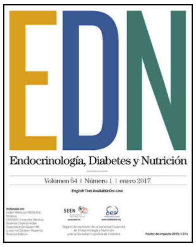
Endocrinología, Diabetes y Nutrición 68 (2021) 170−174

Endocrinología, Diabetes y Nutrición

www.elsevier.es/endo



ORIGINAL ARTICLE

Identification of blood glucose patterns in patients with type 1 diabetes using continuous glucose monitoring and clustering technique☆

Sergio Contador Pachón[a], Marta Botella Serrano[b], Oscar Garnica Alcázar[c], José Manuel Velasco Cabo[c], Aranzanzu Aramendi Zurimendi[b], Remedios Rodríguez Martínez[b], Esther Maqueda Villaizán[d], José Ignacio Hidalgo Pérez[c,*]

[a] *Universidad Rey Juan Carlos, Móstoles, Madrid, Spain*
[b] *Servicio de Endocrinología y Nutrición, Hospital Universitario Príncipe de Asturias, Alcalá de Henares, Madrid, Spain*
[c] *Departamento de Arquitectura de Computadores y Automática, Universidad Complutense de Madrid, Madrid, Spain*
[d] *Servicio de Endocrinología y Nutrición, Hospital Virgen de la Salud, Toledo, Spain*

Received 28 August 2019; accepted 24 December 2019

KEYWORDS
Diabetes mellitus;
Continuous glucose monitoring;
Classification;
Clustering;
Chi-square automatic detection interaction

Abstract
Objective: To show that statistical techniques allow for obtaining a reduced number of four-hour glucose profiles that can identify any glucose behavior in patients with type 1 diabetes mellitus.
Patients and Methods: A retrospective study of 10 patients with type 1 diabetes mellitus was conducted using data collected by continuous glucose monitoring. A data mining technique based on decision trees called CHAID (Chi-square Automatic Interaction Detection) was used to classify glucose profiles into groups using two decision criteria. These were 1, the seven days of the week and 2, different time slots, the day being divided into six sections of four hours each. Clustering was performed according to the glucose levels recorded using the statistically significant differences found.
Results: Significant differences (*P*-value <.05) and dependencies were seen between the glucose profiles classified depending on the independent variables 'day of the week' and 'time slot'. The relationships found were different for each patient, showing the need for individualized studies.

☆ Please cite this article as: Contador Pachón S, Botella Serrano M, Garnica Alcázar O, Velasco Cabo JM, Aramendi Zurimendi A, Rodríguez Martínez R, et al. Identificación de patrones de glucemia en pacientes con diabetes tipo 1 mediante monitorización continua de glucosa y técnicas de clusterización. Endocrinol Diabetes Nutr. 2021;68:170−174.

* Corresponding author.
E-mail address: hidalgo@ucm.es (J.I. Hidalgo Pérez).

https://doi.org/10.1016/j.endien.2021.06.001

Conclusions: The results obtained will facilitate mathematical modeling of glucose, and can be used to develop an individualized classifier for each patient that categorizes glucose profiles based on the day of the week and time slot variables. Using this classifier, it will be possible to predict the glucose levels of the patient knowing on which day of the week and in which time slot he/she is, leading to more precise models. Healthcare professionals will also be able to improve patient habits and therapies.

PALABRAS CLAVE
Diabetes mellitus;
Monitorización continua de glucosa;
Clasificación;
Clusterización;
Detección automática de interacción
Chi-cuadrado

Identificación de patrones de glucemia en pacientes con diabetes tipo 1 mediante monitorización continua de glucosa y técnicas de clusterización

Resumen

Objetivo: Demostrar que mediante técnicas estadísticas es posible obtener un número reducido de perfiles de glucosa de cuatro horas con los que se puede identificar cualquier comportamiento de la glucosa en pacientes con diabetes mellitus tipo 1.

Material y Métodos: Se ha realizado un estudio retrospectivo de diez pacientes con diabetes mellitus tipo 1, con datos adquiridos mediante monitorización continua de glucosa. Se ha utilizado una técnica de minería de datos basada en árboles de decisión denominada CHAID (CHi-square Automatic Interaction Detection), para clasificar los perfiles de glucosa en grupos utilizando dos criterios de decisión. Por un lado, los diferentes días (lunes; martes; miércoles; jueves; viernes; sábado; domingo), por otro, diferentes franjas del día, dividiendo el día en seis tramos de cuatro horas cada uno (00:00 h−04:00 h; 04:00 h−08:00 h; 08:00 h−12:00 h; 12:00 h−16:00 h; 16:00 h−20:00 h; 20:00 h−24:00 h). Las agrupaciones se han realizado de acuerdo a los niveles de glucosa registrados, mediante diferencias estadísticamente significativas encontradas.

Resultados: Se han observado diferencias significativas (P-value <.05) y dependencias entre los perfiles de glucosa clasificados en función de las variables independientes día de la semana y franja horaria, siendo las relaciones encontradas distintas para cada paciente, demostrando la necesidad de hacer un estudio individualizado.

Conclusiones: Los resultados obtenidos facilitarán la modelización matemática de la glucosa, y podrán utilizarse para la creación de un clasificador individualizado para cada paciente, que clasifique los perfiles de glucosa en función de las variables día de la semana y franja horaria. Utilizando este clasificador, se podrán predecir los valores de glucosa del paciente conociendo en que día de la semana se encuentra y en que franja horaria, obteniendo modelos más precisos. También el profesional de la salud podrá mejorar los hábitos y terapias de los pacientes.

Introduction

The collection of methods known as data mining provides technical and methodological solutions to solve problems of medical data analysis and prediction modeling. Data mining is the process of selecting, exploring, and modeling large amounts of data to find and uncover unknown patterns or relationships that provide a clear and useful result.[1] It is a field of science that has developed rapidly in recent years and helps explain data and gain knowledge about it.[2] Decision trees are one of the most powerful statistical classification techniques used in data mining.[3–8] A decision tree is a clear and concise way to review and decide on the potential relationships between data, identifying groups or segments of interest between them.

The purpose of this study was to identify, through the construction of decision trees, glucose profiles classified into groups obtained using as variables the day of the week (Monday; Tuesday; Wednesday; Thursday; Friday; Saturday; Sunday) and the time slot, defined as the division of glucose values into four hour sections (00:00—04:00 h; 04:00—08:00 h; 08:00—12:00 h; 12:00—16:00 h; 16:00—20:00 h; 20:00—24:00 h).

The rest of the article is organized as follows. The following section describes the data used and the technique used for classifying glucose levels. The experimental work and results are shown in the ''Results'' section. The results and conclusions are discussed in the ''Discussion'' section.

Material and methods

Patients

A retrospective study was made of ten patients with type 1 diabetes mellitus. Measurements were recorded every five minutes using Guardian Real Time continuous glucose monitoring (CGM) sensors and Minimed insulin pumps (Medtronic). The carbohydrate estimates by patients trained in the diet by portions process were also recorded. Measurements were not necessarily taken on consecutive days or on the same days for each patient. Only the four-hour intervals with at least 46 values are considered. Table 1 in Appendix A. Supplementary material, shows the characterization of patients with information on sex, age, weight, HbA1c measured in the three months prior to the study, and the time elapsed since diagnosis of the disease and on treatment with the insulin pump. In addition, the same table shows for each patient the mean glucose value, the standard deviation, and the percentages of the time in which the patient has glucose levels below 70 mg/dL, above 250 mg/dL and the time in the range [70,180] mg/dL.

Statistical methods

Decision trees are a data mining technique that explores data to extract information hidden in them. The objective of decision tree construction is to create a model to predict the value of a dependent/target variable from the independent/predictive variables considered. The decision tree has three types of nodes, namely the root node, the internal nodes, and the end nodes, each representing a class characterized by the statistical values of the target variable, and the categories of the predictor variables contained in each node. Each path in construction of the decision tree is associated with a decision rule established by the algorithm itself. Thus, according to the established rules, the data set is recursively divided into separate subsets of smaller data (divisive algorithm). One of the most commonly used algorithms is the CHi-square Automatic Interaction Detection (CHAID) algorithm,[9] used in our study with the predictive analysis software IBM SPSS v.21.[10] This algorithm recursively divides data by a response/target variable using multiple divisions between the different input/predictor variables. A division must reach a threshold significance level between the nominal values of the target variable and the branches, or the node is not divided. The search ends when no more branches may be gathered or there are no significant divisions. The last division is chosen as the solution. It should be noted that the last division does not have to be the most significant division examined, i.e. there could be another more significant clustering different from that shown in the table, although this is an intrinsic property of the algorithm used.

This study used the glucose levels of each patient as the target variable, the day of the week and the time slot as predictor variables, and a 95% confidence level (α = 0.05) as significance threshold. The Snedecor's F distribution was used as division criterion and the Bonferroni adjustment was used for the number of categorical values of the input variable, thus mitigating bias towards inputs with many values.

Results

An individualized study was performed for each patient. To construct the decision trees of each patient, a maximum tree depth of three and minimum numbers of cases of 100 in the parent node and 50 in the child node were selected. The final tree depth, number of nodes, and number of end nodes obtained for each patient are shown in Table 2 of the Appendix A, Supplementary material. The first predictor used in tree construction was the day of the week, and the second predictor was time slot.

Overall, there are seven categories for the day of the week variable and six categories for the time slot variable. The categories of the variables are represented with letters and numbers, as shown in Appendix A Table 3, Supplementary material. Table 4 in Appendix A, Supplementary material represents the groups obtained for each patient. Clusters per day appear in a first level. For example, four different blood glucose patterns are seen in patient 1: one for Tuesday, another for Thursday, another for Friday, and a final pattern for Monday-Wednesday-Saturday-Sunday where the lowest mean glucose value is for Fridays (163.53 ± 58.06 mg/dL) and the highest value is for Tuesdays (179.01 ± 55.61 mg/dL). Clusters by time slot appear in the second level. The size of clusters at this level is separated by color, and it can be seen that in some patients there are clusters with a large number of slots. In the case of patient 1, we can say that he usually has two different behaviors on Friday, one for the [20:00—04:00 h] slot and another for the rest of the day. The same analysis can be done for all other patients.

Table 5 in Appendix A, Supplementary material includes information of glycemic control on the results of Table 4 in Appendix A, Supplementary material. To promote interpretation of results, information of patient 10 is only included. Data from all other patients are included as Supplementary material in the Appendix A.

The relationships found between independent variables and glucose values are also shown as pie charts. Charts were created using the circlize library,[11] with free statistical analysis software R version 3.5.2.[12]

Figures 1 and 2 in Appendix A, Supplementary material shows the pie charts with the results in Table 4 in Appendix A, Supplementary material. For each patient, a pie chart is created divided into seven segments, one for each day, and each segment is divided into another six segments corresponding to each slot. Each segment has a letter and a number identifying the day and the slot (Table 3 in Appendix A, Supplementary material) and the lines represent the relationships found between days and slots.

Discussion

Significant differences were found in the glucose profiles classified by the variables day of the week and time slot in each patient. Automatic classification found similarities between glucose profiles of different categories. Category is defined as all profiles corresponding to the same time slot of the same day of the week (e.g. Mondays from 00:00 to 04:00 h). A glucose profile is made up of the glucose values measured in that time slot.

The clusters obtained with the day of the week variable are heterogeneous, i.e. are made up of one, two, three, and up to four categories. The same applies to clusters with respect to the time slot variable. The most commonly repeated size of clusters for the day of week variable was two categories (in all patients) and the least common cluster sizes were three (patients 6 and 7) and four categories (patients 1 and 9). With regard to the time slot variable, the result was similar, though a greater number of similarities were found between the profiles, i.e. more clusters than for the day of the week variable and for a greater number of patients analyzed.

From the analysis of Tables 4 and 5 in Appendix A, Supplementary material, conclusions of clinical applicability may be drawn. It can be seen that patient 10 is more hyperglycemic on Saturdays than all other days (Table 4 in Appendix A, Supplementary material), and based on information in Table 5 of the Appendix A, Supplementary material for Saturdays, actions may be taken to reduce this behavior, such as changing his insulin regimen or diet.

An analysis of the pie charts in Figures 1 and 2 in Appendix A, Supplementary material shows that centers of the charts for patients 3 and 6 are line-free. This is because the clusters in these patients are made up of few elements, and there are differences between the glucose profiles obtained for both days and slots. The opposite is true for patients 4 and 9. Clusters are made up of several elements and chart centers show many lines. In this case, the glucose profiles obtained are similar, and glycemic control is greater in these patients (longer time in range).

However, there are charts (patients 3, 5, and 6) with few connections because the algorithm detects small differences between glucose values for the different days that it previously grouped in a single cluster, and no lines appear therefore in the charts. In other words, the algorithm performs a classification with greater tree depth (Table 2 of Appendix A, Supplementary material). For example, in patient 3 there is a first level that clusters Mondays and Tuesdays, and then clusters in a second level slots 0–3, 2–5, 1 and 4. If the chart was performed for this second level, a higher number of connections between categories would be seen. It is therefore necessary to establish the number of levels before the algorithm is applied.

Conclusions

The conclusions drawn from this study are as follows:

- Significant differences and dependencies were seen between the glucose profiles classified based on the variables day of the week and time slot.
- The clusters found were different for each patient, showing the need for individualized study.
- Tables 4 and 5 in Appendix A, Supplementary material allow for searching significant differences to correct and improve habits or therapies in patients, and to achieve more accurate models using machine learning and artificial intelligence techniques.
- The results obtained suggest that the techniques applied can facilitate mathematical modeling of glucose, and can be used to create an individualized classifier for each patient that classifies glucose profiles based on the day of the week and time slot variables. Use of this classifier will allow for predicting glucose levels of patients knowing in what day and in what time slot he/she is, obtaining more accurate models.

Funding

This study was partially funded by:

Grants from the Eugenio Rodríguez Pascual Foundation 2019: Development of adaptive and bioinspired systems for blood glucose control with continuous subcutaneous insulin infusion sets and continuous glucose monitors.

Project B2017/BMD3773 (GenObIA-CM) and Y2018/NMT-4668 (Micro-Stres - MAP-CM), financed by the Community of Madrid and co-financed with the EU Structural Funds.

Project Y2018/NMT-4668 (Micro-Stres - MAP-CM), financed by the Community of Madrid and co-financed with the EU Structural Funds.

Grant RTI2018-095180-B-I00 from the Spanish Ministry of Economy and Competitiveness.

Authorship

Marta Botella, José Ignacio Hidalgo, and Oscar Garnica designed the study and participated in writing and review of the manuscript.

Sergio Contador Pachón scheduled data processing and clustering techniques and participated in the writing of the article.

Jose Manuel Velasco produced the clustering charts and participated in the writing of the article.

Esther Maqueda participated in review of the article.

Aranzazu Aramendi and Remedios Martínez performed data collection and patient training.

Conflicts of interest

The authors state that they have no conflicts of interest.

Acknowledgments

Almudena Sánchez Sánchez participated in the discussions and in the origin of the work.

References

1. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform. 2008;77:81–97.
2. Witten IH, Frank E, Hall MA, Pal CJ. Data mining: practical machine learning tools and techniques. Morgan Kaufmann; 2013.
3. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med. 2005;34:113–27, http://dx.doi.org/10.1016/j.artmed.2004.07.002.
4. Chen HY, Chuang CH, Yang YJ, Wu TP. Exploring the risk factors of preterm birth using data mining. Expert Syst Appl. 2011;38:5384–7, http://dx.doi.org/10.1016/j.eswa.2010.10.017.
5. Chang CD, Wang CC, Jiang BC. Using data mining techniques for multi- diseases prediction modeling of hypertension and hyperlipidemia by common risk factors. Expert Syst Appl. 2011;38:5507–13, http://dx.doi.org/10.1016/j.eswa.2010.10.086.
6. Meng XH, Huang YX, Rao DP, Zhang Q, Liu Q. Comparison of three data mining models for predicting diabetes or predi-

abetes by risk factors. Kaohsiung J Med Sci. 2013;29:93–9, http://dx.doi.org/10.1016/j.kjms.2012.08.016.

7. Breault JL, Goodall CR, Fos PJ. Data mining a diabetic data warehouse. Artif Intell Med. 2002;26:37–54, http://dx.doi.org/10.1016/S0933-3657(02)00051-9.
8. Ramezankhani A, Pournik O, Shahrabi J, Khalili D, Azizi F, Ha- daegh F. Applying decision tree for identification of a low risk population for type 2 diabetes. Tehran lipid and glucose study. Diabetes Res Clin Pract. 2014;105:391–8, http://dx.doi.org/10.1016/j.diabres.2014.07.003.
9. Kass GV. An exploratory technique for investigating large quantities of categorical data. J R Stat Soc Ser C Appl Stat. 1980;29:119–27, http://dx.doi.org/10.2307/2986296.
10. Field A. Discovering statistics using IBM SPSS statistics. Sage; 2013.
11. Gu Z, Gu L, Eils R, Schlesner M, Brors B. Circlize implements and enhances circular visualization in R. Bioinformatics. 2014;30:2811–2.
12. The R Development Core Team. http://www.R-project.org/, 2013.