Data in Brief 55 (2024) 110559

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

HUPA-UCM diabetes dataset

J. Ignacio Hidalgo[a,*], Jorge Alvarado[b], Marta Botella[c], Aranzazu Aramendi[c], J. Manuel Velasco[a], Oscar Garnica[a]

[a] Universidad Complutense de Madrid, Profesor José García Santesmases 9, Madrid, Spain
[b] Universidad de Extremadura, Avda. de Elvas s/n, Badajoz, Spain
[c] Hospital Universitario Príncipe de Asturias, Alcalá de Henares, Spain

ARTICLE INFO

Article history:
Received 29 April 2024
Accepted 21 May 2024
Available online 27 May 2024

Dataset link: HUPA-UCM Diabetes Dataset (Original data)

Keywords:
Diabetes
T1DM
Glucose prediction
Machine learning

ABSTRACT

This dataset provides a collection of Continuous Glucose Monitoring (CGM) data, insulin dose administration, meal ingestion counted in carbohydrate grams, steps, calories burned, heart rate, and sleep quality and quantity assessment ac- quired from 25 people with type 1 diabetes mellitus (T1DM). CGM data was acquired by FreeStyle Libre 2 CGMs, and Fitbit Ionic smartwatches were used to obtain steps, calories, heart rate, and sleep data for at least 14 days. This dataset could be utilized to obtain glucose prediction models, hypoglycemia and hyperglycemia prediction models, and research on the relationships among sleep, CGM values, and the rest of the mentioned variables. This dataset could be used directly from the preprocessed version or customized from raw data. The data set has been used previously with different machine learning algorithms to predict glucose values, hypo, and hyperglycemia and to analyze influences among the features and the quality and quantity of sleep in people with T1DM.

* Corresponding author.
E-mail address: hidalgo@dacya.ucm.es (J.I. Hidalgo).
Social media: @at.evoHidalgo (J.I. Hidalgo), @at.AbsysGroup (J.I. Hidalgo; J. Alvarado; M. Botella; A. Aramendi; J.M. Velasco; O. Garnica)

https://doi.org/10.1016/j.dib.2024.110559

Specifications Table

Subject	Health and medical sciences
Specific subject area	Endocrinology, Diabetes and Metabolism
Type of data	Raw, Processed
Data collection	Medical study - 25 real patients with T1DM - Raw data obtained from GCM and activity smartwatches - Data preprocessed and curated
Data source location	- Institution: Hospital Principe de Asturias - City: Alcalá de Henares - Country: Spain
Data accessibility	Repository name: HUPA-UCM Diabetes Dataset Direct URL to data: https://data.mendeley.com/datasets/3hbcscwz44/1 Data identification number: 10.17632/3hbcscwz44.1 Instructions for accessing these data: Not applicable

1. Value of the Data

- This dataset provides data from a medical study of 25 people with type 1 diabetes mellitus (T1DM). The data includes information from CGMs, personal notes, insulin infusion systems, and Fitbit Ionic smartwatches.
- The data collected can be used for research involving data from people with diabetes. Working with simulated data in the medical field is usually necessary because obtaining real data is difficult due to the need for consent from the patients and the medical team.
- The dataset can be used to develop predictive models of glucose, the development of artificial pancreas, and alert systems for hypoglycemic and hyperglycemic events, among others that allow better disease control for people with diabetes.

2. Background

People with Diabetes Mellitus should manage their glucose values with the primary objective of keeping it in a healthy range. Glycemic control is important to prevent glucose from reaching dangerous levels and to prevent hypoglycemic and hyperglycemic events.

Thanks to advances in machine learning (ML) and artificial intelligence, interest in developing the perfect artificial pancreas has increased. Large datasets are vital to creating predictive models with high accuracy. The difficulty of research in the medical field is that some researchers can only work with simulated data because obtaining real data is only sometimes possible due to requirements such as ethical committee approval, patient consent, etc.

3. Data Description

The dataset data format is organized into CSV files for each patient. A 5-min interval has been used for the records. Each file contains the following columns separated by semicolons:

- **time:** time of data recording (yyyy-MM-dd'T'HH:mm:ss format)
- **glucose:** blood glucose value (mg/dL)
- **calories:** calories burned in the time Interval
- **heart_rate:** heart frequency
- **steps:** steps taken in the time interval
- **basal_rate:** basal insulin infusions in the time interval
- **bolus_volume_delivered:** insulin bolus injections in the time interval
- **carb_input:** servings of carbohydrates ingested in the time interval (1 serving = 10 g)

4. Experimental Design, Materials and Methods

4.1. Data

The dataset presented in this paper contains the real data collected from 25 patients with T1DM. Interstitial glucose, insulin infusions and injections, carbohydrate intakes, heart rate, calories, steps, and sleep data were collected for each patient. Glucose data were collected using Abbott FreeStyle Libre GCMs every 15 min. Information about insulin and carbohydrate intakes was recorded using two different methods, depending on the insulin administration mode. Participants wearing an automatic continuous insulin infusion system (CIIS) obtain this information directly from the device (Medtronic or Roche systems). Participants under multiple doses of insulin (MDI) therapy recorded information about basal insulin, insulin boluses, and carbohydrate intakes using a mobile application. Heart rate, calories, steps, and sleep information were obtained using Fitbit activity smartwatches. The Fitbit device was used for passive physiological collection during the subjects' daily living. Participants were trained to wear the Fitbit device during the exercise, and inconsistencies in the data were curated in the preprocessing step.

The data was recorded in free-living conditions. After a training session with the clinician staff, the participants estimated the carbohydrate count of meals. We proceeded in this way to replicate the real free-living conditions. The error in the estimation should be taken as a limitation of the dataset.

The clinical characteristics of the participants taking part in the study are: sex (52.0 % female), age (39.23 +/− 11.84), weight (69.06 +/− 14.12) kg, height (169.04 +/− 10.41 cm), mean HBA1c (glycosylated hemoglobin) 7.37 (+/− 0.82), years of illness 17.8 (+/− 10.5), treatment: 56.0 % continuous subcutaneous insulin infusion (CSII) and 44.0 % multiple dose insulin (MDI). Table 1 shows the characteristics of the patients in the study, and Fig. 1 shows the percentage of time in range for each patient.

4.2. Data preprocessing

The data have been collected using different sensors, as described in the previous section. The variability of the sensors means that the format of the raw files obtained and the data collection intervals are different for each variable. For this reason, the data has been processed and cleaned.

The glUCModel tool developed by the Adaptive and Bioinspired Systems (ABSys) research group of the Complutense University of Madrid has been used for data preprocessing. This application allows the upload of data from different sensors, the classification of the data by patients, and the visualization of the data, graphs, etc.

The preprocessing for each of the variables is detailed below:

- **Interstitial glucose values:** data were rounded to the nearest 5 min because recordings do not occur at exact times. The data were then subsampled to convert the time series with a frequency to 15-min intervals. Finally, linear interpolation was applied to fill the data gaps and obtain records every 5 min.
- **Insulin infusions and injections:** data from insulin boluses were resampled at 5-min intervals by the sum of all interval values. Basal insulin data were calculated by dividing by 288 (24 h $\times$ 12 records per hour) each basal insulin infusion to obtain records at 5-min intervals. In the case of overlapping basal insulin intervals, the values of records that coincided in time were summed. Finally, gaps in the time series without data were set to zero.
- **Carbohydrate intakes:** the carbohydrate data were resampled to obtain records every 5 min by summing all values within the interval. Values recorded in grams were converted to servings by dividing by 10 (1 serving = 10 g). A zero value was set for the gaps without data.

Table 1
Clinical characterization of each patient: ID, Gender, HbA1c, Age, Years of disease, Weight, Height, and Treatment (MDI: Multiple doses of Insulin; CSII: Continuous Subcutaneous Insulin Infusion).

Measure	HUPA0001P	HUPA0002P	HUPA0003P	HUPA0004P	HUPA0005P	HUPA0006P	HUPA0007P	HUPA0009P	HUPA0010P
Gender	Female	Male	Male	Male	Female	Male	Male	Female	Female
HbAc [%]	8.2	7.1	7.3	7.8	6.9	7.8	6.6	7.6	6.0
Age [years]	56.3	48.6	43.4	41.2	41.9	22.1	37.6	41.2	41.9
DX time [years]	15.5	36.5	12.5	8.5	39.5	13.5	10.1	30.7	15.2
Weight [kg]	59.0	82.4	62.0	88.0	58.5	71.0	102.6	64.0	51.0
Height [cm]	161	186	182	180	161	170	183	165	164
CSII/MDI	CSII	CSII	CSII	CSII	CSII	CSII	CSII	CSII	CSII
Measure	HUPA0011P	HUPA0014P	HUPA0015P	HUPA0016P	HUPA0017P	HUPA0018P	HUPA0019P	HUPA0020P	HUPA0021P
Gender	Female	Female	Female	Female	Female	Female	Male	Male	Female
HbAc [%]	7.8	8.5	6.4	6.5	8.2	7.2	7.1	9.7	7.5
Age [years]	35.0	50.0	43.1	29.9	26.3	32.3	18.0	45.7	48.6
DX time [years]	27.3	12.9	11.2	20.1	24.2	25.6	7.6	13.5	2.2
Weight [kg]	56.0	61.0	58.6	64.9	61.8	57.2	69.7	71.6	57.0
Height [cm]	153	155	162	157	167	167	168	168	153
CSII/MDI	CSII	MDI	MDI	CSII	MDI	CSII	MDI	MDI	MDI
Measure	HUPA0022P	HUPA0023P	HUPA0024P	HUPA0025P	HUPA0026P	HUPA0027P	HUPA0028P		
Gender	Male	Male	Male	Male	Female	Male	Female		
HbAc [%]	6.7	7.7	8.3	7.0	7.2	7.0	6.1		
Age [years]	59.6	22.9	47.9	38.1	61.8	26.4	21.2		
DX time [years]	14.6	0.8	35.9	20.3	21.5	23.7	2.0		
Weight [kg]	77.6	55.5	80.5	104.8	80.0	76.0	56.0		
Height [cm]	179	173	174	188	165	185	160		
CSII/MDI	CSII	MDI	MDI	CSII	MDI	MDI	MDI		

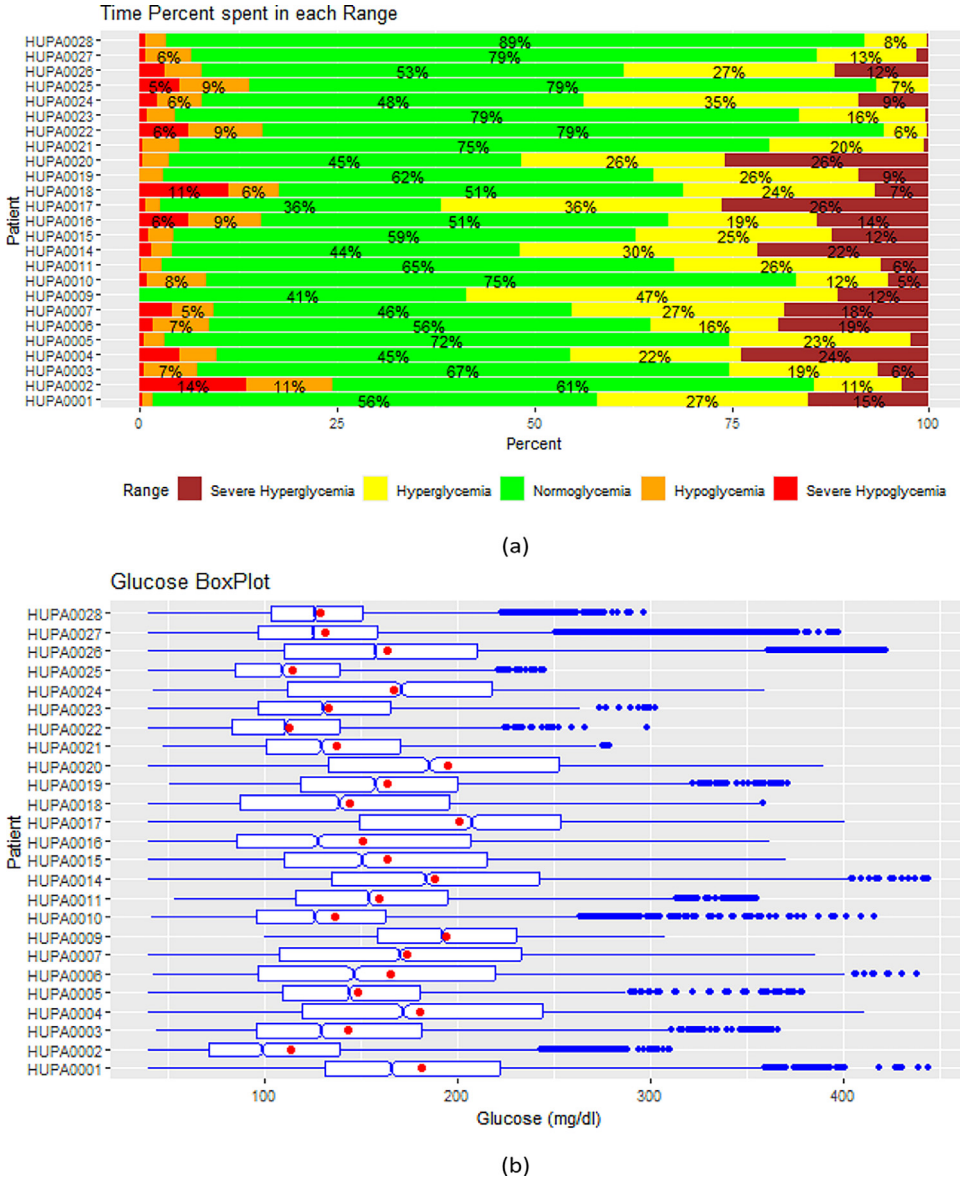

Fig. 1. Figure (a) shows the percentage of time the patient has a very low glucose level (< 54 mg/dL), low ([54, 70) mg/dL), in range ([70, 180] mg/dL), high ((180, 250] mg/dL), and very high (> 250 mg/dL). Figure (b) shows the interquartile ranges of glucose. The mean value is shown with a red dot. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

- **Heart rate:** the data were resampled and rounded to the nearest 5 min because values were recorded without following a regular time interval. A linear interpolation was applied to cover the time series's missing values and obtain records at 5-min intervals.
- **Calories:** calories values were recorded by the sensor at one-minute intervals. Resampling has been applied for processing, consisting of adding every five records by one to obtain data at 5-min intervals. Values with no data were set to zero.

- **Steps:** steps have been recorded in 1-min intervals by the sensor. Therefore, a resampling consisting of summing every five values has been applied to obtain records at 5-min intervals. Gaps in the time series with no data were set to zero.

To generate the files of this dataset, we selected the continuous range of records that the patients had both glucose and heart rate values, discarding the beginning and end of the data that did not contain valid values of these two variables. For the rest of the variables, the empty or invalid values were set to 0 to fill the gaps in the data.

4.3. Previous studies

This dataset has been used in the past in different studies. The authors developed different glucose prediction models using this dataset in [1] and [2]. In [3], data were used to discover Difference Equations with Structured Grammatical Evolution for Postprandial Glycaemia Prediction. Data was also applied in [4] to predict hypoglycemia events by combining wavelet transform and convolutional neural. This dataset has been used to test a hardware implementation of a low-power LSTM neural network wearable medical device designed to predict blood glucose at a 30-min horizon. The hardware was implemented on a Xilinx Virtex-7 FPGA [5]. We can also find a study on the influence of sleep quality and quantity on glycemic control in adults with type 1 diabetes using this dataset [6].

Limitations

This data is limited by the error in estimating carbohydrates. The Fitbit device collects physiological data at different rates, and inconsistencies in the data were curated in the preprocessing step. CGM data was collected every 15 min, while Fitbit data was preprocessed at five-minute intervals.

Ethics Statement

The data were collected with the collaboration of the Hospital Príncipe de Asturias in Alcalá de Henares (Spain) and Universidad Complutense de Madrid (Spain). The study was approved by the ethical committee of Hospital Universitario Príncipe de Asturias de Madrid, and patients signed an informed consent form. Protocol Number: EC/11/2018, Date of approval: December 12th 2018.

Data Availability

HUPA-UCM Diabetes Dataset (Original data) (Mendeley Data).

CRediT Author Statement

J. Ignacio Hidalgo: Conceptualization, Methodology, Resources, Funding acquisition, Writing – review & editing; **Jorge Alvarado:** Data curation, Writing – original draft; **Marta Botella:** Supervision, Conceptualization; **Aranzazu Aramendi:** Resources; **J. Manuel Velasco:** Visualization, Software, Investigation; **Oscar Garnica:** Data curation, Funding acquisition.

Acknowledgments

This work has been supported by the Spanish Ministry of Economy and Competitiveness under projects AEI grants PID2021-125549OB-I00 and PDC2022-133429-I00. AEI/10.13039/501100011033 and of the European Union Resilience and Recovery Mechanism. Funded by the European Union - NextGenerationEU.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] L. Ingelse, J.I. Hidalgo, J.M. Colmenar, N. Lourenço, A. Fonseca, A comparison of representations in grammar-guided genetic programming in the context of glucose prediction in people with diabetes, 2023 available at Research Square, doi:10.21203/rs.3.rs-3596625/v1.

[2] F. Tena, O. Garnica, J. Lanchares, J.I. Hidalgo, Ensemble models of cutting-edge deep neural networks for blood glucose prediction in patients with diabetes, Sensors 21 (21) (2021) 7090.

[3] D. Parra, D. Joedicke, J.M. Velasco, G. Kronberger, J.I. Hidalgo, Learning difference equations with structured grammatical evolution for postprandial glycaemia prediction, IEEE J. Biomed. Health Inf. 28 (5) (2024) 3067–3078.

[4] J. Alvarado, J.M. Velasco, F. Chavez, F. Fernández-de Vega, J.I. Hidalgo, Combining wavelet transform with convolutional neural networks for hypoglycemia events prediction from CGM data, Chemometr. Intell. Lab. Syst. 243 (2023) 105017.

[5] F. Tena, O. Garnica, J.L. Davila, J.I. Hidalgo, An lstm-based neural network wearable system for blood glucose prediction in people with diabetes, IEEE J. Biomed. Health Inf. (2023) 1–12, doi:10.1109/JBHI.2023.3300511.

[6] M. Botella-Serrano, J.M. Velasco, A. Sánchez-Sánchez, O. Garnica, J.I. Hidalgo, Evaluating the influence of sleep quality and quantity on glycemic control in adults with type 1 diabetes, Front. Endocrinol. 14 (2023), doi:10.3389/fendo.2023.998881.