end of 2023. The measurements that are used in the current study started on the eleventh of June 2016 until the sixteenth of June 2022. The current existing database consists of 95 patients in which 78 were eligible for this study.

Method Overview

To give an overview of the various steps taken to obtain the final clusters, a flowchart is provided in Figure 1. In the start of the study, raw CGM data was obtained from the patients CGM devices. CGM- based exclusion criteria were applied and the data was pre-processed by formatting the data correctly and filling in missing data. Feature extraction consisted of obtaining the selected 28 features from the CGM data. Before the clustering analysis could begin, the extracted features had to be scaled, the correlation matrix had to be computed and the highly correlating features had to be removed. When the feature selection was concluded, the exact method to execute clustering analysis had to be determined. The agglomerative coefficient showed that Ward's minimum variance method was the best linkage method. The Euclidean distance metric was used to compute the distance matrix and literature study was performed to determine that the 'agnes' function would be the most suitable. A combination of these three parameters resulted in being able to execute the agglomerative hierarchical clustering method. The last thing left to do was establishing the optimal number of clusters, which was done by realizing the elbow plot and the dendrogram. Combining the agglomeritive hierarchical clustering method and the knowledge of the number of clusters that needed to be computed, led to the formation of the final clusters. Each step depicted by Figure 1 will be explained in more detail in this section.

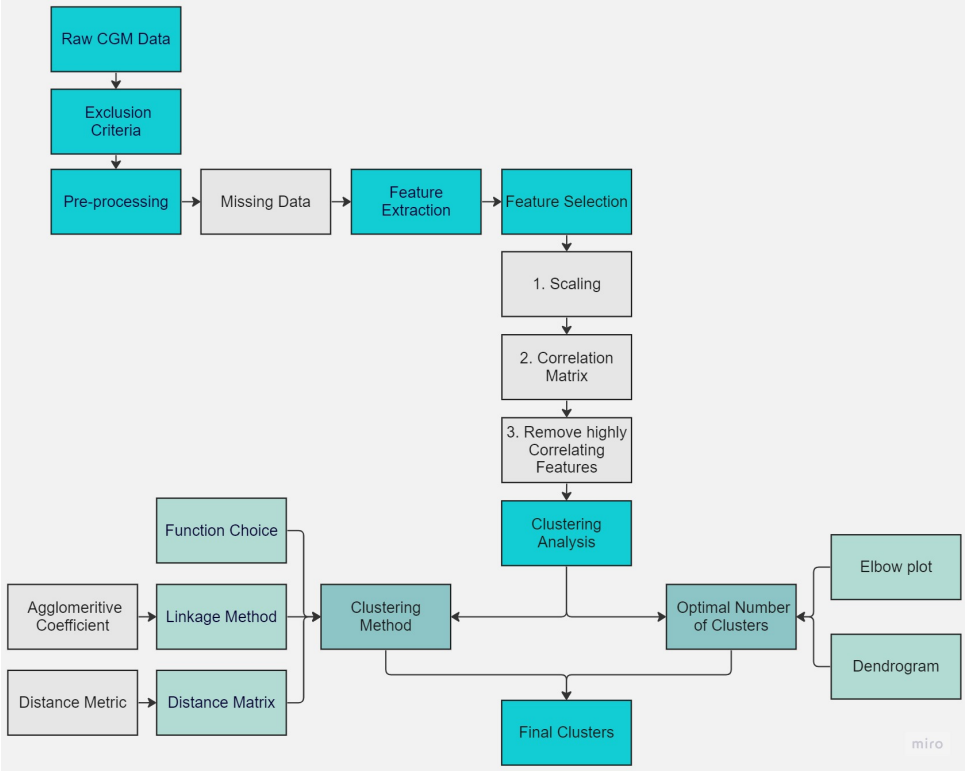


Figure 1: Flowchart which shows the different steps taken to obtain the final clusters. The main steps are indicated with the blue boxes. The grey boxes below feature selection define the steps taken in feature selection and the grey box next to Pre-processing shows an important step of pre-processing.

Exclusion Criteria

To be suitable for this specific study, several CGM-based exclusion criteria were applied. First, every patient should have at least 120 days of CGM data. The gold standard to measure CGM data is 90 days, signifying the minimum number of days required for CGM analysis. An additional 30 days were considered to avoid patients who are not yet familiarised with the used devices. After application of this exclusion criteria, ten patients were excluded, resulting in 85 remaining patients. Second, patients should have at least 80% of data per day for 14 consecutive days. Considering data is obtained every 15 minutes, a minimum of 76 data points per patient per day were needed. Two glucose measurement devices were used, the Dexcom and Freestyle Libre, both measuring at different intervals. To prevent inconsistency Dexcom patients were excluded in this research, coincidentally these patients were already removed by the first exclusion criteria therefore, no additional patients were excluded. After applying the exclusion criteria, 78 patients were left to be included in this study.

Pre-processing

All CGM data was analysed using R (Rstudio version 4.1.1) which is a language and environment for statistical computing [18]. The raw CGM data was obtained in the format shown in Table A.1. This data was formatted correctly and the year with the most available data points was chosen, which was 2021. To be able to compute the metrics for different time blocks an extra column was added to show the time block corresponding to the time at that moment. (Table A.2).

Missing Data

Due to incomplete data, missing data was filled. The data was rounded to the nearest 15 minutes to get consecutive 15-minute intervals for the entire day. Thereafter, the data was padded with the missing data value taking on the previously known glucose value. The number of maximum consecutive missing data points were analysed and found by extracting all not available (NA) values, then finding the sequences of consecutive NA values and calculating the maximum length of the found sequences. The length of maximum consecutive missing data points is kept to a minimum, by checking if it is within reasonable range. The reasonable range was defined at the exclusion criteria based on the minimum number of available data points needed per day of 76. This makes that the acceptable range is from 0-20 consecutive missing data points per day. The maximum length of consecutive data points was taken to show reliability of the different clusters. If more data points had to be estimated, the results would increasingly become less reliable. Additionally, the percentage of missing data points per cluster will be given to give supplementary information about the distribution of missing data between clusters.

Feature Extraction

Metrics were selected based on the key metrics stated by the ATTD congress consensus [6]. The measurable key metrics are; mean glucose value, percentage of time in hypoglycemic ranges (level 1 < 3,0 mmol/L, level 2 < 3,9 mmol/L), percentage of time in target range (3,9-10 mmol/L), percentage of time in hyperglycemic ranges (level 1 > 10 mmol/L, level 2 > 13,9 mmol/L), primary glycemic variability measured by the Coefficient of Variation (CV), secondary glycemic variability measured by the Standard Deviation (SD) , episodes longer than or equal to 15 minutes of hypoglycemia and hyperglycemia and area under the curve (AUC). AUC, level 2 metrics and TIR were not considered due to previously observed high correlation numbers with other key metrics [7]. This resulted in 28 features to measure key metrics shown in Table 1. Other recommendations made by the ATDD consensus statement were also considered including, taking different time blocks, having a minimum of 80% available data per day, minimum length of 15 minutes of the hypoglycemia/hyperglycemia episodes and a minimum CGM data collection period of 14 days.

Table 1: The 28 initial features for hierarchical clustering used to capture hypoglycemia, hyperglycemia and the glycemic variability in the morning, afternoon, evening and night.

Metrics	Features
Mean Blood Glucose Value	Mean Glucose Value, Night[1]
	Mean Glucose Value, Morning[2]
	Mean Glucose Value, Afternoon[3]
	Mean Glucose Value, Evening[4]
Primary Glycemic Variability	Coefficient of Variation, Night
	Coefficient of Variation, Morning
	Coefficient of Variation, Afternoon
	Coefficient of Variation, Evening
Secondary Glycemic Variability	Standard Deviation, Night
	Standard Deviation, Morning
	Standard Deviation, Afternoon
	Standard Deviation, Evening
Hyperglycemia Incidence (periods >= 15 minutes)	Number of Episodes BG[5]>10 mmol/L, Night
	Number of Episodes BG >10 mmol/L, Morning
	Number of Episodes BG >10 mmol/L, Afternoon
	Number of Episodes BG >10 mmol/L, Evening
Hypoglycemia Incidence (periods >= 15 minutes)	Number of Episodes BG <3,9 mmol/L, Night
	Number of Episodes BG <3,9 mmol/L, morning
	Number of Episodes BG <3,9 mmol/L, Afternoon
	Number of Episodes BG <3,9 mmol/L, Evening
Hyperglycemia Exposure (% Time)	Percentage time >10 mmol/L, Night
	Percentage time >10 mmol/L, Morning
	Percentage time >10 mmol/L, Afternoon
	Percentage time >10 mmol/L, Evening
Hypoglycemia Exposure (% Time)	Percentage time <3,9 mmol/L, Night
	Percentage time <3,9 mmol/L, Morning
	Percentage time <3,9 mmol/L, Afternoon
	Percentage time <3,9 mmol/L, Evening

[1] Night defined as 00:00-6:00; [2] Morning defined as 6:00-12:00; [3] Afternoon defined as 12:00-18:00;
[4] Evening defined as 18:00-00:00; [5] BG, Blood Glucose;

Measuring GV could be of help to optimize glucose control by maintaining optimal average glycemia without increasing the risk for hypoglycemia. primary GV was defined as the percentage CV in which 36% is the threshold between stable and unstable glycemia [22]. CV is defined as the standard deviation divided by the mean were the mean and standard deviation are also taken as separate features. Secondary GV was measured by the SD, which is a measure of the spread in glucose readings around the average also known as the variation, calculated by equation 1 [23]. In which x_i is an individual observation, $\bar{x}$ is the mean of observations and k the number of observations. The percentage TAR of level 1 was measured above a value of 10 mmol/L. Percentage TBR of level 1 was measured beneath values of 3,9 mmol/L. The number of hypo-and hpyerglycaemic periods are counted starting from 15 minutes or more.

$$SD = \sqrt{\frac{\sum(x_i - \bar{x})^2}{k-1}} \tag{1}$$

Feature Selection

Features were extracted and calculated from the CGM data for each time block (night, morning, afternoon and evening). To reduce the number of features and avoid repetitive data, feature selection was necessary and therefore executed. Feature selection is to find the best set of features that allows for building optimized models.

Scaling

All features were left continuous and were later normalized to be expressed on the same scale and could be used for the analysis (Table A.4). The 'scale' function in R was used which converts data into z-scores which involves mean centering a variable and then scaling it by the standard deviation (equation 2) [24]. In which x is the real x value, $\bar{x}$ is the mean and SD is the standard deviation. For a subtract of the results obtained after feature calculations, Table A.3.

$$x_{scaled} = \frac{(x - \bar{x})}{SD} \tag{2}$$

Correlation Matrix & Removing highly Correlated Features

A correlation matrix, that shows the correlations between different features, was used to find the highly correlated features (figure 2). One of the pair of correlating features with a correlation value greater than 0,8 was removed to prevent repetitious data. The used function, -"findCorrelation"- determined the mean absolute correlation of each feature and removed the feature with the greatest mean absolute correlation. Mean glucose values in the night, morning, afternoon and evening corresponded with the percentage TAR in the night, morning, afternoon and evening with correlation values of 0,985, 0,975, 0,965, 0,966 respectively. Hypoglycemic exposure of 15 minutes of longer in the night, morning, afternoon and evening corresponded with percentage TBR of the night, morning, afternoon and evening with correlation values of 0,821, 0,838, 0,914, 0,903 respectively. The excluded features are mean glucose of the morning, afternoon and evening, the TAR of the night, hypoglycemic exposure of the night, morning and afternoon and the TBR of the evening, resulting in a remainder of 20 features applicable for clustering analysis.

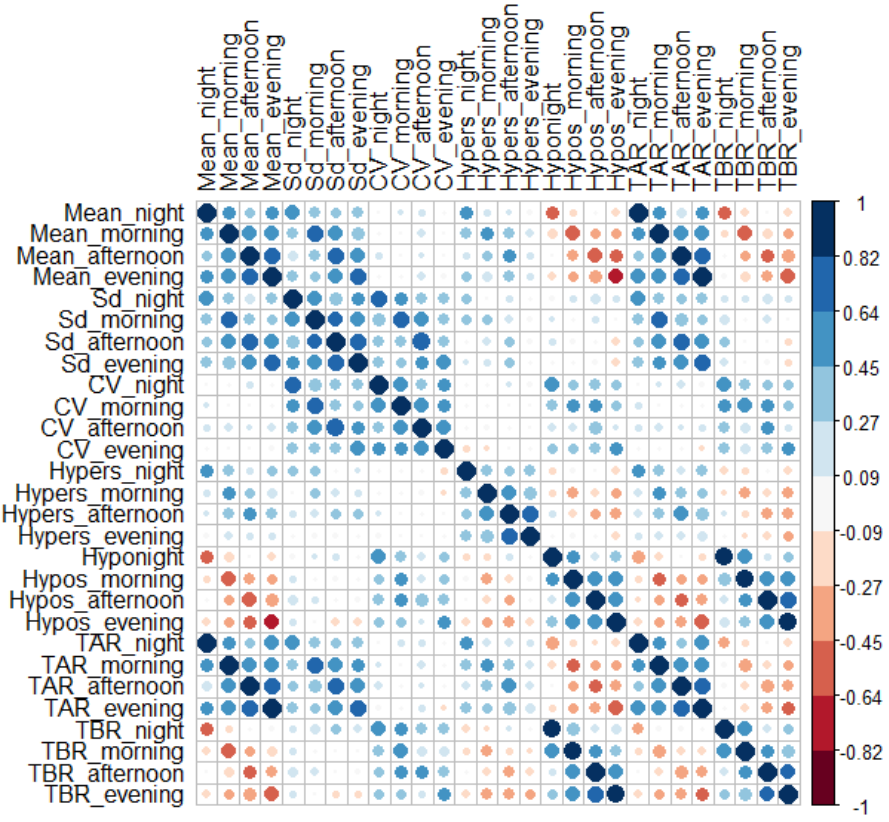


Figure 2: Correlation matrix of the 28 selected features. The colours and size of the dots indicate the correlation value for which the index is shown on the right side. All dots with a dark blue colour were removed. This clearly shows which features corresponded with each other.

Clustering Analysis

The clustering analysis involved placing patients into subgroups based on their relevant features, which were determined previously in feature extraction and feature selection. There are various ways of executing the clustering analysis. The clustering method, the linkage method, a suitable distance metric and the optimal number of clusters had to be selected. A detailed description of each decision will be given below.

Clustering Method

Hierarchical clustering was compared with K-means clustering and a selection of the differences are shown in Table 2. An important factor defining K-means is that prior knowledge is needed about the number of clusters [25, 26]. In this study the number of clusters is not specified beforehand, which makes K-means less suitable. Alongside the fact that a pre-defined number of clusters are needed, having the possibility of variations in the results when re-running the algorithm for K-means is sub-optimal [27]. Variations in results are due to random centroid initialization or due to changing the order of data. K-means generates non-overlapping spherical clusters, however in this study any geometrical shape could be the outcome of the clusters.

K-means defines clusters by randomly selecting a group of centroids, which are the initial centres of clusters [27]. Then the position is optimized by repeating two steps: 1. assigning data points to the nearest cluster centroid and 2. updating the coordinates of the centroid [28]. The sphere is eventually drawn around the outer points of the cluster. Shi *et al.* [29] emphasizes that the repetitive process of calculating the distance between each data object and the cluster centers, affects the efficiency of the clustering algorithm. In combination with the relatively small data set and the aforementioned factors, the apparent decisions was to utilize hierarchical clustering as the clustering method.

Table 2: Comparison of K-means clustering and hierarchical clustering [25–29]

K-means Clustering	**Hierarchical Clustering**
• Prior knowledge required about number of clusters	• Any number of clusters, often based on dendrogram
• Yields varying results with re-running algorithm	• The results are reproducible (more robust)
• Suitable for very large data sets due to low intensity	• More intensive, suitable for smaller data sets
• Only generates non-overlapping spherical clusters	• Any geometrical shape can be the outcome
• Possibility of creating empty clusters	• All clusters contain at least one data point

Hierarchical clustering is a procedure that determines successive clusters based on previously defined clusters. It groups data into a tree of clusters and is often shown in a dendrogram. It results in clusters which are different from the other clusters but the patients in the clusters are similar. There are two types of hierarchical clustering, agglomerative and divisive. Agglomerative hierarchical clustering uses an approach were each patient starts in its own cluster and is then iteratively linked to nearby patients until there is one cluster left [30]. Divisive clustering is the reverse approach of the agglomeritive clustering, were all patients start in one cluster and are then step by step split up into smaller clusters. Divisive methods are rarely used, and in that case they are mainly used for identifying large clusters. In combination with agglomerative clustering being able to handle outliers better than divisive clustering, the agglomerative method will be used in the remainder of this study.

Function Choice

There are two types of functions that can be used in R for agglomerative hierarchical clustering, 'hclust' and 'agnes'. These two functions are roughly the same, however agnes uses fewer shortcuts when updating the distance matrix. Hclust uses two recently merged observations when updating the distance matrix while agnes uses all the observations when updating the matrix [31]. The differences between the two dendrograms can be seen in Figure A.1. The number of patients per cluster is the same for both methods when using different number of clusters. Agnes is the chosen function because it has the advantage of being able to calculate the agglomeritive coefficient.

Linkage Method

The linkage method calculates the similarities or distances between all clusters. The two clusters that are nearest are

combined, reducing the total number of clusters [32]. Various linkage methods are used in agglomerative hierarchical clustering, which are discussed below.

1. Single linkage: two clusters with the closest minimum distance are merged. The formula used to describe this method is stated in equation 3 in which d_{mj} is the distance between clusters m and j, d_{kj} the distance between cluster k and j and j_{lj} is the distance between clusters l and j [33].

$$d_{mj} = min(d_{kj}, d_{lj}) \tag{3}$$

2. Complete linkage: two clusters with the closest maximum distance are merged [33, 34]. The formula used to describe this, is similar to the simple linkage formula except the maximum distance is used (equation 4).

$$d_{mj} = max(d_{kj}, d_{lj}) \tag{4}$$

3. Ward's linkage: joins the two clusters A and B in a way that the increase in the sum of squared erros (SSE) is minimized (equation 5) [32].

$$I_{AB} = SSE_{AB} - (SSE_A + SSE_B) \tag{5}$$

The SSE between and within the clusters are defined as shown in equation 6 [32]. Where a_i, b_i, y_i represent the i^{th} observation vector in cluster A, B and AB respectively. $\overline{a}$, $\overline{b}$, $\overline{y}_{AB}$ represent the centroid of cluster A,B and of the new cluster AB respectively [32].

$$\begin{aligned} SSE_A &= \sum_{i=1}^{n_A} (a_i - \overline{a})'(a_i - \overline{a}) \\ SSE_B &= \sum_{i=1}^{n_B} (b_i - \overline{b})'(b_i - \overline{b}) \\ SSE_{AB} &= \sum_{i=1}^{n_{AB}} (y_i - \overline{y}_{AB})'(y_i - \overline{y}_{AB}) \end{aligned} \tag{6}$$

Ward's method, also called Ward's minimum variance method calculates the distance between patients and the centroid of that cluster, in which the centroid is the point where the sum of squared distances between the point and every other point in the cluster is minimal [32].

4. Average linkage: uses the average pair-wise proximity among pairs of all objects in different clusters, in which they are merged on the lowest average distance [33]. For the average linkage method the formula is denoted by equation 7 with the previously described parameters

$$d_{mj} = (N_k d_{kj} + N_l d_{lj})/N_m \tag{7}$$

Agglomerative Coefficient

The most accurate linkage method was selected based on the agglomerative coefficient. This coefficient measures the dissimilarities of sets as a function of pairwise distances of patients in the sets [35], the closer the value is to one, the more accurate the method is. The function describing this parameter is defined as $1 - m(i)$ in which i is the observation, $m(i)$ is the dissimilarity of that observation with respect to the first cluster it merged with, divided by the dissimilarity of the fuser in the last step of the algorithm [36]. Executing this for all patients results in the agglomeritive coefficient. The "ac" function in R was used to establish the agglomerative coefficient for the different methods. Table 3 shows that Ward's minimum variance method is the most accurate for the used data set. To give an indication of the influence of the linkage method, the dendrogram of each linkage method, computed with the agnes function, is shown in Figure A.2.

Table 3: agglomerative Coefficients for each linkage method, average, single, complete and ward's minimum variance.

Method	Agglomerative coefficient
Average linkage	0,649
Single linkage	0,539
Complete linkage	0,768
Ward's minimum variance	0,849

Distance Matrix
The distance matrix was computed by using the "dist" function in R, which required the distance metric. The distance metric is a cluster similarity measure which gives the degree of closeness or separation of features of different data points [37]. The most frequently reoccurring method of identifying similar and dissimilar measures are the Euclidean and Manhattan distance.

Euclidean distance is the most widely used distance metric [37], calculating the length of the shortest path between two points. Equation 8 shows the formula used for calculating the Euclidean distance [38], with d as a vector of the dimensions of the data used and x, y as the considered vectors. Euclidean distance is proved to be a suitable combination with hierarchical clustering, however it is only valid for continuous variables.

$$d(x,y) = (\sum_{j=1}^{d} (x_j - y_j)^2)^{1/2} \tag{8}$$

The Manhattan distance calculates the distance between two points by adding the horizontal and vertical component (see equation 9) . Manhattan distance travels along coordinates only, resulting in overall higher distances between points than the Euclidean distance. Manhattan distance is preferred over Euclidean distance when there is high dimensionality in the data [39], more features than observations, which does not apply for this study. Weighing these factors made that the more common Euclidean distance metric was applied.

$$d(x,y) = \sum_{j=1}^{d} |x_j - y_j| \tag{9}$$

Optimal Number of Clusters

The optimal number of clusters was found by using a combination of the dendrogram and the elbow plot.

Dendrogram
To find the optimal number of clusters with a dendrogram the largest vertical distance between nodes was found. In the middle of this vertical line a horizontal line can be drawn, the number of vertical lines it crosses is the optimal number of clusters [40]. The dendrogram shows the height of fusion on the vertical axis, thereby showing the (dis)similarity between two patients. The higher the height of the fusion, the less similar the observations are [40].

Elbow plot.
The elbow plot is defined by calculating the total within sum of squares per cluster in a range of number of clusters. The location where the bend (i.e. knee) is located is considered the optimal number. It is not always a clear elbow, which makes that the elbow plot was used in combination with the dendrogram.

Statistical Analysis

When the clusters were computed, the resulting clusters were validated in R. Statistics was used to show overall significance per feature and significance between specific clusters. The statistical significance level used for all analyses was $p < 0.05$. Overall-tests of difference were carried out using the Kruskal-Wallis test. Inter-cluster-tests of difference were performed with the Dunn's test. These methods of showing significance were used for the continuous numeric features.

Seven clinical parameters were compared between groups, these included HBA1c, Creatine, low-density-lipoprotein (LDL), high-density-lipoprotein (HDL), total cholesterol (TC), blood pressure and the Albumin/Creatine-ratio. A high LDL, also known as 'bad' cholesterol, is associated with a higher risk of cardiovascular disease. In contrast a high HDL (i.e.,'good' cholesterol) is associated with a lower risk of cardiovascular disease (CVD). Smoking and being overweight are associated with lower HDL levels and consequently increase risk for heart disease and stroke. Diabetes affects cholesterol levels by increasing 'bad' cholesterol levels and decreasing 'good' cholesterol levels. Target LDL and HDL cholesterol levels for adults with diabetes are < 2.6 mmol/L and > 1.02 mmol/L respectively. The TC target range is to stay below 4.0 mmol/L. The blood pressure targets for T1D patients are below 135/85 mmHg. For HbA1c, the target is

to stay below 48 mmol/mol or 6.5%. For albumin/creatine-ratio the goal is to stay below 2.5 mg/mmol (men) and 3.5 mg/mmol (women).

As a validation method, the silhouette width was computed to note if patients were correctly placed within the right cluster. The silhouette coefficient calculation result can vary from -1 to 1 [41]. The closer the value is to one the more it indicates that the patient is placed properly. If the score is zero, it indicates that it is undecided to which cluster the patient belongs and is probably in between two clusters. If the value drops below zero, the patient is placed incorrectly [42]. For each patient the silhouette width is defined as depicted by equation 10 [41] in which $a(i)$ is the average dissimilarity between i and all other points of the cluster to which i belongs. $b(i) = min_c d(i,C)$ which is the dissimilarity between i and the nearest cluster to which it does not belong. C is defined as all other clusters and $d(i,C)$ is the average dissimilarity of i to all patients of C [41].

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{10}$$

In the feature comparison between clusters, colours were used to indicate if the values were the lowest, low, medium, high or the highest. Values were considered low when they were below 85% of the average value of all patients and values were considered high if they were above 115% of the average value. For example if the average value is 4 then values < 3,4 would be considered low and values > 4,6 high. These colours were given to give a (visually) clear comparison between the clusters.

Results

The results of the executed clustering analysis are depicted in this section. First of all, the patient characteristics per cluster will be compared, following with a comparison between clusters based on the silhouette score and missing data values. Subsequently the features per cluster will be compared based on significance. Finally, the differences between time blocks will briefly be discussed.

Patient Characteristics

The current study included 78 T1D patients. To create similar conditions, only the data of 2021 was used, while the mean available data per patient was 1,94 years (707 days) with a minimum of 141 days and a maximum of 4,80 years (1753 days). Patients were mostly male (52,56%) and had a mean age of 53,14 years, the youngest was 21 years and the oldest 81 years. The patient characteristics are shown for all patients and per cluster in Table 4.

Table 4: Patient characteristics for all patients and per cluster 1-5(CL)

Characteristics	**All**	**CL1**	**CL2**	**CL3**	**CL4**	**CL5**
Amount of Patients	n= 78	n= 27	n= 22	n= 18	n= 6	n= 5
Distribution						
age (years)	53,15	55,48	54,64	46,22	70,83	37,6
Sex male (%)	52,56	55,55	36,36	66,67	33,30	80
BMI[1]weight status (%)						
under or normal weight	28,21	5,13	8,97	6,41	2,56	5,13
overweight	23,08	10,26	6,41	3,85	1,28	1,28
obese	14,10	5,13	1,28	3,13	2,57	0,00
unknown	34,62	14,10	11,54	7,70	1,28	0,00
BSA[2](m^2)	2,03	2,03	1,9	2,14	2,12	1,78
Blood Pressure (mmHg)	131/75	132/75	134/76	126/75	133/73	126/72
Baseline Diabetes Characteristics						
HbA1c[3](mmol/mol)	59,70	56,07	63,63	66,37	52,76	47,83
HbA1c (%)	7,60	7,30	8,00	8,20	7,69	6,50
LDL[4](mmol/L)	2,45	2,20	2,64	3,00	1,62	2,60
HDL[5](mmol/L)	1,65	1,68	1,76	1,45	1,70	1,68
Total Cholesterol (mmol/L)	3,96	3,90	3,83	4,80	2,90	3,80
Albumin/Creatin-ratio	3,51	3,29	1,31	0,74	1,02	6,23
Other						
Smoking (%)						
stopped	24,36	12,82	7,69	3,85	1,28	1,28
currently	12,82	5,13	2,56	2,56	0,00	2,56
never	51,28	16,67	15,38	12,82	5,13	1,28
unknown	11,54	2,56	2,56	3,85	1,28	1,28
Alcohol Intake (%)						
currently	51,28	19,23	15,38	11,54	3,85	1,28
no	23,08	7,69	7,69	5,13	1,28	1,28
unknown	25,64	7,69	5,13	6,41	2,56	3,85

[1] BMI, Body Mass Index; [2] BSA, Body Surface Area; [3] HbA1c, glycated Heamoglobin A1c;
[4] LDL, Low-Density-Lipoprotein; [5] HDL, High-Density-Lipoprotein

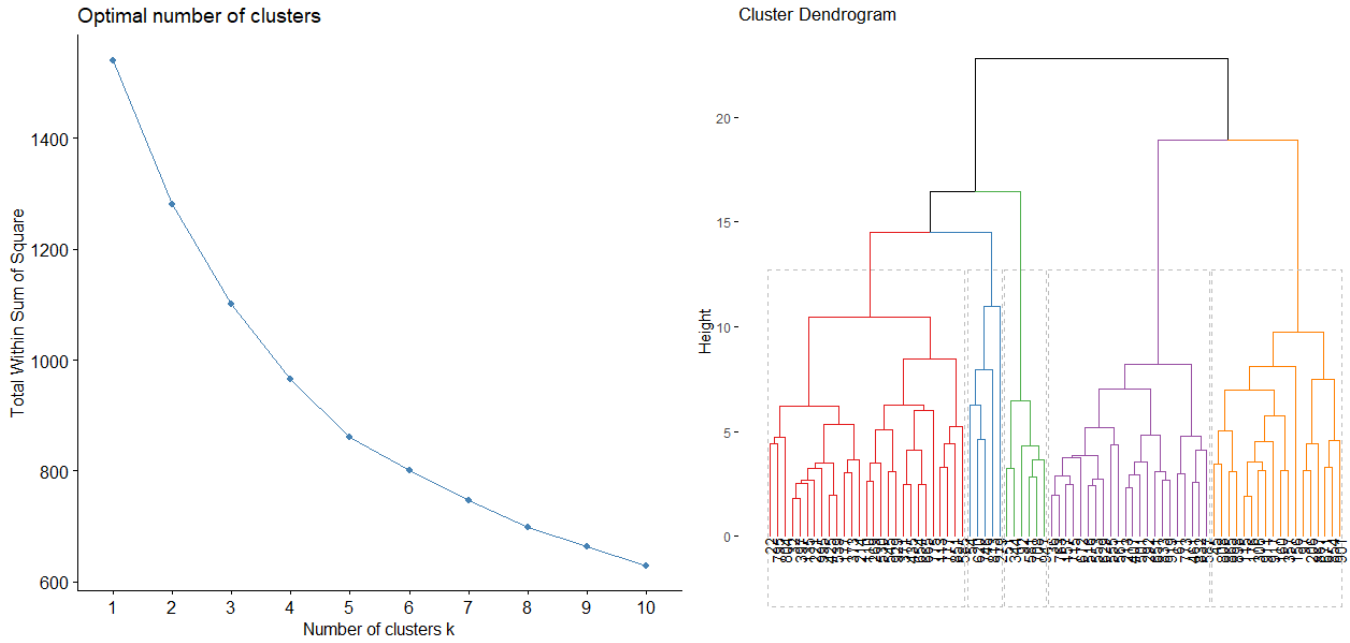
Table 4 shows that the blood pressure values per cluster are within the blood pressure target range of < 135/85 mmHg. The Body Surface Area (BSA) is relatively high for all clusters compared to the average of $1,75m^2$, but correlates with the amount of overweight and obese patients. Cluster 5 can be seen as an healthier example; this cluster shows the lowest

BSA and the highest percentage of normal weight patients. The albumin/creatin values are well within range for cluster 2 to 4, all staying below < 2,5 mg/mmol. Cluster 1 and 5 however severely exceed the set threshold.

Cluster 2, 3 and 5 all exceeded the target value of LDL (< 2,6 mmol/L) , which usually corresponds with increased risk for cardiovascular disease (CVD). Despite this all clusters had sufficient HDL values (> 1,02 mmol/L), which can, in contrast to LDL, Lower the risk for cardiovascular disease and stroke. Cluster 3 however had the highest LDL, lowest HDL and a TC which is exceeding the threshold (4,0 mmol/L), making group 3 the most prone to CVD. In contrast cluster 4 had the lowest LDL, highest HDL and lowest TC, making this group, overall the least prone to CVD. This is surprising considering this cluster has the highest mean age. All clusters display poor metabolic control (surpassing the set threshold), except for cluster 5 which is close the target value (< 46 mmol/mol, < 6,5%). Cluster 3 shows the highest HbA1c values, making this cluster the most likely to develop diabetes complications.

Final Clusters

Five clusters were identified based on the elbow plot, shown in Figure 3a, and the dendrogram, shown in Figure 3b. The bend in the elbow plot curve could involve a factor of interpretation, therefore the dendrogram was also inspected. The bend in the curve is in this case clearly visible at five clusters, which correlates with the outcome of the dendrogram. Cutting the dendrogram at the vertical line with the biggest height difference indicates once again an optimal number of clusters of five.



(a) **(b)**

Figure 3: a) Elbow plot showing the optimal number of clusters. The knee indicates the optimal number of clusters, indicating that five clusters is the optimal amount. b) Dendrogram of the different clusters with the patients numbers on the x-axis and the height on the y-axis. Cluster 1 is indicated with a red colour, cluster 5 is blue cluster 4 is green, cluster 2 is purple and in orange cluster 3. The horizontal line is drawn in the middle of the line with the amplest height difference, resulting in an optimal number of five clusters

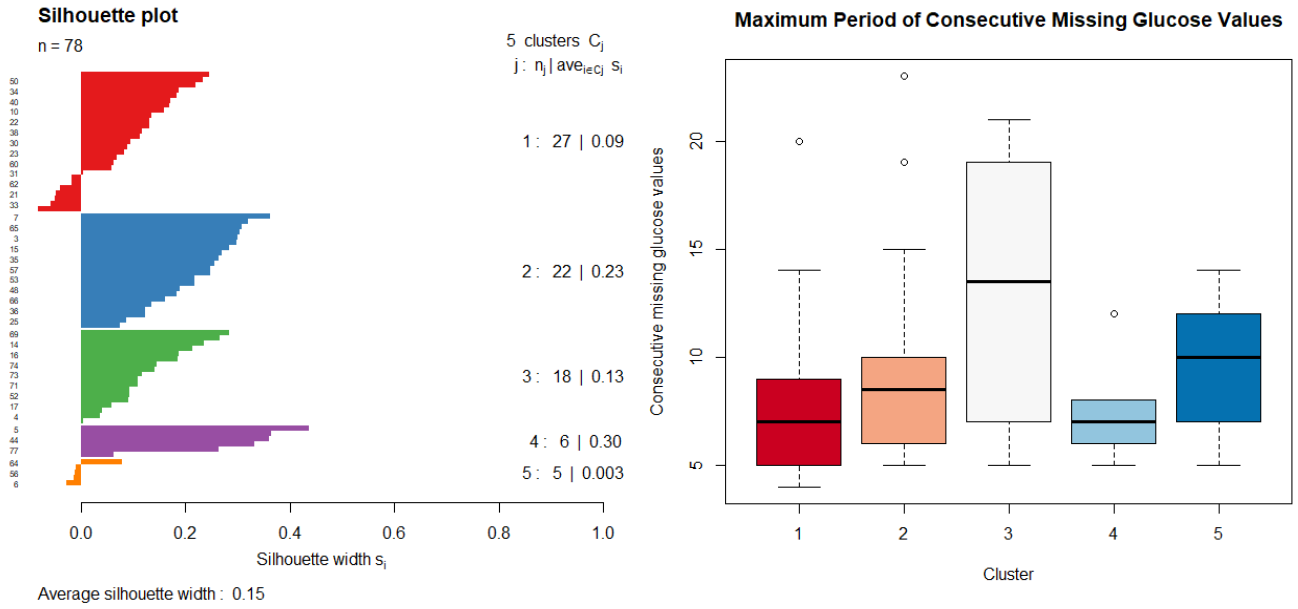
Figure 4a shows the computed silhouette values per patient and per cluster. Additionally the silhouette plot provides the average silhouette width, which is 0,15 for this study. The highest values correspond with cluster 2 and cluster 4, implying that the patients in these clusters are placed the best. The score for cluster 5 is almost zero indicating that cluster 5 patients do not necessarily belong in this cluster, but are in between two clusters. Cluster 1 also has a relatively low score with several patients who have negative values indicating that they are placed incorrectly. Cluster 3 has a silhouette width close to the average width and has no patients with negative values.

Figure 4b shows the maximum length of consecutive missing glucose values per cluster and the intra-cluster division of these values. Cluster 3 has a longer period of missing consecutive data points compared to the other clusters. On the contrary, cluster 4 has the shortest consecutive period of missing data points. Thus cluster 3 is considered the least

accurate, considering the longest length of consecutive estimated data points, combined with the fact that cluster 3 has the highest percentage of missing data (Table 5). Table 5 shows the percentage of consecutive missing data points of all the clusters combined and for the separate clusters. Cluster 1 shows the lowest percentage of total missing data points followed by cluster 2, 4 and 5.

Table 5: Percentage of total missing data points

	All	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Missing Data (%)	3,05	2,17	2,40	5,27	2,53	3,27



(a) **(b)**

Figure 4: a) Silhouette plot which depicts the silhouette widths per patient and per cluster. On the right-hand side the number of patients per cluster and the corresponding average silhouette width of a cluster are shown. On the left-hand side the patient number with corresponding silhouette width is visualized. On the bottom left-hand side the total average silhouette width can be observed. b) Boxplot showing the distribution of maximum consecutive missing data points per cluster. With the cluster numbers on the x-axis and the maximum number of consecutive missing data points on the y-axis. The total considered amount of data points was 1344, which is two weeks of 96 data points per day.

The Kruskal-Wallis significance test revealed that all CGM metrics showed significant differences ($p<0,05$) on an overall base (Table 6), but differed between clusters. An overview of the mean values of all clusters and how the values relate to each other on a scale from lowest to highest can be seen in Table 7. Below the characteristics of all clusters based on the features are discussed.

1. Cluster 1 consisted of 27 patients (34,6%) which showed high hypoglycemia exposure and high hypoglycemia incidence, The hyperglycemia exposure was on the low side while the hyperglycemia incidence was moderate. SD and CV were both moderate resulting in a moderate glycemic variability.

2. Cluster 2 had 22 patients (28,2%) which showed the lowest hypoglycemic exposure and incidence with a moderate hyperglycemia exposure and the highest hyperglycemia incidence. GV was moderate based on both CV and SD.

3. Cluster 3 included 18 patients (23,1%) which showed moderate hypoglycemia exposure and high incidence in the night but low hypoglycemia incidence in the other time blocks. Hyperglycemia exposure was the highest for the entire day, with moderate hyperglycemia incidence. Glycemic variability was moderate in all time blocks and the mean glucose values were the highest.

4. Cluster 4 had 6 patients (7,7%) which showed a low hypoglycemia incidence in the night and afternoon and moderate in the morning and evening. hyperglycemia exposure, hyperglycemia incidence, GV and mean glucose were the lowest in this cluster.

5. Cluster 5 included 5 patients (6,4%) which showed the highest hypoglycemia exposure and incidence during the day, with relatively low hyperglycemia exposure and moderate incidence. Cluster 5 showed the highest glycemic variability based on the CV.

Overall, cluster 4 showed the lowest metric rates and cluster 5 the highest as can be seen in Table 7. Cluster 1 did not excel in any value staying close to the mean of all patients. The coefficient of variation in cluster 3 and 5 exceeds the CV threshold of 36% that has previously been proposed to indicate unstable glycemia and increased risk for hypoglycemia. For cluster 1 this threshold is exceeded in the afternoon and evening.

Table 6: Table containing all metrics with their features, calculated for all patients combined and for the seperate clusters (CL).

	All	CL1	CL2	CL3	CL4	CL5	p[1]
Amount of Patients	n= 78	n= 27	n= 22	n= 18	n= 6	n= 5	
Hypoglycemia Exposure							
% Time below Range, Night (<3.9 mmol)	5,59	6,27	2,00	6,99	3,40	15,29	0,001
% Time below Range, Morning (<3.9 mmol)	3,61	4,45	1,19	2,02	3,47	15,60	<0,0001
% Time below Range, Afternoon (<3.9 mmol)	4,08	5,31	1,85	2,98	3,08	12,44	<0,0001
% Time below Range, Evening (<3.9 mmol)	3,49	4,66	1,00	2,56	4,32	10,48	<0,0001
Hypoglycemia Incidence							
Episodes Hypoglycemia, Night >= 15 minutes	3,14	3,56	2,00	3,22	2,67	6,20	0,0172
Episodes Hypoglycemia, Morning >= 15 minutes	3,27	4,04	1,73	2,11	3,50	9,80	0,00334
Episodes Hypoglycemia, Afternoon >= 15 minutes	5,50	7,33	3,36	3,56	3,83	14,00	<0,0001
Episodes Hypoglycemia, Evening>= 15 minutes	3,87	5,41	1,59	2,72	4,33	9,20	<0,0001
Hyperglycemia Exposure							
% Time above Range, Night (<3.9 mmol)	33,32	24,44	39,06	47,03	9,12	35,67	<0,0001
% Time above Range, Morning (<3.9 mmol)	34,71	28,91	38,16	51,67	7,84	22,08	<0,0001
% Time above Range, Afternoon (<3.9 mmol)	34,18	24,83	37,54	53,89	14,58	22,44	<0,0001
% Time above Range, Evening (<3.9 mmol)	40,37	29,75	47,82	59,11	15,08	27,74	<0,0001
Hyperglycemia Incidence							
Episodes Hyperglycemia, Night >= 15 minutes	8,72	7,67	10,36	9,61	4,83	8,60	0,00219
Episodes Hyperglycemia, Morning >= 15 minutes	12,19	12,30	13,95	13,00	5,67	8,80	0,00014
Episodes Hyperglycemia, Afternoon >= 15 minutes	12,87	11,56	15,41	14,61	6,50	10,20	<0,0001
Episodes Hyperglycemia, Evening>= 15 minutes	12,94	12,74	15,59	12,39	7,00	11,40	<0,0001
Glycemic Variability							
Standard Deviation, Night	3,18	2,82	2,89	4,12	2,03	4,43	<0,0001
Standard Deviation, Morning	3,07	2,96	2,77	3,85	1,81	3,68	<0,0001
Standard Deviation, Afternoon	3,28	3,05	2,97	4,34	2,09	3,51	<0,0001
Standard Deviation, Evening	3,47	3,26	3,17	4,58	2,26	3,37	<0,0001
Coefficient of Variation, Night	36,05	35,89	31,16	41,29	28,28	48,85	<0,0001
Coefficient of Variation, Morning	33,97	34,99	29,72	36,56	25,96	47,48	<0,0001
Coefficient of Variation, Afternoon	36,32	37,75	32,00	40,14	27,28	44,73	<0,0001
Coefficient of Variation, Evening	36,22	38,23	31,40	39,67	29,98	41,69	<0,0001
Mean Glucose Value							
Mean Glucose, Night	8,83	7,91	9,32	10,07	7,15	9,13	<0,0001
Mean Glucose, Morning	9,04	8,50	9,35	10,58	6,99	7,60	<0,0001
Mean Glucose, Afternoon	9,00	8,05	9,25	10,91	7,68	7,73	<0,0001
Mean Glucose, Evening	9,61	8,52	10,15	11,66	7,53	8,20	<0,0001

Grey indicates the average value for all patients, light green, green, yellow, orange and brown indicate the lowest value of all clusters, a low, a moderate, a high or the highest value respectively. [1] significantly different if p <0,05;

The Dunn's significance method was applied to see inter-cluster relations. This method revealed that cluster 5 only correlated significantly with cluster 1 on one feature, standard deviation in the night. The correlation number was 0,027 which is significant but weak. After further consideration it was discovered that when four clusters were chosen instead of five, cluster 5 would merge with cluster 1. Additionally the Dunn's significance method revealed that cluster 4 is

significantly different to cluster 1 and cluster 5 on six features. If three clusters would be computed instead of five, cluster 4 would merge with cluster 1 and 5.

Table 7: Mean values of the considered metrics per cluster (Cl) and of all the patients combined.

	All	Cl1	Cl2	Cl3	Cl4	Cl5
Hypoglycemia Exposure	4,19	5,17	1,51	3,64	3,57	13,45
Hypoglycemia Incidence	3,95	5,08	2,17	2,92	3,58	9,80
Hyperglycemia Exposure	35,64	26,98	40,65	52,92	11,65	26,98
Hyperglycemia Incidence	11,68	11,06	13,83	12,40	6,00	9,75
Glycemic Variability	19,45	19,87	17,01	21,82	14,96	24,72
Mean Glucose Value	9,12	8,25	9,52	10,81	7,33	8,17

Index[1]
Average
Lowest
Low
Medium
High
Highest

[1] Index showing the definition of the used colours.

Time Block Differences

Boxplots of the selected metrics show insight in the division of values within and between clusters. Showing the median width, in colour, the interquartile range. Outliers are indicated by dots and are considered outliers when the value is 1,5 times the interquartile range away from the top or bottom of the box. Only the boxplot of one metric will be discussed, Figure 5, the other boxplots can be found in Figure A.3, A.4, A.5

Figure 5 shows how similar the different time block features are, with every time block indicating similar relations between clusters. The hyperglycaemic incidence is lowest in each time block for cluster 3, and highest for cluster 2. The interquartile ranges are small for cluster 5, especially in the morning while cluster 1 shows large interquartile ranges in the morning.

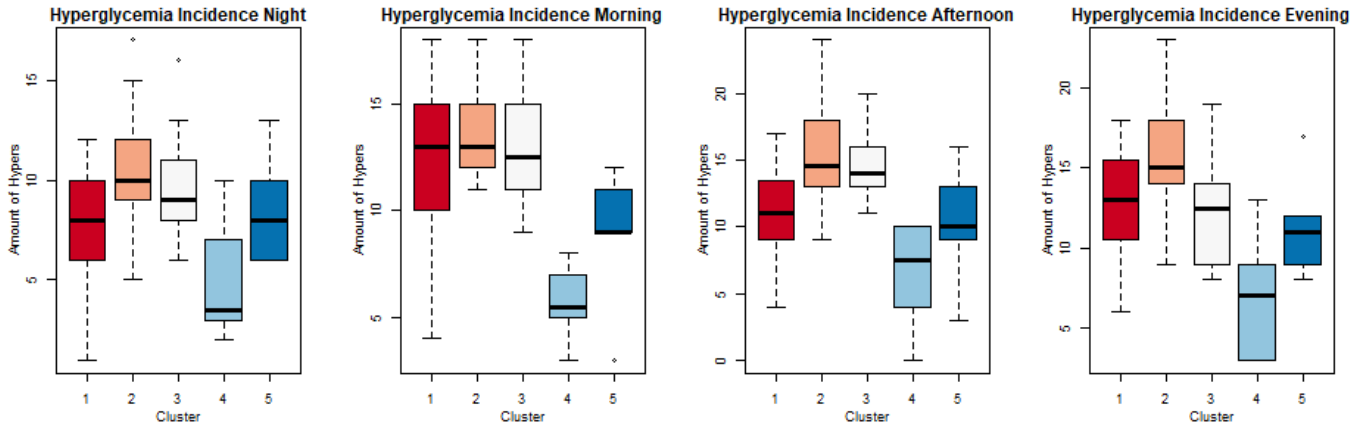


Figure 5: Boxplot of the hyperglycemic exposure lasting 15 minutes or longer of the different time blocks. with on the y-axis the amount of Hypoglycemic episodes and on the x-axis the cluster numbers.

In addition, the current study examined the overall differences in values between time blocks. Figure 6 shows two boxplots with metrics taken across all 78 patients. Hyperglycemic incidence appears to be lower in the night while morning, afternoon and evening appear to be similar. Hypoglycemic incidence is the lowest in the morning, however appeared to not be significantly lower compared to other time blocks. There are quite some outliers in the morning and afternoon, revealing that there are some patients with a lot of hypoglycemic episodes during these time blocks. The evening and nighttime blocks appear to be more steady with just one outlier.

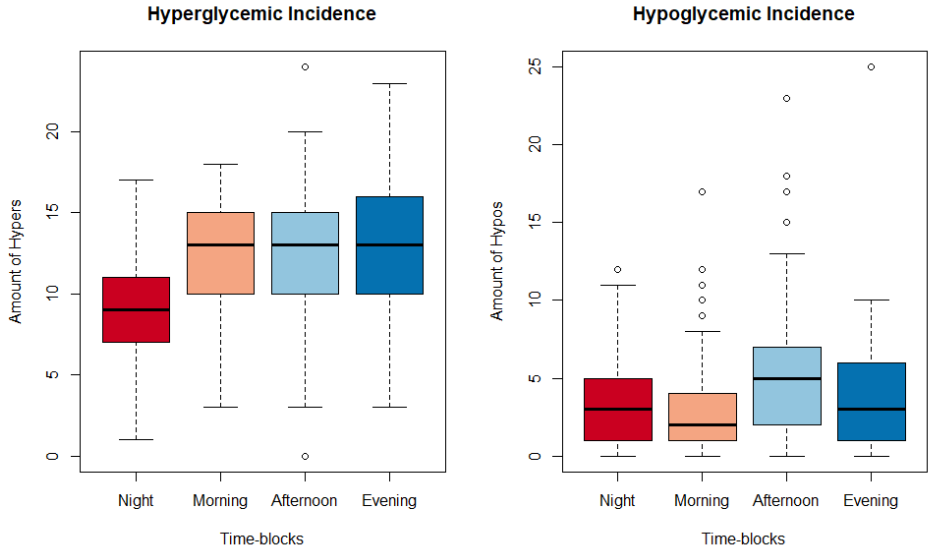


Figure 6: Boxplot of all patients, considering different time blocks, on the left-hand side the hyperglycemic incidence and on the right-hand side the hypoglycemic incidence. On the x-as the time blocks are depicted, with on the y-axis the amount of hyperglycemia episodes and in the left figure the amount of hypoglycemia episodes.

Discussion

This was the first study to explore features stratified by four time blocks in combination with CGM cluster analysis. It appeared that there were no mayor differences between time blocks regarding hyper-and hypoglycemia incidence, except for a lower hyperglycemic incidence in the night. This could be due to overall lower dietary intake during the night. Additionally this study used a combination of features, exclusion criteria and cluster analysis methods that is thus far unique.

The current study took a first step in identifying different phenotypes of T1D. Five clusters were identified, in which cluster 2, 3 and 4 differed significantly from each other. Cluster 1 has an increased risk of developing kidney disease due to a high albumin/creatin ratio. Cluster 2 is the least prone to severe hypoglycemic complications such as seizures or coma due to low hypoglycemia features. Cluster 3 had increased risk of diabetes complications such as increased risk of CVD or kidney, eye and heart failure due to high cholesterol values, the highest hyperglycemic incidence and high HbA1c values. This high HbA1c value corresponded with a observed high mean glucose value. On the contrary, cluster 4 had the lowest; cholesterol values, hyperglycemia features, GV and lowest mean glucose values which makes this cluster the least prone to develop diabetes related complications. Finally, cluster 5 had the highest hypoglycemia features and the highest GV which is associated with increased risk of seizures or CVD.

Feature Selection and Reduction

Some features recommended by the ATTD consensus were not taken into account given the time-restraints of the current study. Therefore, future studies should include all features mentioned by the ATTD consensus to further investigate their importance. Moreover, only metrics given by the ATTD consensus should be considered but also other metrics, such as metrics derived from the frequency domain, could be important for evaluation of raw CGM data. Composite metrics from the overview of Nguyen *et al.* [43] can for example be used to describe the quality of a set of CGM data for defining glycemic control. Fico *et al.* [44] explored the frequency domain of CGM to find important metrics, and still more could be discovered in future research.

Differences between level one TBR (glucose <3,0 mmol/L) and level two TBR (glucose < 3,9 mmol/L) were assessed but level one was not included in the study because of a previous study by Kahkoska *et al.* [7] who mentioned that they were highly correlated. Differences between level one TAR (glucose >13,9 mmol/L) and level two TAR (glucose > 10 mmol/L) were also assessed but the same conclusion was drawn as for TBR, level one was excluded.

For feature reduction, a correlation plot was used in combination with a correlation matrix showing the correlation values. Then the "Findcorrelation" function in R showed the correlating features and decided, based on the means, which feature was excluded. However, the exact mechanism behind this function was unclear. Therefore, an alternative method to reduce feature dimensionality such as the Principal Component Analysis (PCA) in combination with a scree plot could be considered. PCA allows a massive amount of information enclosed in initially correlated data to be transformed into a set of new orthogonal components [45]. The components found by PCA could consist of one feature or a combination of multiple features. Pham *et al.* [46] used this method to place diabetic patients into subgroups and used three principal components as input for clustering analysis. PCA is the most common used method for filtering out repetitive data and finding the most important components representing the original data. In figure A.6 an indication is given on how the found clusters in this study are visualized based on two PCA components.

Identifying Clusters

In the current study, many decisions had to be made for executing the analysis that led to the identification of clusters. This involved, for example, the number of consecutive days of sufficient data needed, the exclusion of seasons in the analysis, the data analysis period and certain methodological steps such as the linkage method. In doing the analysis, there are no optimal choices yet and all decisions made impacted the result. Therefore, it is important to investigate the effect of alternative choices on the results.

Some alternative choices were investigated to explore which results would be affected. First, the effect of including seasons was investigated, which would make the overall circumstances were patients are in more comparable. The affected parameters are displayed in Table A.5. The optimal linkage method did not change for any season or any

other choice. Second, the change in consecutive days and its effects on the parameters were investigated (Table A.5). Interestingly, the optimal cluster number varies with alternative choices made. This table shows how every choice has an effect on the outcome, making it complicated to find the optimal number of clusters. Additionally, there are other methods and indexes for determining the optimal number of clusters. For example, Pollé *et al.* [18] used a scree plot based on the dendrogram height to determine the number of clusters and Lugner *et al.* [47] used, among others techniques, the gap-statistics method to establish the optimal number of clusters.

As mentioned before, the elbow method and dendrogram were chosen as methods for determining the optimal number of clusters and different methods are available to determine this. For example-, the gap-statistics method would have given two clusters as the optimal number of clusters, but is more often used in combination with K-means clustering so was therefore discarded in the current study. Because every method resulted in a different optimal number of clusters, the best option was chosen each time and was later evaluated on appropriateness. It appeared that there could be an error in the method of defining the optimal number of clusters, because the number of clusters could also be four based on the non-significance between cluster 1 and 5 and the low silhouette width of these clusters.

Due to the complexity of handling missing data, it was decided to choose the method that involved rounding the given times to the nearest 15 minutes. For example 17:08 then it became 17:15, causing a maximum deviation of seven minutes. Although this deviation is sub-optimal, it was the most feasible option given the time-restraints of the current study. However, further research, should use the exact given timestamps and implement those in the cluster analysis.

Another limitation concerning missing data is the method of filling in missing data points. The missing data points were replaced with the previously known value. An exception was when the first values of the data were missing, then the next known value was taken. Future studies should take the value of the previous and succeeding value of the missing data point and take the mean, as this would be a better method for replacing missing data points. The maximum missing data points was observed to be just below 30 data points. This is however not favourable because this means more than seven hours of data was missing and all these hours just one value was assumed. In future research this is something to consider and possibly define a separate exclusion criteria stating the maximum length of allowed consecutive missing glucose values over a certain time period.

The current study used data from two weeks of the year 2021. However, these two exact weeks differed per patient. To ensure more comparable circumstances of patients, the suggestion is to use the same two weeks for each patient if sufficient data is available. This ensures the exclusion of inconsistencies regarding public holidays, weather and ratio of weekdays and weekends. E. Gecili *et al.* [12] showed that glucose levels are higher on weekends compared to weekdays, so this is something to take into account in future research.

Cluster Evaluation

Seventy-eight patients were eligible for this study, which is a relatively small group of people to compare, especially considering there were five clusters. Two of the five cluster had six or less patients, resulting in deviating values for these specific clusters. To confirm and/or add to the findings of the current study, a larger scale study should be executed. Fortunately, such a study will be executed starting this year at the ZGT hospital in Hengelo and Almelo.

Not every known clinical parameter of the patients was included. This was mainly due to a lack of consistency, the selection of available parameters were different between patients. For example, only a few patients had albumin values, making it an unreliable parameter to compare. Insulin regulation (basal-and bolus volume, carbohydrate intake and frequency of insulin administration) was initially intended to be included in this study but while conducting the research, it appeared that insufficient data was accessible. Only a few patients per cluster had insulin regulation information available, with cluster five having zero available data. Insulin regulation however is an interesting parameter to consider and could be included in future research for cluster evaluation. Patients could significantly differ on insulin regulation, possibly forming a distinct phenotype. An interesting question could be, 'how does insulin regulation differ per cluster and what other remarkable features does that cluster have that could relate to this regulation'.

Relatively, there was a significant amount of data missing regarding the BMI, smoking habits and alcohol intake, conceivably due to the limitation placed on the time span of considered data. It could be that the patient was weighted in

the foregoing year and that the patient indicated they stayed on the same weight, which meant that no new measurement was taken. The alcohol intake is given by either indicating yes or no, which means not a full picture of alcohol intake is provided. A more intuitive approach would be asking how many alcohol consumptions the patient takes across a certain time period. This lack of appropriate alcohol consumption estimate makes that no conclusions can be drawn regarding alcohol intake. Hospitals should consider to change the method for assessing alcohol consumption during patient consults and, should thereby improve knowledge of their patients and a reliable measure of alcohol consumption could consequently be used to improve parameter analysis between clusters.

Clusters could be assessed by having more patient specific information, involving information about physical activity, dietary intake and psychosocial characteristics (e.g. motivation, quality of life and depression symptoms). This would give a more complete picture of the patient, leading to quicker identification of causes of hypoglycemic and hyperglycemic episodes. However, Kahkoska *et al.* [7] found no significant differences between clusters based on psychosocial characteristics but discussed that the lack of statistically significant correlations could be the small sample size which may limit statistical power.

Albumin/Creatin-ratio was, compared to other clusters, high in cluster 5, due to one patient having a value larger than 30 mg/mmol, which is associated with a possibility of kidney disease. It could be the case that this patients was ill at the time of the measurement without knowing. Since cluster 5 only has five patients, the influence was substantial. The same applied in some degree for cluster 1 where two patients had an Albumin/Creatin-ratio greater than 30 mg/mmol and thus resulted in an impaired overall mean value of Albumin/Creatin-ratio for cluster 1.

It is unsure whether five clusters is actually the optimal number of clusters as certain observations took this into question. The Dunn's index revealed that there was only one significantly different feature between cluster 1 and 5. Additionally the silhouette width values revealed that the patients in cluster 5 were placed incorrectly and that some patients in cluster 1 were placed incorrectly. Currently, cluster 1 appears to consist of the patients which did not belong to another cluster, possibly indicating an extra cluster could be formed or clusters could be merged. Because of the lack of dissimilarities and the low silhouette value, possibly, cluster 1 and 5 could be merged to form one cluster. Besides, cluster 4 only had a few significantly different features but in contrast did have a reasonable silhouette width, which leads to the need of more advance research.

In hierarchical clustering, outliers are forced to merge with the nearest cluster [48], but the downside is that this could generate noise. There is a method that uses a threshold of similarity in which, outliers will be discarded when the distance between outliers and one sample in clusters is more than the threshold. Hu *et al.* [48] discovered a method to merge hierarchical clustering and the threshold of similarity into a new method called 'threshold-based hierarchical clustering'. This method includes the forming of a single-sample cluster when the threshold is exceeded [48]. Future studies should consider this threshold since outliers could be discarded which would optimize results. Currently, outliers are included even though they could impair the results.

In the current study, the conclusion if different phenotypes are found are solely based on if features differ significantly and if the silhouette width is reasonable. To validate if distinct phenotypes are found and how this can be concluded more extensive studies are needed.

As mentioned in the introduction, this study was based on the two previously done studies by pollè *et al.* [18], and Kahkoska *et al.* [7]. Kahkoska *et al.* [7] identified three distinct, clinically meaningful clusters sharing phenotypes defined by different exposure to and incidence of hypoglycemia and glycemic variability. Similar metrics were used to distinguish clusters in this study but in this study five different clusters were identified. pollè *et al.* [18] identified four clusters who were found distinct across the PCA and also showed distinctive glycemic patterns. The groups all differed significantly from each other for all CGM metrics ($p < 0.05$), which was not the case in the current study. Although there was an overall significant difference, the between cluster differences were not always significant. This gives a short indication of how the various researches have different results. It is however difficult to compare to other studies due to the method differences and the difference in used metrics and analysis methods.

Clinical Relevance

Multiple points of clinical relevance for the results can be discussed. The identification of key combinations of clinical metrics from CGM data would contribute to the improvement of the process of identifying possible diabetes related issues. This study involves a new combination of features to define ATTD metrics and shows how these can be used to establish different clusters within T1D patients. This could lead to diabetes care givers being able to recognize patients as certain clinical phenotypes. Subsequently tailored individual approaches for T1D patients such as, insulin regulation advice for clusters characterized with severe hyperglemia, could be implemented to improve care. The availability of CGM data keeps increasing with the use of devices such as the freestyle libre becoming more available and normalized. More data could be collected, and better digital record storage could lead to advance databases for finding clinical relevant phenotypes. Possibly, apps could be developed to give direct standardized advice based on the phenotypes, thereby decreasing the glycemic variability and improving HbA1c. Eventually CGM data analysis could become the new gold standard for the prognosis and diagnosis of DM.

Conclusion

In conclusion, this study showed that CGM data analysis can be used to discover different subgroups of T1D, which could eventually result in more personalized health care. Five clusters were identified by using agglomerative hierarchical clustering analysis on the Diabase cohort, based on glycemic features. The aims set at the start of the study were met. Twenty relevant features for clusters analysis were identified which can be used for cluster analysis. Different methods for the identification of clusters such as different clustering methods, linkage methods and methods to determine the optimal number of clusters, were explored. When the clusters were identified, the third aim of the study, to compare clusters based on known clinical parameters, was met. A select number of differences such as LDL, HDL and HbA1c differences between clusters were found. Finally, clusters were tested for significant differences, which unravelled that not all clusters had significant differences. However, to validate if distinct phenotypes of T1D were found, additional research is needed.

References

[1] diabetes federation I. IDF Diabetes Atlas 2021; 2021. https://diabetesatlas.org/.

[2] Mark A Atkinson AWM George S Eisenbarth. Type 1 diabetes. The Lancet. 2014;383(9911). https://www.sciencedirect.com/science/article/pii/S0140673613605917.

[3] Soumya D SB. Late Stage Complications of Diabetes and Insulin Resistance. Diabetes Metabolism. 2011;2(9). https://www.iomcworld.com/open-access/late-stage-complications-of-diabetes-and-insulin-resistance-2155-6156.1000167.pdf.

[4] Battelino T BRASBRBTBE Danne T. Clinical Targets for Continuous Glucose Monitoring Data Interpretation: Recommendations From the International Consensus on Time in Range. Diabetes Care. 2019;42(8):1593-602. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6973648/.

[5] Vettoretti M AGFASG Cappon G. Continuous Glucose Monitoring: Current Use in Diabetes Management and Possible Future Applications. Journal of Diabetes Science and Technology. 2018;12(5):1064-71. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6134613/.

[6] Danne T BTBRea Nimri R. International Consensus on Use of Continuous Glucose Monitoring. Diabetes care. 2017;40(12):1631-40. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6467165/.

[7] Kahkoska AR AABKBJCJMDNCKMMDE Adair LA. Identification of clinically-relevant dysglycemia phenotypes based on continuous glucose monitoring data from youth with type 1 diabetes and elevated hemoglobin A1c. Pediatr Diabetes. 2019;20(5):556-66. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6625874/.

[8] : Tao R LJea Yu X. Multilevel clustering approach driven by continuous glucose monitoring data for further classification of type 2 diabetes. s BMJ Open Diab Res Care. 2021;9(2146):1593-602. https://drc.bmj.com/content/bmjdrc/9/1/e001869.full.pdf.

[9] Richesson RL WDBBFMMMHWCRSS Rusincovitch SA. A comparison of phenotype definitions for diabetes mellitus. Am Med Inform Assoc. 2013;20(2). https://pubmed.ncbi.nlm.nih.gov/31603912/.

[10] Deutsch AEUMS A J. Phenotypic and genetic classification of diabetes. Diabetologia). 2022;65. https://link.springer.com/article/10.1007/s00125-022-05769-4.

[11] Oana P Zaharia ASGJBYKSAKB Klaus Strassburger. Risk of diabetes-associated diseases in subgroups of patients with recent-onset diabetes: a 5-year follow-up study. The Lancet Diabetes Endocrinology). 2019;7(9). https://www.sciencedirect.com/science/article/pii/S2213858719301871.

[12] Gecili E KJKEAMBKSR Huang R. Functional data analysis and prediction tools for continuous glucose-monitoring studies. Journal of Clinical and Translational Science. 2020;5(1). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8057494/.

[13] Kahkoska AR JXea Nguyen CT. Characterizing the weight-glycemia phenotypes of type 1 diabetes in youth and young adulthood. BMJ Open Diabetes Research Care). 2020;8. https://drc.bmj.com/content/8/1/e000886.

[14] Szczesniak RD DLAMMMKJ Li D. Longitudinal patterns of glycemic control and blood pressure in pregnant women with type 1 diabetes mellitus: phenotypes from functional data analysis. Am J Perinatol. 2016;13. https://pubmed.ncbi.nlm.nih.gov/27490775/.

[15] D R. Continuous Glucose Monitoring: A Review of Recent Studies Demonstrating Improved Glycemic Outcomes. Diabetes technology and therapeutics. 2017;19(S3):S25-37. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5467105/.

[16] Battaglia M AMAMBD Ahmed S. Introducing the Endotype Concept to Address the Challenge of Disease Heterogeneity in Type 1 Diabetes. Diabetes care. 2020;43(1):5-12. https://pubmed.ncbi.nlm.nih.gov/31753960/.

[17] H Hall ABPLRKTMMS D Perelman. Glucotypes reveal new patterns of glucose dysregulation. PLOS BIOLOGY. 2018;7(16). https://journals.plos.org/plosbiology/article/metrics?id=10.1371/journal.pbio.2005143.

[18] Pollé OG MMLJGIdBMSNLMMTGLLP Delfosse A. Glycemic Variability Patterns Strongly Correlate With Partial Remission Status in Children With Newly Diagnosed Type 1 Diabetes. Diabetes care. 2022;45(10):2360-8. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9862313/.

[19] Kietsiriroje N OLWDARARCM Pearson SM. Glucose variability is associated with an adverse vascular profile but only in the presence of insulin resistance in individuals with type 1 diabetes: An observational study. Diabetes and vascular disease research. 2019;19(3). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9168893/.

[20] Lobo B SMKB Farhy L. A Data-Driven Approach to Classifying Daily Continuous Glucose Monitoring (CGM) Time Series. IEEE Trans Biomed Eng. 2022;69(2):654-65. https://pubmed.ncbi.nlm.nih.gov/34375274/.

[21] YINAN MAO ASPWSATAARC KYLE XIN QUAN TAN. Stratification of Patients with Diabetes Using Continuous Glucose Monitoring Profiles and Machine Learning. Health data science. 2022. https://spj.science.org/doi/10.34133/2022/9892340?permanently=true.

[22] RM B. Understanding Continuous Glucose Monitoring Data. Role of Continuous Glucose Monitoring in Diabetes Treatment. 2018. https://www.ncbi.nlm.nih.gov/books/NBK538967/.

[23] Krishna SMK Surabhi Kota. Glycemic variability: Clinical implications. Indian journal of endocrinology and metabolism. 2013;17(4). https://pubmed.ncbi.nlm.nih.gov/23961476/.

[24] Nanyi Fei ZLTX Yizhao Gao. Z-Score Normalization, Hubness, and Few-Shot Learning. International Conference on Computer Vision. 2021. https://openaccess.thecvf.com/content/ICCV2021/html/Fei$_Z$ − $Score_Normalization_Hubness_and_Few-Shot_Learning_ICCV_2021_paper.html$.

[25] Raykov YP BFLM Boukouvalas A. What to Do When K-Means Clustering Fails: A Simple yet Principled Alternative Algorithm. PLoS One. 2016;11(9). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5036949/.

[26] Ankita Dubey AC. A Systematic Review on K-Means Clustering Techniques. International Journal of Scientific Research Engineering Technology. 2017;6(6).

[27] Kehar Singh NS Dimple Malik. Evolving limitations in K-means algorithm in data mining and their removal. International Journal of Computational Engineering Management,. 2011;12. http://ijcem.org/papers42011/42011$_2$6.pdf.

[28] Shukla S, S N. A Review ON K-means DATA Clustering APPROACH. International Journal of Information Computation Technology. 2014;4(17). https://www.ripublication.com/irph/ijict$_spl/ijictv4n17spl_1$5.pdf.

[29] Shi XGY Na Liu. Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm. 3rd International Symposium on Intelligent Information Technology and Security Informatics. 2010. https://www.researchgate.net/publication/221600313$_Research_on_k$ − $means_Clustering_Algorithm_An_Improved_k$ − $means_Clustering_Algorithm$.

[30] H Whittingham SA. The Era of Artificial Intelligence, Machine Learning, and Data Science in the Pharmaceutical Industry. Academic Press. 2021:81-102. https://doi.org/10.1016/B978-0-12-820045-2.00006-4.

[31] Fionn Murtagh PL. Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm. Journal of Classification. 2014;31(3). https://arxiv.org/pdf/1111.6285.pdf.

[32] Strauss T vMM. Generalising Ward's Method for Use with Manhattan Distances. PLoS One. 2017;12(1). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5235383/.

[33] Bu J PZLK Liu W. Comparative Study of Hydrochemical Classification Based on Different Hierarchical Cluster Analysis Methods. nt J Environ Res Public Health. 2020;17(24). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7766391/.

[34] Olson CF. Parallel algorithms for hierarchical clustering. Parallel Computing. 1995;21(8):1313-25. https://doi.org/10.1016/0167-8191(95)00017-I.

[35] Hartigan JA. Statistical Clustering. International Encyclopedia of the Social Behavioral Sciences. 2001:15014-9. https://doi.org/10.1016/B0-08-043076-7/00400-9.

[36] agnes.object: Agglomerative Nesting (AGNES) Object;. https://www.rdocumentation.org/packages/cluster/versions/2.1.4/top

[37] Absalom E Ezugwu OOOLAJOACIEAAA Abiodun M Ikotun. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. Engineering Applications of Artificial Intelligence. 2022;110. https://doi.org/10.1016/j.engappai.2022.104743.

[38] l Abdalla H. A Brief Comparison of K-means and Agglomerative Hierarchical Clustering Algorithms on Small Datasets. WCNA. 2022. https://link.springer.com/chapter/10.1007/978-981-19-2456-9_64.

[39] Aggarwal HAKDA C C. On the Surprising Behavior of Distance Metrics in High Dimensional Space. Lecture Notes in Computer Science. 2021;1973. https://link.springer.com/chapter/10.1007/3-540-44503-x_27.

[40] M Forina VR C Armanino. Clustering with dendrograms on interpretation variables. Analytica Chimica Acta. 2002;454(1):13-9. https://www.sciencedirect.com/science/article/abs/pii/S0003267001015173.

[41] Lin Z SC Laska E. A general iterative clustering algorithm. Statistical Analysis and Data Mining. 2022;15(4). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9438941/.

[42] Fatikhurrizqi AW Akhmad Wijayanto. COMPARISON OF REGIONAL CLUSTER ANALYSIS ACCORDING TO INCLUSIVE DEVELOPMENT INDICATORS IN JAVA ISLAND 2018 BETWEEN HIERARCHICAL AND PARTITIONING CLUSTERING STRATEGIES. JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer). 2021;6(2). https://www.researchgate.net/publication/$350110921_COMPARISON_OF_REGIONAL_CLUSTER_ANALYSIS_ACCORDING_TO_I$

[43] Michelle Nguyen EKSBPK Julia Han, Klonoff DC. A Review of Continuous Glucose Monitoring-Based Composite Metrics for Glycemic Control. Diabetes Technology Therapeutics. 2020;22(8). https://www.liebertpub.com/doi/full/10.1089/dia.2019.0434.

[44] Giuseppe Fico JC Liss Hernández. Exploring the Frequency Domain of Continuous Glucose Monitoring Signals to Improve Characterization of Glucose Variability and of Diabetic Profiles. Journal of Diabetes Science and Technology. 2017;11(4). https://journals.sagepub.com/doi/pdf/10.1177/1932296816685717.

[45] Changsheng Zhu WF Christian Uwa Idemudia. Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. Informatics in Medicine Unlocked. 2019;17. https://www.sciencedirect.com/science/article/pii/S2352914819300139.

[46] et al HNP. Predicting Hospital Readmission Patterns of Diabetic Patients using Ensemble Model and Cluster Analysis. International Conference on System Science and Engineering (ICSSE). 2019. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=arnumber=8823441.

[47] Lugner SSNSAMMMEOKEBFS Moa Gudbjörnsdottir. Comparison between data-driven clusters and models based on clinical features to predict outcomes in type 2 diabetes: nationwide observational study. Diabetologia. 2021;64. https://www.researchgate.net/publication/$352013666_comparison_between_data-driven_clusters_and_models_based_on_clinical_features_to_predict_outcomes_in_type_2_diabetes_nationwide_observational_study.$

[48] Hu M WYGY Zeng K. Threshold-Based Hierarchical Clustering for Person Re-Identification. Entropy. 2021;23(5). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8145342/.

A Appendix

A.1 Figures and Tables

Table A.1: Table which shows an extract of the original raw CGM data. With in the first column the patients number, the second the date, third the time, fourth column the glucose value and the last column showing the CGM device that was used.

22	4-6-2020	21:59:00	12.00	FreeStyle Libre
22	4-6-2020	22:16:00	11.10	FreeStyle Libre
24	3-11-2020	15:31:00	5.60	FreeStyle Libre
24	3-11-2020	15:46:00	6.30	FreeStyle Libre

Table A.2: Pre-processed data with added column names and a column showing the time of day.

DIN	Date	Time	Glucosewaarde	device	Time_of_day
22	2021-01-01	00:13:00	10,6	FreeStyle Libre	Night
22	2021-01-01	00:28:00	10,8	FreeStyle Libre	Night
22	2021-01-01	00:43:00	10,4	FreeStyle Libre	Night

Table A.3: Results after feature calculations. This is a sample of just the first features and patients to given an example of what the data looked like.

Patients	Meannight	Meanmorning	Meanafternoon	Meanevening	Sdnight	Sdmorning	Sdafternoon
22	7,624405	8,674405	7,263393	8,698512	2,608608	3,115441	2,603923
32	8,611607	8,202976	8,297917	7,823214	3,599637	2,890301	2,711473
36	10,17917	9,353274	9,641667	10,31161	2,891105	3,019042	2,959516

Table A.4: Scaled results, with the highly correlated features removed. Substract of input data for cluster analysis

Meannight	Sdnight	Sdmorning	Sdafternoon	Sdevening	CVnight	CVmorning
-0,7299	-0,61309	0,06218	-0,7428	-0,2608	-0,22711	0,30213
-0,13032	0,448866	-0,24132	-0,62454	-0,97595	0,71259	0,196303
0,821736	-0,31037	-0,06777	-0,35179	-0,26341	-0,94704	-0,26351

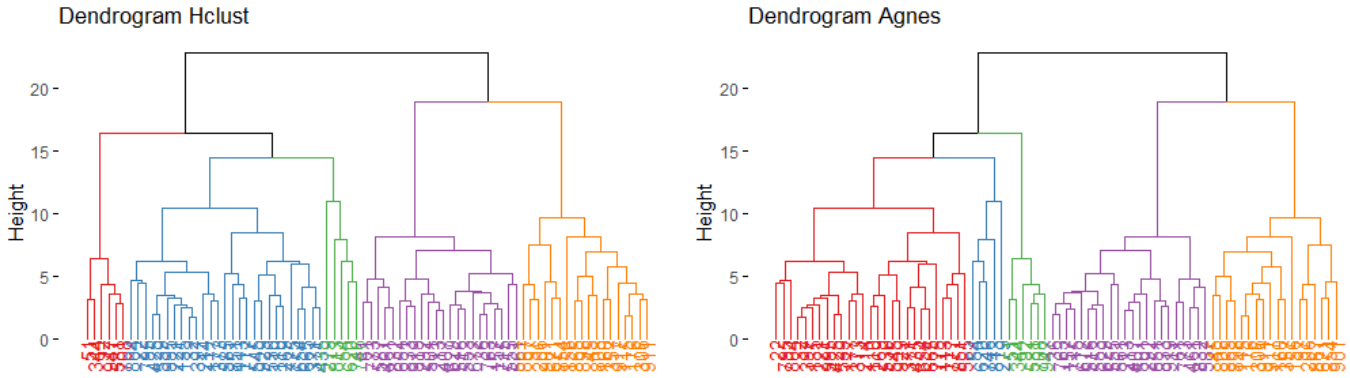


Figure A.1: Dendrograms for the agglomerative clustering functions, Agnes and Hclust, computed with the Ward's linkage method. On the y-axis the dendrogram height and on the x-axis the patients numbers. The colours indicate the different clusters with in red cluster 4, blue cluster 1, green cluster 5, purple cluster 2 and orange cluster 3.

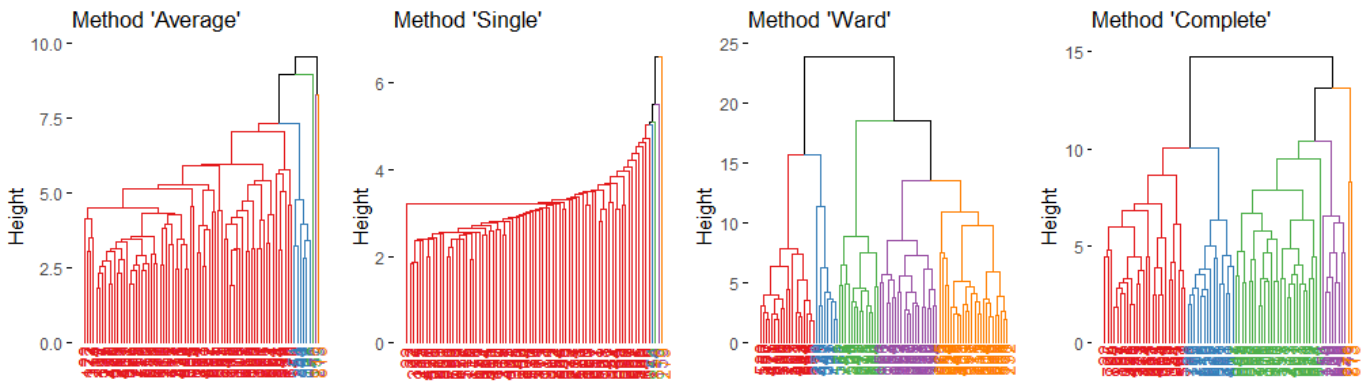


Figure A.2: Dendrograms for the different linkage methods, with Agnes as the used function

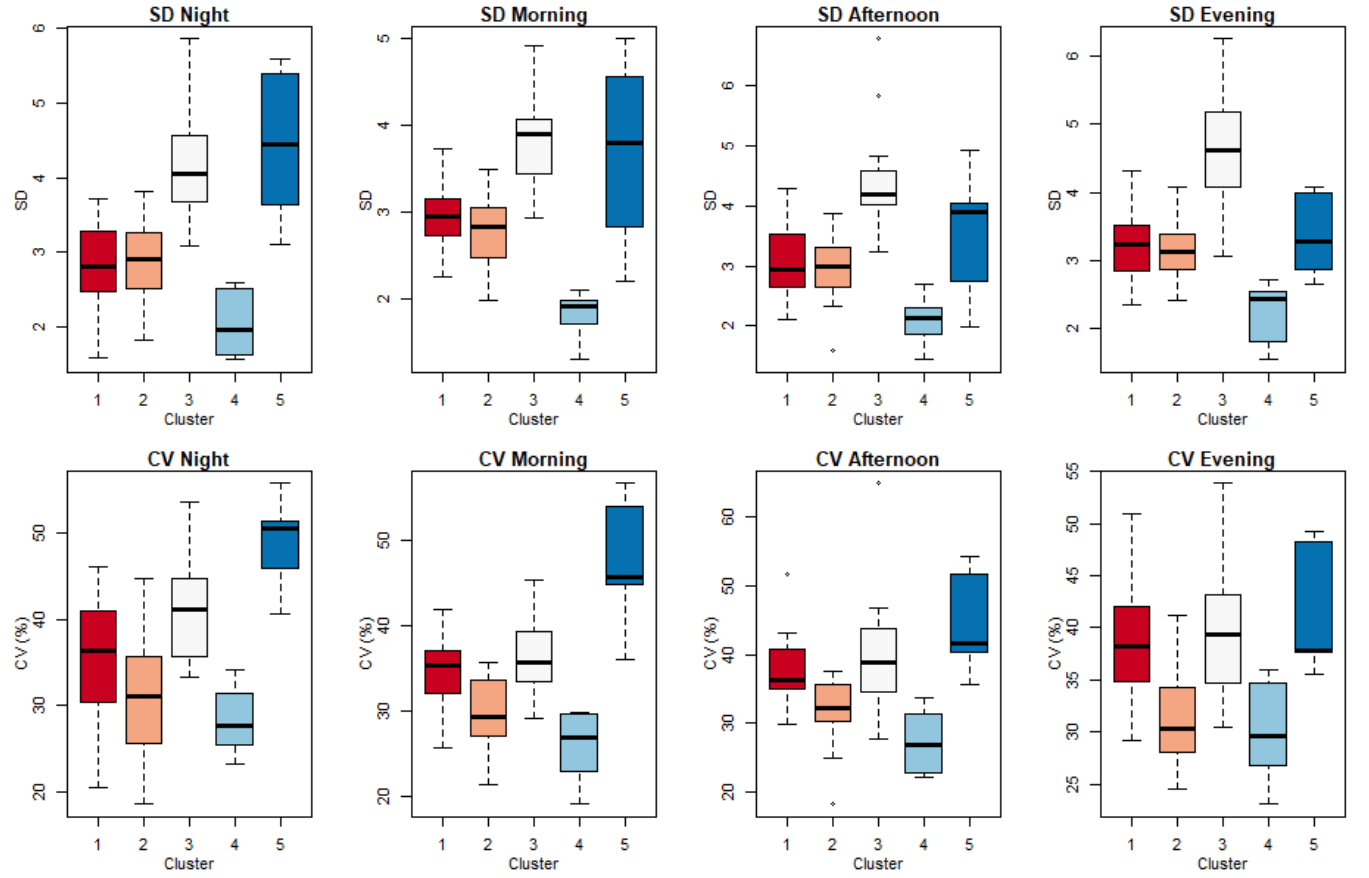


Figure A.3: Boxplot showing glycemic variability based on standard deviation (SD) and percentage coefficient of variation (CV).

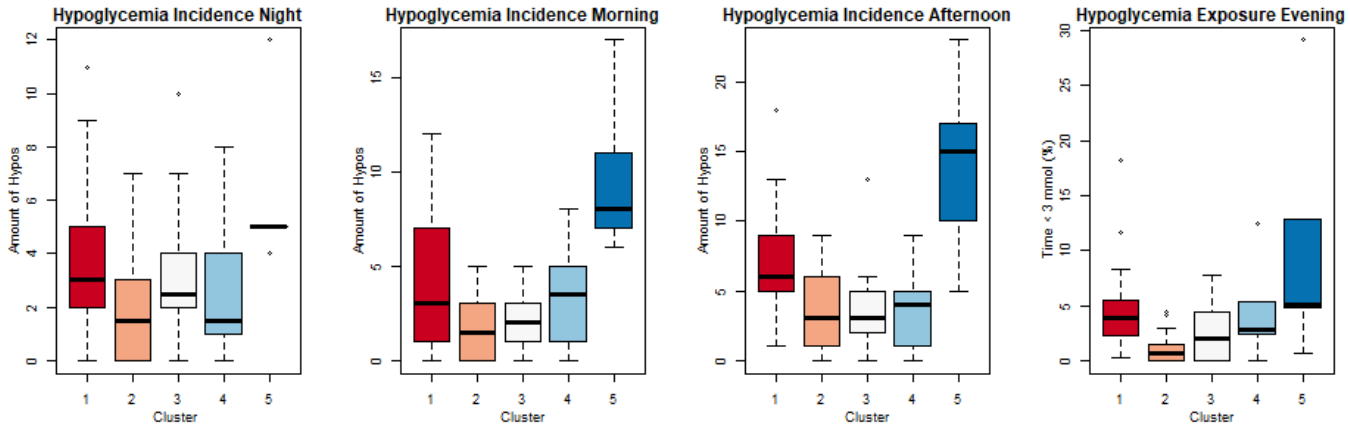


Figure A.4: Boxplot of the hypoglycemic incidence and on the right-side a boxplot of hypoglycemic exposure of the evening based on time below rang (TBR) of < 3,9 mmol/L. of the evening

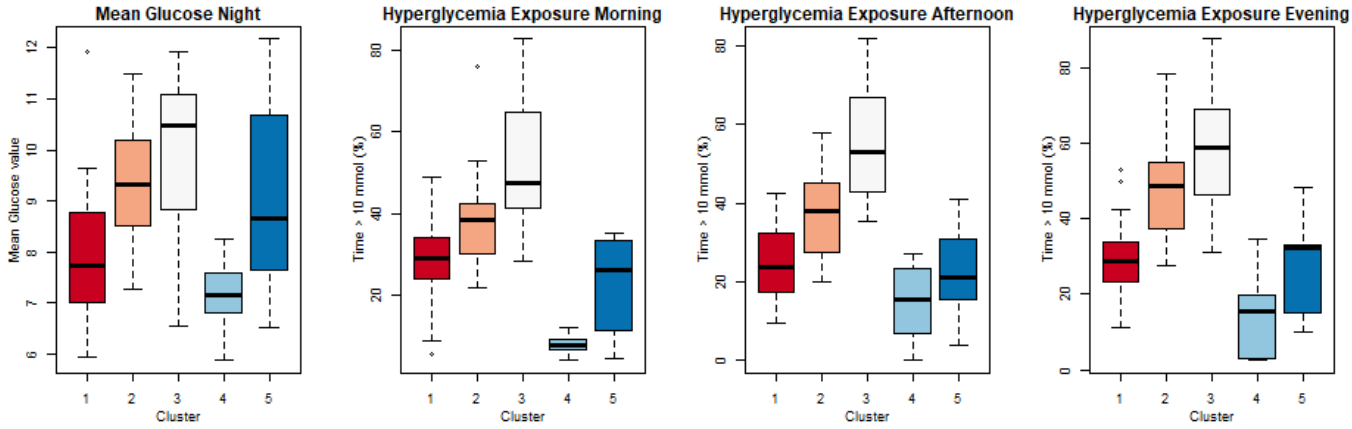


Figure A.5: Boxplot of the mean glucose value of the night on the left-hand side. The hyperglycemic exposure of the other time blocks are shown based on the percentage time above range (TAR)

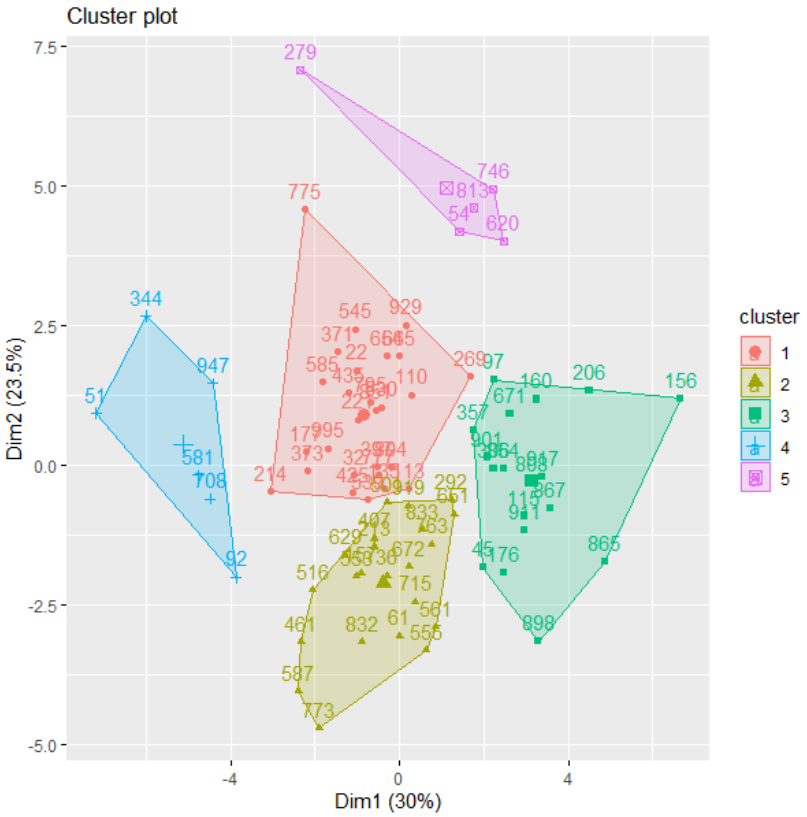


Figure A.6: Cluster plot showing the distribution of patients in clusters based on the first two principal components, which are automatically generated by the 'fvizcluster' function in R.

	Eligible patients	Features excluded	Optimal cluster amount	Patients per cluster
Season:				
Spring	69	9	4, 4	17-16-3-33
Summer	67	11	4, 4	38-12-6-11
Autumn	58	8	6, 5	28-9-4-15-2
Winter	57	7	2/4, 4	16-27-3-11
7 consecutive days	82	7	4, 4	27-18-7-17
21 consecutive days	69	8	3, 3	27-25-17

Table A.5: Table showing the different changes made and which parameters were affected. The optimal cluster amount shows two values, were the first is the optimal amount of clusters taken from the elbowplot and the second taken from the dendrogram.