scientific data

OPEN

DATA DESCRIPTOR

Chinese diabetes datasets for data-driven machine learning

Qinpei Zhao[1,2,6], Jinhao Zhu[1,6], Xuan Shen[3,6], Chuwen Lin[3,6], Yinjia Zhang[4], Yuxiang Liang[1], Baige Cao[3], Jiangfeng Li[1,7] ✉, Xiang Liu[5], Weixiong Rao[1,7] ✉ & Congrong Wang[3,7] ✉

Data of the diabetes mellitus patients is essential in the study of diabetes management, especially when employing the data-driven machine learning methods into the management. To promote and facilitate the research in diabetes management, we have developed the *ShanghaiT1DM* and *ShanghaiT2DM* Datasets and made them publicly available for research purposes. This paper describes the datasets, which was acquired on Type 1 (n = 12) and Type 2 (n = 100) diabetic patients in Shanghai, China. The acquisition has been made in real-life conditions. The datasets contain the clinical characteristics, laboratory measurements and medications of the patients. Moreover, the continuous glucose monitoring readings with 3 to 14 days as a period together with the daily dietary information are also provided. The datasets can contribute to the development of data-driven algorithms/models and diabetes monitoring/managing technologies.

Background & Summary

Diabetes is a chronic disease that could lead to cardiovascular disease, neuropathy, retinopathy, kidney failure and even mortality. Rapid socioeconomic changes and unhealthy lifestyle habits have led to the increasing prevalence of diabetes worldwide. Type 1 diabetes mellitus (T1DM) and Type 2 diabetes mellitus (T2DM) are the two main types of diabetes. T1DM is a chronic autoimmune disease resulting from destruction or damaging of the pancreatic beta cells[1]. T2DM is caused by insulin resistance and relative insulin deficiency[2]. T1DM accounts for only 5–10% of all diabetes worldwide, but varies geographically with the annual incidence of adult-onset T1DM about 1 per 100,000 in China[3], while T2DM is the most common subtype of diabetes, accounting for over 90% of all the diabetes worldwide and in China[3,4]. It is shown that good blood glucose (BG) control significantly reduces the development or progression of chronic complications in T1DM and T2DM[5–7]. Thus, BG measurement plays a key part in diabetes care, which allows patients to adjust their food intake, physical activity and medications with the help of physicians (clinicians)[8]. Self-monitoring of blood glucose (SMBG) is a measurement that uses blood to collect blood glucose information at many time points[9]. Recently, a continuous glucose monitoring (CGM) technology is used to continuously monitor the BG levels in more or less real time[10,11].

The use of CGM technology makes it possible to obtain a large amount of continuous BG data. However, there were relatively few publicly available BG datasets, as the data may have ethical restrictions and privacy concerns. There have been many studies[12,13] on the BG prediction using different datasets. A rigorous literature review[12] was conducted to develop a compact guide regarding machine learning methods on BG prediction in T1DM. The review included 55 papers from 2000 to 2018 and showed their subject, type of input, data source, input pre-processing methods, machine learning algorithms, prediction horizon and performance metrics. A systematical review[13] on the literature from 2014 to 2020 was performed to study the data-based algorithms and models using real data for BG and hypoglycaemia prediction in T1/T2DM. The existing datasets in T1/T2DM for the BG prediction have been listed in the review. However, the T2DM datasets are much less studied than the T1DM datasets, e.g., 6 of 63 publications included T2DM in the review[13]. For real data, the data size was relatively small. In the review[13], 27 papers (42.9%) present small samples (n < 10), 19 papers (30.2%) with small-medium samples (n = 11–50) and 17 papers (27%) with relatively large samples (n > 50). In another

[1]School of Software Engineering, Tongji University, Shanghai, China. [2]AIway Oy, Helsinki, Finland. [3]Department of Endocrinology & Metabolism, Shanghai Fourth People's Hospital, School of Medicine, Tongji University, Shanghai, China. [4]Department of Computer Science, School of Science, Aalto University, Helsinki, Finland. [5]Zhejiang Yugu Medical Technology Ltd, Zhejiang, China. [6]These authors contributed equally: Qinpei Zhao, Jinhao Zhu, Xuan Shen, Chuwen Lin. [7]These authors jointly supervised this work: Jiangfeng Li, Weixiong Rao, Congrong Wang. ✉e-mail: lijf@tongji.edu.cn; wxrao@tongji.edu.cn; crwang@tongji.edu.cn

Datasets	Type	Study period (days)	No.of patients	Data Availability	CGM / CBG	Food	Exercise	Insulin Use	Published Year
UVA/Padova[14,16–18,41,42]	T1DM	customized	30	Open[43]	✓/✓	✓	×	✓	2018
OhioT1DM[18,19,21,22,44]	T1DM	56	12	Credentialed[45]	✓/✓	✓	✓	✓	2020
D1NAMO[23]	T1DM	4	9	Credentialed[46]	✓/×	✓	✓	✓	2018
ABC4D[18,24]	T1DM	180	10	not accessible	✓/×	✓	×	✓	2020
Weinstock[25]	T1DM	12	201	not accessible	✓/×	×	×	✓	2016
KDD18[26]	T1DM	1095	40	Open[47]	✓/×	×	×	×	2018
Yang[29]	T1/T2DM	1–7	49/51	not accessible	✓/×	N/A	N/A	N/A	2018
Maryland[27]	T2DM	365	N/A	not accessible	×/✓	×	×	✓	2015
Maastricht study[28,30]	T2DM	2	851	Credentialed[48]	✓/✓	×	✓	×	2021

Table 1. A summary on existing diabetes data in the literature. CBG, capillary blood glucose; CGM, continuous glucose monitoring; N/A, not available; T1DM, Type 1 diabetes mellitus; T2DM, Type 2 diabetes mellitus.

review for T1DM[12], 51.7% were with small samples, 29.3% with small-medium samples, 17.2% with simulated data and 1.7% with samples over 50 patients. Another limitation pointed out by the reviews was the low free access data availability. Most data are credentialed or not accessible due to ethical restrictions and data privacy. We summarized recently studied and popular T1DM and T2DM datasets in Table 1.

In T1DM, both real and simulated patient data in silico were well studied. Simulators can conveniently provide and customize detailed data of virtual diabetic patients from their dietary and treatment strategies. UVA/Padova T1DM simulator[14] was widely employed, which was approved by Food and Drug Administration (FDA) and provided 30 different virtual patients freely. Virtual diabetes simulators were studied in tasks such as glycemic events identification, BG control[15] and predictions[14,16–18]. The simulators were able to generate as many BG instances as possible for each patient[14].

As a public dataset, *OhioT1DM*[18–22] was a comprehensive dataset of real T1DM patients in the United States, which was publicly released by Ohio University and contained data of 12 real patients. Compared to the *OhioT1DM*, *D1NAMO*[23] dataset focused on diabetes management. This dataset was composed of 20 real healthy people and nine real T1DM patients with additional patient information such as BG measurements, food pictures, breathing signals and accelerometer outputs. A clinical data[18,24] including 10 T1DM adults from the *ABC4D* project using CGM sensors was used in a deep learning framework for accurate glucose forecasting. *Weinstock*[25] collected diabetes-related data from adult type 1 diabetes ($>=60$ years of age, diabetes duration $>=20$ years). This dataset consisted of 14 days' CGM data, information of insulin, other medications and patient demographics from 201 patients. This dataset was proposed to analyze the risk factors that can cause severe hypoglycemia in old patients. Fox *et al.*[26] collected CGM records from 40 T1DM patients over three years (data size $>=1900$ days of BG measurements, $>=550$k distinct glucose measurements) and developed a deep multi-output forecasting algorithm.

T2DM datasets were less common than T1DM datasets[27,28]. A CGM data from both the T1DM and T2DM patients were employed to predict future BG levels for preventing hyperglycemia or hypoglycemia[29], which was collected over a period ranging from 1.3 to 7 days. The *Maryland* data[27] contained 56,000 SMBG data points collected in a 1-year prospective study. In this study, patients were treated with a variety of medications, including oral antihyperglycemic agents and insulin. The *Maastricht Study*[28,30], an observational, prospective, population-based cohort study, focused on the aetiology, pathophysiology, complications and comorbidities of T2DM, and was characterized by an extensive phenotyping approach.

The existing diabetes data are used not only in BG prediction[31], but also in other diabetes-related fields, such as the generation of BG control strategies[15] and the study of the influence of external factors on blood glucose level. However, the limitations of many diabetes datasets in terms of the number of patients, the racial regions where they are collected, and the types of diabetes mellitus have led to the restrictions in diabetes-related research.

It is known that dietary intake, exercise and medication are the main factors affecting the BG level[32,33]. The collection on these external information is therefore essential in the datasets, which is a tedious task. More specifically, eating habits are quite influenced by ethnic groups and regions, e.g., the Chinese dietary habits are very complicated[34]. Therefore, two datasets from T1DM and T2DM patients in Shanghai, China with dietary information, clinical characteristics, laboratory measurements and medications of the patients were constructed. To the best of our knowledge, these are the first publicly available datasets to include rich information for people with T1DM and T2DM in China. The datasets could contribute to the research in data-driven machine learning.

Methods

Study population. A registry study on Diabetes Data Registry and Individualized Lifestyle Intervention (DiaDRIL) was initiated in Shanghai East Hospital and Shanghai Fourth People's Hospital affiliated to Tongji University since 2019. The aims of this project were to provide evidence for personalized lifestyle recommendations and optimize glycemic control.

In this study, the patients were recruited from DiaDRIL in Shanghai East Hospital (September 2019 to March 2021) and Shanghai Fourth Peoples Hospital (June 2021 to November 2021), respectively. The inclusion criteria were as follows: patients with diagnosed diabetes according to the 1999 World Health Organization (WHO) criteria;

more than 18 years of age, willing to sign the informed consent form and with CGM recording for at least 3 days. Patients were excluded if they reported alcohol or drug abuse, were unable to comply with the study, or were not suitable to attend this study judged by the investigators. Data was anonymous to protect the sensitive information of the patients.

Clinical and laboratory measurements. A standard questionnaire was conducted by trained research staff to obtain demographic information. Information on diagnosis and treatment of diabetes, duration of diabetes, laboratory measurements, comorbidities and pharmacologic treatments were collected from medical records. Each patient underwent a physical examination including measurement of height and weight. Body mass index (BMI) was calculated as weight divided by height squared (kg/m^2). Each patient wore a flash glucose monitoring device (FreeStyle Libre H, Abbott Diabetes Care, Witney, UK) to measure interstitial glucose levels continuously for up to 14 days. CGM glucose data were automatically stored on the sensor every 15 minutes. The data can be obtained by scanning the glucose sensor with the reader and uploaded using the device software. Available laboratory measurements ($\leq$6 months before or after CGM) including glucose metabolism, lipid profile and renal function were obtained from medical records. Any dietary intake including the exact time at consumption and weighed food record was reported by the patients. Hypoglycemic medications during CGM were also recorded.

This study was approved by the Ethics Committee of Shanghai Fourth People's Hospital and Shanghai East Hospital affiliated to Tongji University in accordance with the Declaration of Helsinki. The informed consent was obtained from all the patients.

CGM parameters. Time in range (TIR), one of the critical CGM-derived metrics, reflects the glucose variability and evaluates the quality of glycemic control[35]. It is associated with microvascular complications and macrovascular outcomes of diabetes. TIR is defined as the percentage of time spent in the target glucose range of 70–180 mg/dL. Time below range (TBR) and time above range (TAR) are the percentage of time when blood glucose is below 70 mg/dL and above 180 mg/dL, respectively. For most patients with T1DM or T2DM, the recommended CGM targets by the Advanced Technologies & Treatments for Diabetes (ATTD) consensus were $\geq$70% for TIR, $\leq$25% for TAR and $\leq$4% for TBR[36].

Analysis for CGM data. A clinical important task in diabetes management is the prevention of hypo/hyperglycemic events[37]. The algorithms to prevent the hpyo/hyperglycemic events can be obtained by generating hpyo/hyperalerts on the basis of ahead-of-time prediction of glucose concentration by using past CGM data and suitable time-series models.

Auto-correlation[38] represents the degree of similarity between a given time series and a lagged version of itself over successive time intervals. It can help to uncover hidden patterns in data. Additionally, analyzing the autocorrelation function (ACF) and partial autocorrelation function (PACF) in conjunction is necessary for selecting the appropriate time-series models, e.g., ARIMA[39].

$$\rho_k = \frac{E\left[\left(x_t - \mu\right)\left(x_{t-k} - \mu\right)\right]}{\sigma^2}$$

where x_t is the observation at time t, k is lag, E is the expected value operator, μ is the mean and σ^2 is the variance of the time series. ρ_k can show the correlation between two observations with a lag k in the time series.

Data Records

The datasets *ShanghaiT1DM* and *ShanghaiT2DM* comprise two folders named "Shanghai_T1DM" and "Shanghai_T2DM" and two summary sheets named "Shanghai_T1DM_Summary" and "Shanghai_T2DM_Summary". The datasets can be downloaded through *Figshare* repository[40].

The "Shanghai_T1DM" folder and "Shanghai_T2DM" folder contain 3 to 14 days of CGM data corresponding to 12 patients with T1DM and 100 patients with T2DM, respectively. Of note, for one patient, there might be multiple periods of CGM recordings due to different visits to the hospital, which were stored in different excel tables. In fact, collecting data from different periods in one patient can reflect the changes of diabetes status during the follow-up. The excel table is named by the patient ID, period number and the start date of the CGM recording. Thus, for 12 patients with T1DM, there are 8 patients with 1 period of the CGM recording and 2 patients with 3 periods, totally equal to 16 excel tables in the "Shanghai_T1DM" folder. As for 100 patients with T2DM, there are 94 patients with 1 period of CGM recording, 6 patients with 2 periods, and 1 patient with 3 periods, amounting to 109 excel tables in the "Shanghai_T2DM" folder. Overall, the excel tables include CGM BG values every 15 minutes, capillary blood glucose (CBG) values, blood ketone, self-reported dietary intake, insulin doses and non-insulin hypoglycemic agents. The blood ketone was measured when diabetic ketoacidosis was suspected with a considerably high glucose level. Insulin administration includes continuous subcutaneous insulin infusion using insulin pump, multiple daily injections with insulin pen, and insulin that were given intravenously in case of an extremely high BG level.

Each excel table in the "Shanghai_T1DM" folder and "Shanghai_T2DM" folder contains the following data fields: <Date> Recording time of the CGM data. <CGM> CGM data recorded every 15 minutes. <CBG> CBG level measured by the glucose meter. <Blood ketone> Plasma-hydroxybutyrate measured with ketone test strips (Abbott Laboratories, Abbott Park, Illinois, USA). <Dietary intake> Self-reported time and weighed

Characteristics	ShanghaiT1DM (n = 12)	ShanghaiT2DM (n = 100)	p value
Age, years	57.83 ± 11.12	60.17 ± 13.71	0.571
Women, n (%)	7 (58.3%)	44 (44.0%)	0.346
BMI, kg/m^2	20.95 [17.87–24.21]	23.69 [22.12–25.54]	0.017
Duration of diabetes, years	8.50 [2.25–16.75]	7.00 [1.00–14.75]	0.614
Fasting plasma glucose, mg/dL	184.08 [117.00–262.35]	158.40 [126.00–194.40]	0.410
2-hour postprandial plasma glucose, mg/dL	297.00 [248.76–348.84]	250.65 [196.16–317.88]	0.218
HbA1c, mmol/mol	71 [63–122]	69 [54–97]	0.223

Table 2. The characteristics of the T1DM and T2DM patients in the ShanghaiT1DM and ShanghaiT2DM. Data are presented as mean ± SD, median [interquartile range], or number(percentage%). BMI, body mass index; HbA1c, hemoglobin A1c; T1DM, Type 1 diabetes mellitus; T2DM, Type 2 diabetes mellitus.

Datasets	Type of diabetes	Study period (days)	Monitoring interval (minutes)	No. of patients	No. of recording file	Total CGM measurements
ShanghaiT1DM	T1DM	4–14	15	12	16	15,695
ShanghaiT2DM	T2DM	3–14	15	100	109	112,475
SimulatorT1DM	T1DM	56	5	30	unlimited	482,610
OhioT1DM	T1DM	56	5	12	12	191,605

Table 3. General characteristics of the datasets. CGM, continuous glucose monitoring; T1DM, Type 1 diabetes mellitus; T2DM, Type 2 diabetes mellitus; No., number.

food intake <Insulin dose-s.c.> Subcutaneous insulin injection with insulin pen. <Insulin dose-i.v.> Dose of intravenous insulin infusion. <Non-insulin hypoglycemic agents> Hypoglycemic agents other than insulin. <CSII-bolus insulin> Dose of insulin delivered before a meal through insulin pump. <CSII-basal insulin> The rate (iu/per hour) at which basal insulin was continuously infused through insulin pump.

The summary sheets summarize the clinical characteristics, laboratory measurements and medications of the patients included in this study, with each row corresponding to one excel table in "Shanghai_T1DM" and "Shanghai_T2DM" folders. Clinical characteristics include patient ID, gender, age, height, weight, BMI, smoking and drinking history, type of diabetes, duration of diabetes, diabetic complications, comorbidities as well as occurrence of hypoglycemia. Laboratory measurements contain fasting and 2-hour postprandial plasma glucose/C-peptide/insulin, hemoglobin A1c (HbA1c), glycated albumin, total cholesterol, triglyceride, high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, creatinine, estimated glomerular filtration rate, uric acid and blood urea nitrogen. Both hypoglycemic agents and medications given for other diseases before the CGM reading were also recorded.

Technical Validation

The characteristics of the Chinese diabetes datasets. The detailed characteristics of the patients in the *ShanghaiT1DM* and *ShanghaiT2DM* datasets were summarized in Table 2. The age of the *ShanghaiT1DM* group and the *ShanghaiT2DM* group was 57.8 ± 11.1 and 60.2 ± 13.7 years, respectively. There was no statistically significant difference in age between the *ShanghaiT1DM* group and *ShanghaiT2DM* group. This is because most of the patients (10/12) in the *ShanghaiT1DM* group belonged to a subtype of T1DM called "latent autoimmune diabetes in adults", which is characterized by slow autoimmune β-cell destruction and an older mean age at onset of diabetes[1]. Women accounted for 58.3% of the *ShanghaiT1DM* group and 44% of the *ShanghaiT2DM* group, respectively. Besides, data concerning fasting plasma glucose, 2-hour postprandial plasma glucose and HbA1c were comparable between the two groups. However, the *ShanghaiT2DM* group had higher BMI values than the *ShanghaiT1DM* group ($p < 0.05$).

To show the size of these two datasets more intuitively, we listed the patient's type, the study period, sampling interval of CGM devices, number of patients, total number of recording files and total CGM measurements of the *ShanghaiT1DM* and *ShanghaiT2DM* in Table 3. For a given patient, he or she may have more than one recording period. In Fig. 1, we showed the number of recording files with different CGM data size in days in the *ShanghaiT1DM* and *ShanghaiT2DM*. The collected CGM data size varied from 3 days to 14 days.

We summarized the hypo/hyperglycemia events and calculated the auto-correlation coefficient on the BG values of the two datasets in time series. Hypoglycemia and hyperglycemia events are two potential risk factors for complications in diabetes. Hence, the time percentages of hypoglycemia (TBR) and hyperglycemia (TAR) events for each patient were calculated in Fig. 2. The horizontal axis represented each recording file of the patients with an order of TBR increasing, while the vertical axis represented the percentage of time (TAR, TIR and TBR) during the data collection period. The higher values of the TAR and TBR indicated that the patient's condition was more serious. To give a clearer view of the TBR, TIR and TAR in the two datasets, we calculated the mean ± standard deviation of these values for the two datasets. For the *ShanghaiT1DM*, the mean ± standard

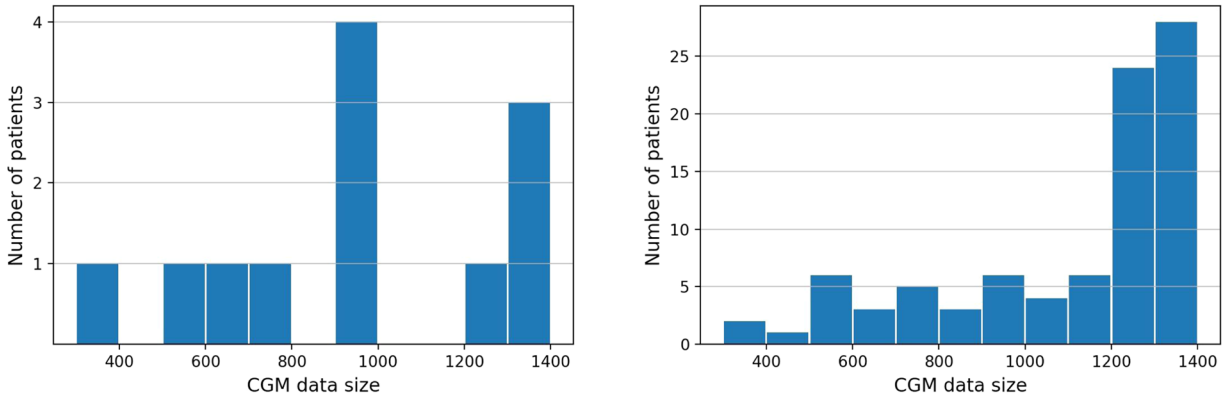

Fig. 1 The number of recording files with different CGM data size in days (**a**) ShanghaiT1DM dataset (**b**) ShanghaiT2DM dataset.

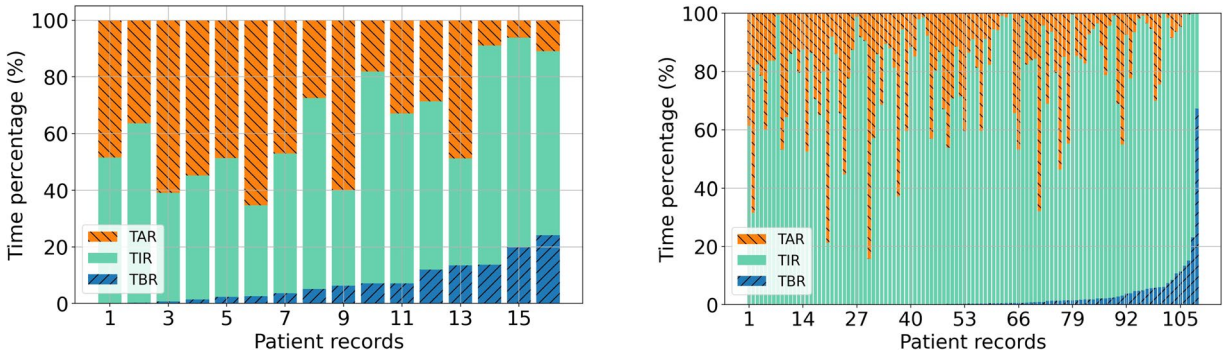

Fig. 2 The average percentage of TBR (time below range), TIR (time in range) and TAR (time above range) for CGM in two datasets. (**a**) ShanghaiT1DM: TAR ($37.8 \pm 18.8\%$), TIR ($54.7 \pm 14.5\%$), TBR ($7.5 \pm 7.0\%$). (**b**) ShanghaiT2DM: TAR ($20.0 \pm 18.4\%$), TIR ($77.7 \pm 18.1\%$), TBR ($2.4 \pm 7.2\%$). Data are presented as mean ± SD.

deviation of the TIR were $54.7 \pm 14.5\%$ and $77.7 \pm 18.1\%$ for the *ShanghaiT2DM*. We noted that the average TIR was higher in T2DM patients than in T1DM patients (Fig. 2).

Besides, as the collection on individual patient's behavior information in each dataset was different, we randomly chose three patients from each dataset for the auto-correlation graph of the BG time series in Fig. 3. The auto-correlation coefficients identify seasonality and trend in time series data. It can be found that patients in *ShanghaiT2DM* (Fig. 3b) showed a more noticeable 24-hour periodic pattern than those in *ShanghaiT1DM* (Fig. 3a).

Since there might be discrepancy in BG levels by different blood glucose monitoring methods, we conducted a comparative analysis of the blood glucose measured by the CGM and CBG in Fig. 4, 5. The collection of the CBG was more sparse than that of the CGM, we only plotted the time stamps with both of the measurements. Two patients were randomly selected from each dataset. The results showed that the CBG values were usually greater than those of CGM readings.

Comparison to other datasets. There have been widely used datasets such as the *SimulatorT1DM* and the *OhioT1DM* (see Table 3). In order to show more specifically the difference between the newly constructed datasets and other existing data, the comparisons were performed in Table 3, figs. 3c,d & 6.

The auto-correlation coefficients of the *ShanghaiT1DM* (Fig. 3a) and *OhioT1DM* (Fig. 3d) indicated that the two real T1DM datasets shared similar trend and periodic pattern, which made it possible to combine the two datasets together in certain research. The *SimulatorT1DM* (Fig. 3c) had strong regularity as it was simulated.

Achieving higher TIR has been shown to reduce the percentages of time in the hypoglycemic and hyperglycemic range and complications of diabetes. In Fig. 6, we found that the patients in the *OhioT1DM* had lower mean TBR values compared to those in the *ShanghaiT1DM* (Fig. 2), which means that they have better control of hypoglycemia. In addition, patients in the *ShanghaiT2DM* (Fig. 2) had the highest mean TIR values, which suggests that people with T2D have better glycemic control overall than people with T1D. The virtual patients from the UVA/Padova (Fig. 6) had worse control of hypoglycemia, which may be due to the fact that the glycemic control strategy of the virtual patients was based on a fixed formula and therefore could not produce a

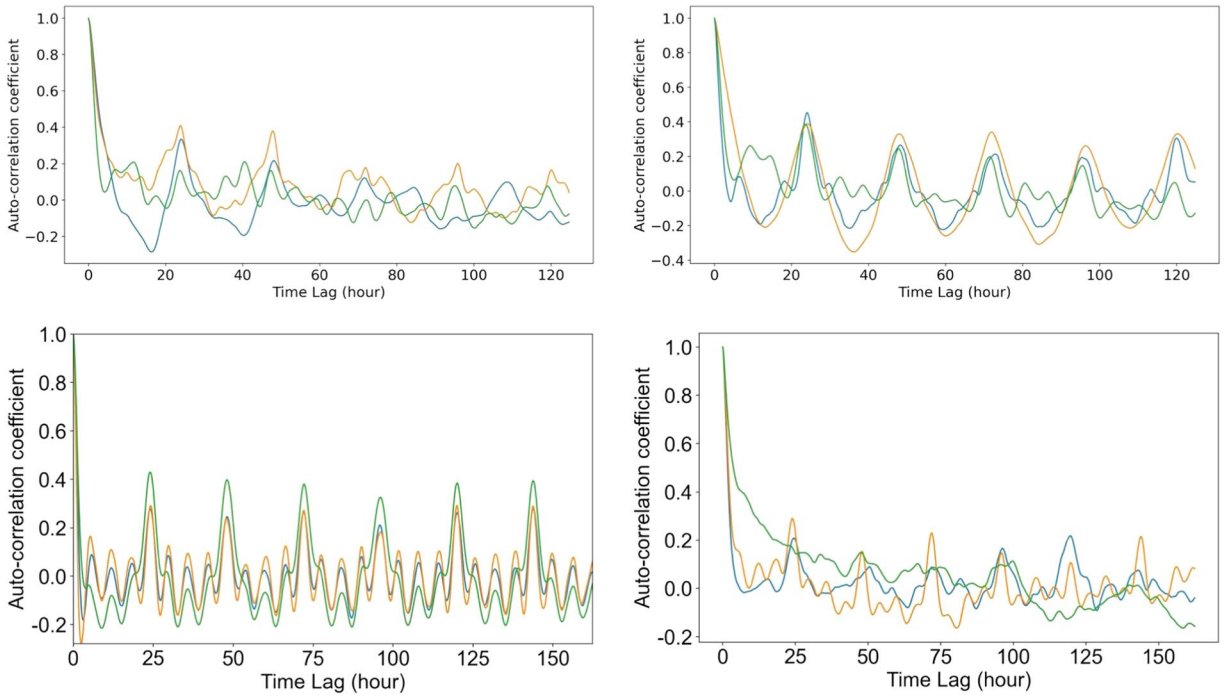

Fig. 3 Auto-correlation coefficient of randomly picked three patients from the (**a**) ShanghaiT1DM, (**b**) ShanghaiT2DM, (**c**) SimulatorT1DM and (**d**) OhioT1DM.

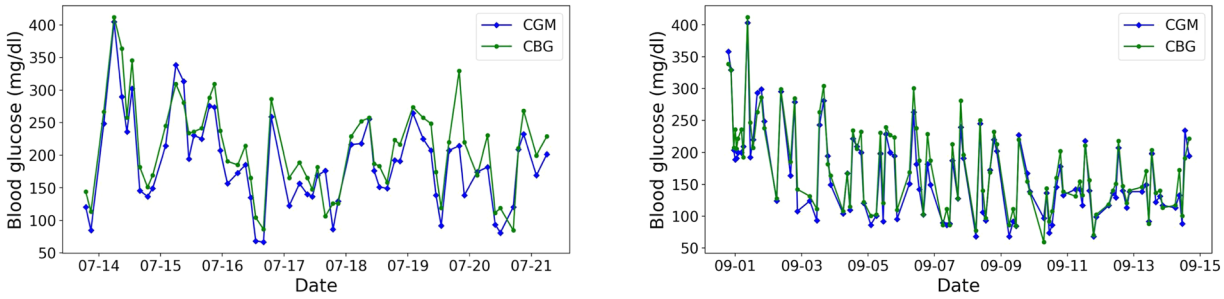

Fig. 4 Randomly selected patients (**a**) 1008_0_20210713 and (**b**) 1003_0_20210831 in the ShanghaiT1DM for the distributions of glucose values of CGM readings and CBG. (CGM, continuous glucose monitoring; CBG, capillary blood glucose).

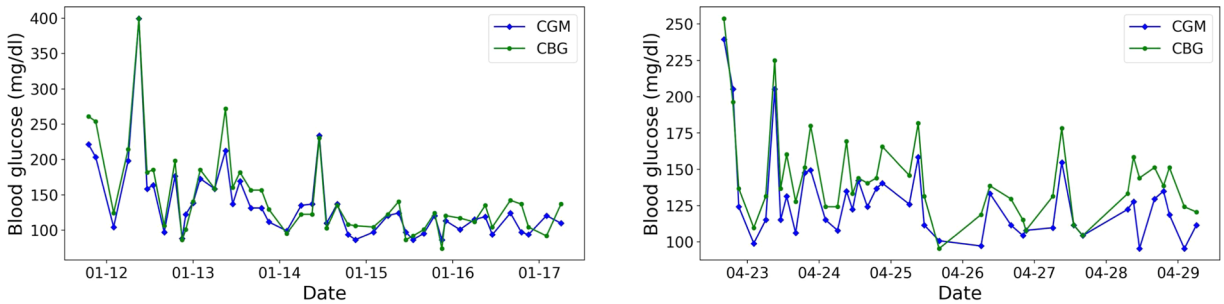

Fig. 5 Randomly selected patients (**a**) 2010_0_20220111 and (**b**) 2022_0_20210419 in the ShanghaiT2DM for the distributions of glucose values of CGM readings and CBG. (CGM, continuous glucose monitoring; CBG, capillary blood glucose).

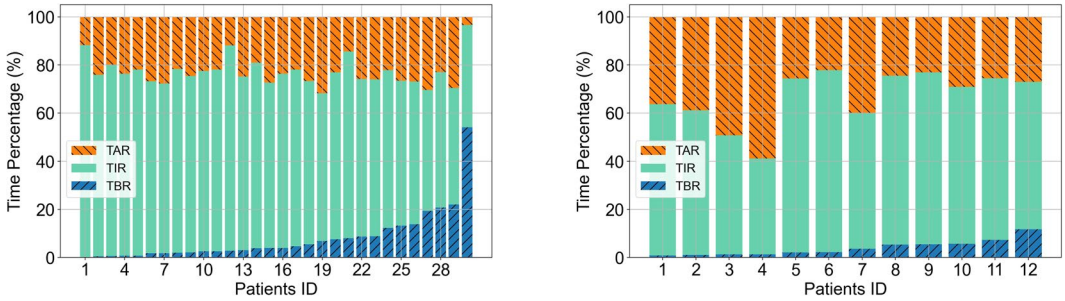

Fig. 6 The average percentage of TBR (time below range), TIR (time in range) and TAR (time above range) for CGM (continuous glucose monitoring) in two datasets. (**a**) SimulatorT1DM: TAR (22.9 ± 5.9%), TIR (69.1 ± 10.2%),TBR (8.0 ± 10.5%). (**b**) OhioT1DM: TAR (33.4 ± 11.1%),TIR (62.6 ± 9.9%),TBR (4.0 ± 3.1%). Data are presented as mean ± SD.

timely response to the hypoglycemia. By comparing the *ShanghaiT1DM* and *OhioT1DM* (Fig. 6), we found that the standard deviations of TBR, TIR and TAR in the *ShanghaiT1DM* were higher than those in the *OhioT1DM*.

Code availability

The code for the analysis of the datasets and the generation of the figures and tables can be accessed in the Figshare repository[40], which is a JUPYTER notebook named "data_analysis.ipynb". The script can be executed with Python 3.6 and allows for reproducibility and code reuse.

Received: 14 April 2022; Accepted: 6 January 2023;
Published online: 19 January 2023

References

1. American Diabetes Association. Professional Practice Committee. 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes-2022. *Diabetes Care* **45**, S17–S38, https://doi.org/10.2337/dc22-S002 (2022).
2. Kahn, S. E., Hull, R. L. & Utzschneider, K. M. Mechanisms linking obesity to insulin resistance and type 2 diabetes. *Nature* **444**, 840–846 (2006).
3. *IDF DIABETES ATLAS*, 10th edn. (Brussels: International Diabetes Federation, 2021).
4. Chinese Diabetes Society. Guideline for the prevention and treatment of type 2 diabetes mellitus in china (2020 edition). *Chin J Diabetes Mellitus* **13**, 315–409, https://doi.org/10.3760/cma.j.cn115791-20210221-00095 (2021).
5. Diabetes Control and Complications Trial Research Group. *et al*. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *N Engl J Med.* **329**, 977–986 (1993).
6. UK Prospective Diabetes Study (UKPDS) Group. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). *Lancet* **352**, 837–853 (1998).
7. Holman, R., Paul, S., Bethel, M., Matthews, D. & Neil, H. 10-year follow-up of intensive glucose control in type 2 diabetes. *N Engl J Med.* **359**, 1577–1589 (2008).
8. American Diabetes Association. Introduction: Standards of medical care in diabetes-2022. *Diabetes Care* **45**, S1–S2, https://doi.org/10.2337/dc22-Sint (2022).
9. Benjamin, E. M. Self-monitoring of blood glucose: The basics. *Clinical Diabetes* **20**, 45–47 (2002).
10. Bao, Y. *et al*. Chinese clinical guidelines for continuous glucose monitoring (2018 edition). *Diabetes/metabolism research and reviews* **35**, e3152 (2019).
11. Galindo, R. J. & Aleppo, G. Continuous glucose monitoring: the achievement of 100 years of innovation in diabetes technology. *Diabetes Research and Clinical Practice* **170**, 108502 (2020).
12. Woldaregay, A. Z. *et al*. Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes. *Artificial Intelligence in Medicine* **98**, 109–134 (2019).
13. Felizardo, V., Garcia, N. M., Pombo, N. & Megdiche, I. Data-based algorithms and models using diabetics real data for blood glucose and hypoglycaemia prediction-a systematic literature review. *Artificial Intelligence in Medicine* **118**, 102120 (2021).
14. Visentin, R. *et al*. The UVA/Padova type 1 diabetes simulator goes from single meal to single day. *J Diabetes Sci Technol.* **12**, 273–281 (2018).
15. Zhu, J. *et al*. Reinforcement learning for diabetes blood glucose control with meal information. In Wei, Y., Li, M., Skums, P. & Cai, Z. (eds.) *Bioinformatics Research and Applications*, 80–91 (Springer International Publishing, Cham, 2021).
16. Pompa, M., Panunzi, S., Borri, A. & De Gaetano, A. A comparison among three maximal mathematical models of the glucose-insulin system. *PloS one* **16**, e0257789 (2021).
17. Contreras, I., Oviedo, S., Vettoretti, M., Visentin, R. & Veh, J. Personalized blood glucose prediction: A hybrid approach using grammatical evolution and physiological models. *PloS one* **12**, e0187754 (2017).
18. Li, K., Liu, C., Zhu, T., Herrero, P. & Georgiou, P. GluNet: A deep learning framework for accurate glucose forecasting. *IEEE Journal of Biomedical and Health Informatics* **24**, 414–423 (2020).
19. Marling, C. & Bunescu, R. The OhioT1DM dataset for blood glucose level prediction: Update 2020. In *CEUR workshop proceedings*, **vol. 2675**, 71 (NIH Public Access, 2020).
20. Marling, C. & Bunescu, R. C. The OhioT1DM dataset for blood glucose level prediction. In *KHD@ IJCAI* (2018).
21. Xie, J. & Wang, Q. Benchmarking machine learning algorithms on blood glucose prediction for type I diabetes in comparison with classical time-series models. *IEEE Transactions on Biomedical Engineering* **67**, 3101–3124 (2020).
22. Martinsson, J., Schliep, A., Eliasson, B. & Mogren, O. Blood glucose prediction with variance estimation using recurrent neural networks. *Journal of Healthcare Informatics Research* **4**, 1–18 (2020).
23. Dubosson, F. *et al*. The open D1NAMO dataset: A multi-modal dataset for research on non-invasive type 1 diabetes management. *Informatics in Medicine Unlocked* **13**, 92–100 (2018).
24. Reddy, M. *et al*. Clinical safety and feasibility of the advanced bolus calculator for type 1 diabetes based on case-based reasoning: A 6-week nonrandomized single-arm pilot study. *Diabetes Technol Ther* **487** (2016).
25. Weinstock, R. S. *et al*. Risk factors associated with severe hypoglycemia in older adults with type 1 diabetes. *Diabetes Care* **39**, 603–610 (2016).

26. Fox, I., Ang, L., Jaiswal, M., Pop-Busui, R. & Wiens, J. Deep multi-output forecasting: Learning to accurately predict blood glucose trajectories. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1387–1395 (2018).
27. Sudharsan, B., Peeples, M. & Shomali, M. Hypoglycemia prediction using machine learning models for patients with type 2 diabetes. *Journal of Diabetes Science & Technology* **9**, 86 (2015).
28. van Doorn, W. P. *et al*. Machine learning-based glucose prediction with use of continuous glucose and physical activity monitoring data: The maastricht study. *PloS one* **16**, e0253125 (2021).
29. Yang, J., Li, L., Shi, Y. & Xie, X. An ARIMA model with adaptive orders for predicting blood glucose concentrations and hypoglycemia. *IEEE Journal of Biomedical and Health Informatics* **23**, 1251–1260 (2018).
30. Schram, M. T. *et al*. The maastricht study: an extensive phenotyping study on determinants of type 2 diabetes, its complications and its comorbidities. *European Journal of Epidemiology* **29**, 439–451 (2014).
31. Zhu, T., Yao, X., Li, K., Herrero, P. & Georgiou, P. Blood glucose prediction for type 1 diabetes using generative adversarial networks. *CEUR Workshop Proceedings* **2675**, 90–94 (2020).
32. Pan, X. *et al*. Effects of diet and exercise in preventing NIDDM in people with impaired glucose tolerance. The Da Qing IGT and Diabetes Study. *Diabetes Care* **20**, 537–544 (1997).
33. Tuomilehto, J. *et al*. Finnish diabetes prevention study group. prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. *N Engl J Med.* **344**, 1343–1350 (2001).
34. Mora, N. & Golden, S. H. Understanding cultural influences on dietary habits in asian, middle eastern, and latino patients with type 2 diabetes: A review of current literature and future directions. *Curr Diab Rep.* **17**, 126 (2017).
35. Danne, T. *et al*. International consensus on use of continuous glucose monitoring. *Diabetes Care* **40**, 1631–1640 (2017).
36. Battelino, T. *et al*. Clinical targets for continuous glucose monitoring data interpretation: Recommendations from the international consensus on time in range. *Diabetes Care* **42**, 1593–1603 (2019).
37. Sparacino, G. *et al*. Glucose concentration can be predicted ahead in time from continuous glucose monitoring sensor time-series. *IEEE Transactions on Biomedical Engineering* **54**, 931–937 (2007).
38. Yin, J. *et al*. Experimental study of multivariate time series forecasting models. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2833–2839 (2019).
39. Zhang, G. P. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing* **50**, 159–175 (2003).
40. Zhao, Q. *et al*. Diabetes Datasets, ShanghaiT1DM and ShanghaiT2DM, *figshare*, https://doi.org/10.6084/m9.figshare.c.6310860 (2022).
41. Turksoy, K. *et al*. Meal detection in patients with type 1 diabetes: a new module for the multivariable adaptive artificial pancreas control system. *IEEE Journal of Biomedical and Health Informatics* **20**, 47–54 (2015).
42. Haidar, A. The artificial pancreas: How closed-loop control is revolutionizing diabetes. *IEEE Control Systems Magazine* **36**, 28–47 (2016).
43. Xie, J. *Simglucose v0.2.1*. https://github.com/jxx123/simglucose (2018).
44. Veh, J., Contreras, I., Oviedo, S., Biagi, L. & Bertachi, A. Prediction and prevention of hypoglycaemic events in type-1 diabetic patients using machine learning. *Health Informatics Journal* **26**, 703–718 (2020).
45. Marling, C. & Bunescu, R. OhioT1DM, http://smarthealth.cs.ohio.edu/OhioT1DM-dataset.html (2020).
46. Dubosson, F. *et al*. The open D1NAMO dataset: A multi-modal dataset for research on non-invasive type 1 diabetes management. *Zenodo* https://doi.org/10.5281/zenodo.1421616 (2018).
47. Fox, I., Ang, L., Jaiswal, M., Pop-Busui, R. & Wiens, J. Learning to accurately predict blood glucose trajectories, https://github.com/igfox/multi-output-glucose-forecasting (2018).
48. Stehouwer, C. *et al*. Maastricht study, https://www.demaastrichtstudie.nl/research (2014).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 61972286, 82070913), the Shanghai Municipal Science and Technology Major Project (2021SHZDZX0100), the Natural Science Foundation of Shanghai, China (Grant No. 20ZR1460500, 22511104300), the Shanghai Science and Technology Development Funds (Grant No. 20ZR1446000, 22410713200), the Fundamental Research Funds for the Central Universities and the Research fund from Shanghai Fourth People's Hospital (sykyqd01801, SY-XKZT-2021-1001). Finally, thanks Ms. Xiongbaixue Yan for her previous efforts on the management of the project.

Author contributions

Q.Z., J.Z. and C.W. had the initial idea for this study. C.L., X.S., B.C. and C.W. established the datasets, i.e., *ShanghaiT1DM* and *ShanghaiT2DM*. Y. Liang verified the food data. Q.P. and J.Z. designed and performed the technical validation. J.Z., X.S. and Q.Z. drafted the paper. J.L., C.W. and W.Rao jointly supervised the work. All authors participated in verifying the data and revising the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.L., W.R. or C.W.