Compositional Data Analysis of Type 1 Diabetes Data

Lyvia Biagi[1,2], **Arthur Bertachi**[1,2], **Josep Antoni Martín-Fernández**[1], **Josep Vehí**[1,3]
[1] University of Girona, Spain
[2] Federal University of Technology - Paraná (UTFPR), Guarapuava, Brazil
[3] Centro de Investigación Biomédica en Red de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM), Spain
lyviar@utfpr.edu.br, abertachi@utfpr.edu.br, josepantoni.martin@udg.edu, josep.vehi@udg.edu

Abstract

Type 1 Diabetes (T1D) is a chronic disease characterized by a substantial reduction or no production of insulin by the pancreas. Subjects with T1D need to infuse insulin exogenously to avoid high blood glucose levels (hyperglycemia). However, if insulin is delivered exaggeratedly, subjects may experience low glucose levels (hypoglycemia). Both conditions are undesirable, and physicians prescribe individualized insulin therapy to cope with the disturbances that jeopardize glycemic control, e.g. meals, physical activity, stress, illness. This work presents a novel methodology for the individual classification of glucose profiles obtained from continuous glucose monitoring. Daily glucose profiles were discretized into time spent in distinct glucose ranges according to different glucose levels and formed a composition. Compositional data (CoDa) analysis was applied to the data and the discovery of groups of similar compositions was possible. Data was expressed in coordinates and k-means algorithm was applied to the coordinates to classify different patterns of days. This classification allowed the extraction of information of the data, which can assist physicians to adjust patients' treatments.

1 Introduction

Individuals with type 1 diabetes (T1D) need to inject insulin to assure blood glucose (BG) control. Insulin must be properly dosed either by multiple daily injections or continuous subcutaneous insulin infusion (CSII) in order to avoid both hypo- and hyperglycemia. These situations over time can result in different complications, such as seizures, coma, kidney, liver and heart damage, and even death [Agiostratidou *et al.*, 2017].

Even though present T1D technology combines continuous glucose monitoring (CGM) and CSII, achieving optimal glycemic control in T1D patients is still a hurdle due to the great intra-patient variability, which is resultant from characteristics of individual behavior and a complex metabolic system [Kovatchev and Cobelli, 2016].

The integration of CGM and CSII allows not only the visualization of glucose data in real time, which permits that the patient take measures in order to pursue glucose control, but also the acquisition of data, which supports the extraction of information that can help physicians to improve and adapt the current insulin therapy. Contreras *et al.* [2016] developed a tool for profiling BG dynamics of T1D patients based on different levels of BG. Even though the authors did not consider any insulin information in advance, they obtained groups of glucose profiles with different insulin requirements. The possibility of categorizing daily glycemic profiles according to the glycemic control can assist physicians to find different patterns of days and determine proper therapies to deal with day-to-day variations.

The analysis of time spent in distinct BG ranges defined by different levels during one day can be performed with Compositional Data Analysis (CoDA). The analysis of Compositional Data (CoDa) deals with vectors in the form of proportions to some whole [Aitchison, 1982]. The time spent in different glucose ranges are relative contributions to the 24-h time budget, and therefore, are compositional data, which have important characteristics that must be considered. The aim of this paper is to present a proof of concept of the feasibility of the usage of CoDA for the categorization of glucose profiles in patients with T1D.

2 Compositional Data Analysis of Glucose Time Series

Considering a compositional vector $\mathbf{x} = [x_1, x_2, \ldots, x_D]$, where D is the number of parts, $x_1, x_2, \ldots, x_D$ are positive components and $\sum_{i=1}^{D} x_i = C$, where C is a closure constant. The set of real positive vectors closed to a constant C is called simplex (S^D) [Aitchison, 1982]. As data carries only relative information, the time spent in each glucose range could be measured either as percentages of the day or minutes, or hours. The relevant information of a composition is contained in the ratios between its components.

This work considers a dataset obtained from one T1D patient who wore for approximately eight weeks the Paradigm Veo system with second generation of the Enlite CGM sensor (Medtronic Minimed, Northridge, CA, USA). Glucose data obtained from the CGM was used to validate the proposal. Insulin data obtained from the pump was also considered during the preprocessing of the data and analysis of results. The CGM system used by the patient recorded BG measurements every five minutes, thus, a complete day is supposed to have 288 samples. For several reasons, such as CGM or insulin pump malfunction during periods of the days, the quantity of samples per day, starting at 00:00 and ending at 23:55 is sometimes non-uniform, due to the missing values.

Data was preprocessed as follows: it was decided to consider days as valid only if each six-hour period contained at least 70% of the possible values for both glucose and insulin data. After that, valid days starting at 00:00 and ending were selected for analysis.

The 24-h glucose time series is analyzed considering five glucose ranges defined according to the standardized clinical levels of hypo- and hyperglycemia described in Agiostratidou *et al.* [2017]:

- Hypoglycemia
  Level 1: 54 mg/dL $\leq$ BG $<$ 70 mg/dL
  Level 2: BG $<$ 54 mg/dL
- Hyperglycemia
  Level 1: 180 mg/dL $<$ BG $\leq$ 250 mg/dL
  Level 2: BG $>$ 250 mg/dL

Thus, the glucose ranges considered were: <54 mg/dL, 54-70 mg/dL, 70-180 mg/dL, 180-250 mg/dL, >250 mg/dL, obtaining the composition $\mathbf{x}$ = (G_54, G_54_70, G_70_180, G_180_250, G_250). Missing values were assumed to be evenly distributed between the existing ranges of the day in analysis. The 24-h glucose profile was split into time spent in the five aforementioned glucose ranges. Time spent in different ranges are relative contributions to the 24-h glucose profile, codependent, and therefore, should be analyzed as CoDa.

2.1 Treatment of zeros

The log-ratio methodology, which is the basis of CoDA, must be preceded by a proper treatment of zero values. That is because both operations, logarithms and ratios, require non-zero elements in the data matrix. Martín-Fernández *et al.* [2011] describe different zero problems in their work. One example of a zero problem are rounded zeros, which cannot be observed because their value is is below some detection limit (DL).

The analysis of the zeros of the dataset was performed [Palarea-Albaladejo and Martín-Fernández, 2015]. Given the sampling frequency of the CGM (one sample every 5 minutes), and that the missing data was evenly distributed among the ranges of the data available for that day, the DL was set to 5 minutes.

The imputation of rounded zeros from continuous data can be done with the log-ratio Expectation-Maximization (EM) replacement [Palarea-Albaladejo and Martín-Fernández, 2015]. The values imputed by this method incorporate the information of the relative covariance structure.

2.2 Representation in Coordinates

CoDa must not be analyzed through standard multivariate analyses, which are designed for unconstrained multivariate data [Aitchison, 1982]. CoDa can be analyzed in the real space after the expression in coordinates, which allows the application of traditional statistical methods [Aitchison, 1986]. The *centred log-ratio* (clr) transformation projects S^D to the real space R^D. It was introduced in Aitchison [1986] as the logarithm of the ratio of each part over the geometric mean, and it is defined in (1). The *isometric log-ratio* (*ilr*) transformation expresses $\mathbf{x}$ in terms of its orthonormal log-ratio coordinates. It was introduced in Egozcue *et al.* [2003]. An *ilr* vector can be viewed as the coordinates of a composition with respect to an orthonormal basis $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}$ on the simplex [Pawlowsky-Glahn *et al.*, 2015], as described in (2).

$$\operatorname{clr}(\mathbf{x}) = \left[\ln\left(\frac{x_1}{g(\mathbf{x})}\right), \ln\left(\frac{x_2}{g(\mathbf{x})}\right), \ldots, \ln\left(\frac{x_D}{g(\mathbf{x})}\right)\right] \quad (1)$$

$$\operatorname{ilr}(\mathbf{x}) = \operatorname{clr}(\mathbf{x}) \cdot \mathbf{\Phi}^t \quad (2)$$

where $g(\mathbf{x})$ is the geometric mean of $\mathbf{x}$, $\mathbf{\Phi}$ is the $(D-1) \times D$-matrix whose i-th row is the vector $\operatorname{clr}(\mathbf{e}_i)$, for $i = 1, \ldots, D-1$.

2.3 Compositional biplot

A representation of any matrix by means of a 2-rank approximation is possible with a biplot [Gabriel, 1971]. The adaptation of the biplot for use with CoDa is a useful exploratory tool [Pawlowsky-Glahn *et al.*, 2015]. The quality of the $(D-1)$-dimensional *ilr*-transformed representation in a two-dimensional graph is expressed by the cummulative total variance of the biplot of the *clr*-transformed data. The discovery of potential clusters of compositions is possible with the *clr*-biplot [Pawlowsky-Glahn *et al.*, 2015; Palarea-Albaladejo *et al.*, 2012].

Figure 1 shows the biplot of days of the patient in analysis, obtained with CoDaPack [Comas-Cufí and Thió-Henestrosa, 2011]. The origin of the biplot represents the centre of the compositional dataset. Five vertices, each one related to the *clr*-coordinate of the part correspondent to the time spent in each glucose range are connected to the origin through five rays (in red). Each marker represents a single day. Markers close to a ray of determined *clr*-coordinate indicate that those days are characterized by high values in that *clr*-coordinate, i.e. it means that the individual spent relatively high time in the glucose range correspondent to that *clr*-coordinate, comparing to the geometric average of time spent in all glucose ranges. The cummulative total variance retained by the biplot is equal to 92.19%, the high value of the variance retained means that the biplot provides a good representation of the data in the real space. It is possible to infer the existence of groups.

According to Palarea-Albaladejo *et al.* [2012], it is inappropriate to apply the ordinary approach of clustering, due to the constraints of the simplex space. However, it is possible to use distance based clustering techniques in CoDA, considering that the Aitchison distance between compositions is equal to the Euclidean distance between the log-ratio coordinates. One way to obtain the log-ratio coordinates is through the SBP [Egozcue and Pawlowsky-Glahn, 2005].

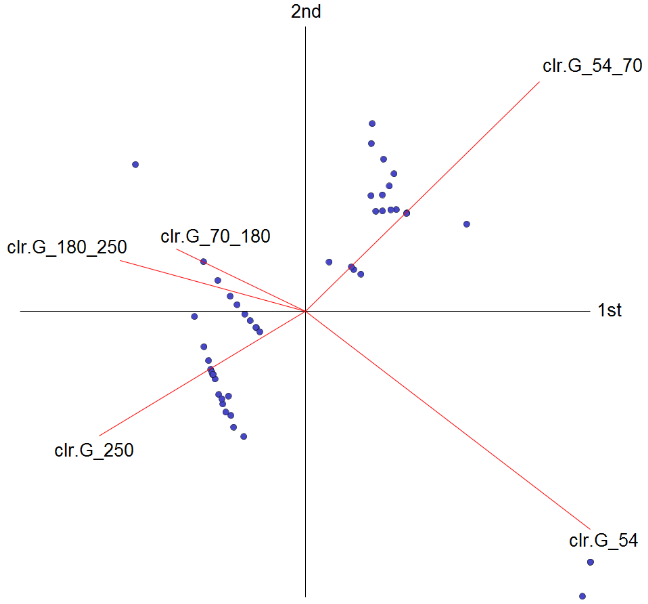

Figure 1: *clr*-biplot of days of one patient. Cummulative total variance retained = 92.19%.

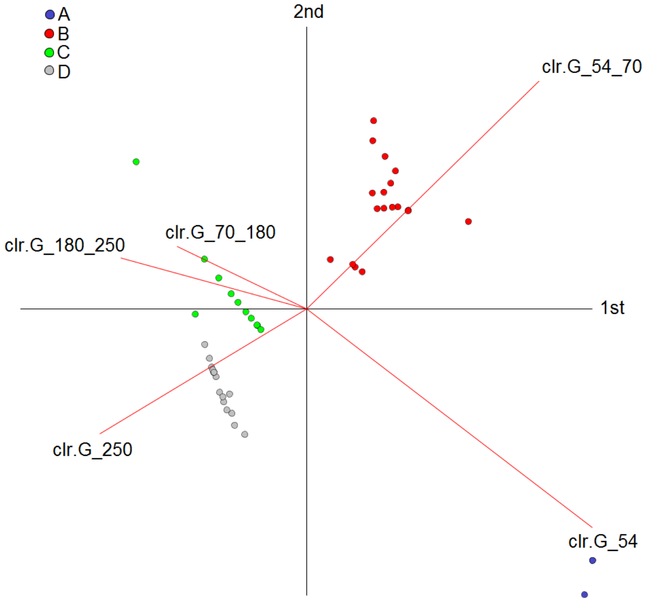

Figure 2: *clr*-biplot of days of one patient. Four groups obtained after clustering, indicated by different colors and the letters A, B, C and D.

2.4 Cluster Analysis

The log-ratio coordinates were obtained through the SBP of the data and k-means algorithm [Hartigan and Wong, 1979] was applied to the coordinates to check for different patterns of days. This algorithm is based on squared Euclidean distance and we considered with 25 random repetitions of the selection of initial centres.

K-means was tested for several numbers of groups. We analyzed the groups of days obtained per patient in terms of minimums and maximums of parts (times in different glucose ranges) and ratios (coordinates). The choice of number of groups took into account the representation obtained in the *clr*-biplot. The following measurements were also summarized per group: average blood glucose (Avg BG), BG variation (BGV), number of level 1 and 2 hypoglycemic events per day, where a hypoglycemic event is defined as three or more CGM readings under the referred level, average time of level 1 and 2 hypo- and hyperglycemia per day. The information regarding insulin therapy is also provided: total basal and bolus insulin, expressed in units of insulin (UI), number of bolus per day (# Bolus), carbohydrates intake (CHO), expressed in exchanges (ex, where 1 exchange is equivalent to 10 grams of CHO), relation between bolus insulin and carbohydrates (bolus:C) and time of pump suspension (min/day).

3 Results

Figure 2 shows the *clr*-biplot of the days of one patient. Four groups are represented. Days of group A are very close to the ray of variable G_54. This suggest that days of this group are characterized by high values in this variable, i.e., days of this group are characterized by relatively high values in the part corresponding to BG $<$ 54 mg/dL. Days of group B are very close to the ray of variable G_54_70. This suggests that days of this group are characterized by high values in the part corresponding to BG between 54-70 mg/dL. Days of group C are close to the rays of variables G_70_180 and G_180_250. This suggest that days of group C are characterized by relatively high values in the parts corresponding to BG between 70-180 mg/dL and 180-250 mg/dL. Days of group D are characterized by relatively high values in the parts corresponding to BG $>$250 mg/dL.

Figure 3 shows the compositional geometric mean barplot, which is useful as visualization tool for the analysis of grouped data. The compositional centre of each group of data is compared with the centre of the whole dataset. Thus, positive bars reflect relatively mean values of a part above the overall composition. Analogously, negative bars reflect relatively mean values of a part below the overall composition.

Days of group A are characterized by relatively high values in the parts corresponding to BG $<$54 mg/dL and between 54-70 mg/dL.

Days of group B are characterized by relatively high values in the parts corresponding to between 54-70 mg/dL and relatively low values in the part corresponding to BG $>$250 mg/dL. Days of group B are characterized by the existence of level 2 hypoglycemic events.

Days of group C are characterized by relatively low values in the parts corresponding to BG $<$54 mg/dL, between 54-70 mg/dL and $>$250 mg/dL. On days of group C, the individual spent relatively less time in the extreme glucose ranges (high and low).

Days of group D are also characterized by relatively low values in the parts corresponding to BG $<$54 mg/dL and between 60-70 mg/dL, however, unlike days of groups C, days of group D are characterized by relatively high values in the part corresponding to BG $>$250 mg/dL. On days of group D,

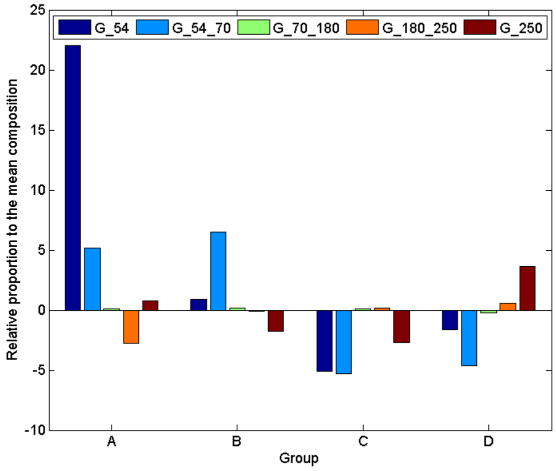

Figure 3: Compositional geometric mean barplot of groups of days of one patient.

the individual spent relatively more time in the highest glucose range and less time in low glucose ranges.

Table 1 show some clinically relevant outcomes considering the classification of the days. As suggested by the biplot of Figure 2, days of group A are those related to the occurrence of level 2 hypoglycemic events. Days of this group also presented fewer boluses than days of other groups and also the highest time with the insulin delivery suspended by the pump, however, the highest bolus:C ratio is presented for days of this group.

Days of group B are characterized by the existence of level 1 hypoglycemic events, but without the occurrence of level 2 hypoglycemic events. In these days, patients delivered more boluses when compared with group A and also consumed more CHO.

Days of group C present the lowest BGV between all groups and there is no incidence of hypoglycemic events nor level 2 hyperglycemic events. However, days of this group are also characterized by the occurrence of level 1 hyperglycemia.

Even achieving the highest Basal, Bolus, # Bolus between all groups, days of group D present the highest Avg BG and in days of this group, individuals also spent more time in hyperglycemia. Days of this group are characterized by relatively high values in the parts corresponding to BG >250 mg/dL, as showed both by the biplot of Figure 2 and by the barplot of Figure 3. It is very likely that patients's behavior during these days require an adjustment on his/her insulin therapy, such as increasing the total amount of basal or increasing bolus:C ratio.

4 Conclusion

The aim of this paper is to present the feasibility of the usage of CoDa for the analysis of glucose profiles of patients with T1D. This methodology was based on the discretization of daily glucose profiles considering different ranges. The time spent in different glucose ranges are relative contributions to the 24-h period, and therefore, should be analyzed as CoDa.

Table 1: Summary of clinically relevant outcomes for each group.

	A	**B**	**C**	**D**
# days per group	3	18	11	16
Avg BG (mg/dl)	150.28	140.40	151.33	181.44
BGV (mg/dl)	40.52	40.50	36.34	49.86
# Level 1 Hypo events (events/day)	0.67	0.89	0.00	0.00
Time Level 1 Hypoglycemia (min/day)	23.11	33.53	0.00	0.00
# Level 2 Hypo events (events/day)	0.33	0.00	0.00	0.00
Time Level 2 Hypoglycemia (min/day)	14.06	0.00	0.00	0.00
Time Level 1 Hyperglycemia (min/day)	166.44	296.72	363.72	434.32
Time Level 2 Hyperglycemia (min/day)	107.39	18.61	0.65	199.55
Basal (UI)	18.50	19.46	19.91	19.95
Bolus (UI)	21.27	20.65	20.22	24.33
# Bolus	4.33	7.00	6.64	7.25
CHO (ex)	12.17	14.43	13.23	13.71
bolus:C (U/ex)	1.93	1.47	1.55	1.81
Time Pump Suspension (min/day)	98.33	67.50	49.55	30.63

Although the proposed methodology has been applied in a dataset collected from a single T1D patient in a retrospective way, this novel approach was able to classify the days into different groups that reflect patient's behavior. With this information, physicians may use this classification as an analysis tool to adjust insulin therapy for each group of days to improve overall glycemic control. The method provides tools for personalized analysis of the data, making possible the comparison of characteristics of groups of days with the whole dataset. Even though more effort is required for the categorization of days in real time, the analysis allows extraction of information of different groups of days, and consequently, the inference of correction measures that should be taken for each specific group.

Acknowledgments

This project has been partially supported by the Spanish Government MINECO through Grant DPI-2016-78831-C2-2-R and the National Council of Technological and Scientific Development, CNPq Brazil through Grants 202050/2015-7 and 207688/2014-1. The authors thank the Spanish Consortium on Artificial Pancreas and Diabetes Technology for sharing their database.

References

[Agiostratidou *et al.*, 2017] Gina Agiostratidou, Henry Anhalt, Dana Ball, Lawrence Blonde, Evgenia Gourgari, Karen N. Harriman, Aaron J. Kowalski, Paul Madden, Alicia H. McAuliffe-Fogarty, Molly McElwee-Malloy, Anne Peters, Sripriya Raman, Kent Reifschneider, Karen Rubin, and Stuart A. Weinzimer. Standardizing clinically meaningful outcome measures beyond hba1c for type 1 diabetes: A consensus report of the american association of clinical endocrinologists, the american association of diabetes educators, the american diabetes association, the end…. *Diabetes Care*, 40(12):1622–1630, 2017.

[Aitchison, 1982] J Aitchison. The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society. Series B (Methodological) J. R. Statist. Soc. B*, 44(2):139–177, 1982.

[Aitchison, 1986] J. Aitchison. *The statistical analysis of compositional data*. Monographs on statistics and applied probability. Chapman and Hall, 1986. Reprinted 2003 with additional material by The Blackburn Press.

[Comas-Cufí and Thió-Henestrosa, 2011] Marc Comas-Cufí and Santi Thió-Henestrosa. CoDaPack 2.0: a stand-alone, multi-platform compositional software. In J.J. Egozcue, R. Tolosana-Delgado, and M.I Ortego, editors, *CoDaWork'11: 4th International Workshop on Compositional Data Analysis*, Sant Feliu de Guíxols, 2011.

[Contreras *et al.*, 2016] Iván Contreras, Carmen Quirós, Marga Giménez, Ignacio Conget, and Josep Vehi. Profiling intra-patient type I diabetes behaviors. *Computer Methods and Programs in Biomedicine*, 136:131–141, 2016.

[Egozcue and Pawlowsky-Glahn, 2005] J J Egozcue and V Pawlowsky-Glahn. Groups of Parts and Their Balances in Compositional Data Analysis. *Mathematical Geology*, 37(7), 2005.

[Egozcue *et al.*, 2003] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300, Apr 2003.

[Gabriel, 1971] K R Gabriel. The Biplot Graphic Display of Matrices with Application to Principal Component Analysis. *Source: Biometrika Biometrika Trust*, 58(3):453–467, 1971.

[Hartigan and Wong, 1979] J. A. Hartigan and M. A. Wong. A k-means clustering algorithm. *JSTOR: Applied Statistics*, 28(1):100–108, 1979.

[Kovatchev and Cobelli, 2016] Boris Kovatchev and Claudio Cobelli. Glucose variability: Timing, risk analysis, and relationship to hypoglycemia in diabetes. *Diabetes Care*, 39(4):502–510, 2016.

[Martín-Fernández *et al.*, 2011] Josep Antoni Martín-Fernández, Javier Palarea-Albaladejo, and Ricardo Antonio Olea. Dealing with Zeros. *Compositional Data Analysis: Theory and Applications*, pages 43–58, 2011.

[Palarea-Albaladejo and Martín-Fernández, 2015] J Palarea-Albaladejo and Josep Antoni Martín-Fernández. zCompositions – R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems*, 143:85–96, 2015.

[Palarea-Albaladejo *et al.*, 2012] Javier Palarea-Albaladejo, Josep Antoni Martín-Fernández, and Jesús A. Soto. Dealing with distances and transformations for fuzzy c-means clustering of compositional data. *Journal of Classification*, 29(2):144–169, Jul 2012.

[Pawlowsky-Glahn *et al.*, 2015] Vera Pawlowsky-Glahn, Juan José Egozcue, and Raimon Tolosana-Delgado. *Modeling and Analysis of Compositional Data*. John Wiley & Sons, 2015.