HHS Public Access

Author manuscript

CEUR Workshop Proc. Author manuscript; available in PMC 2021 February 13.

Published in final edited form as:

CEUR Workshop Proc. 2020 September ; 2675: 71–74.

The OhioT1DM Dataset for Blood Glucose Level Prediction: Update 2020

Cindy Marling[1], **Razvan Bunescu**[1]

[1]Ohio University, USA

Abstract

This paper documents the OhioT1DM Dataset, which was developed to promote and facilitate research in blood glucose level prediction. It contains eight weeks' worth of continuous glucose monitoring, insulin, physiological sensor, and self-reported life-event data for each of 12 people with type 1 diabetes. An associated graphical software tool allows researchers to visualize the integrated data. The paper details the contents and format of the dataset and tells interested researchers how to obtain it.

The OhioT1DM Dataset was first released in 2018 for the first Blood Glucose Level Prediction (BGLP) Challenge. At that time, the dataset was half its current size, containing data for only six people with type 1 diabetes. Data for an additional six people is being released in 2020 for the second BGLP Challenge. This paper subsumes and supersedes the paper which documented the original dataset.

1 INTRODUCTION

Accurate forecasting of blood glucose levels has the potential to improve the health and wellbeing of people with diabetes. Knowing in advance when blood glucose is approaching unsafe levels provides time to proactively avoid hypo- and hyper-glycemia and their concomitant complications. The drive to perfect an artificial pancreas [2] has increased the interest in using machine learning (ML) approaches to improve prediction accuracy. Work in this area has been hindered, however, by a lack of real patient data; some researchers have only been able to work on simulated patient data.

To promote and facilitate research in blood glucose level prediction, we have curated the OhioT1DM Dataset and made it publicly available for research purposes. To the best of our knowledge, this is the first publicly available dataset to include continuous glucose monitoring, insulin, physiological sensor, and self-reported life-event data for people with type 1 diabetes.

The OhioT1DM Dataset contains eight weeks' worth of data for each of 12 people with type 1 diabetes. These anonymous people are referred to by randomly selected ID numbers. All data contributors were on insulin pump therapy with continuous glucose monitoring (CGM). They wore Medtronic 530G or 630G insulin pumps and used Medtronic Enlite CGM sensors

marling@ohio.edu, bunescu@ohio.edu.

throughout the 8-week data collection period. They reported life-event data via a custom smartphone app and provided physiological data from a fitness band. The first cohort of six individuals wore Basis Peak fitness bands. Data for this cohort was released in 2018. The second cohort of six individuals wore the Empatica Embrace. Data for this cohort is included in the 2020 release. Table 1 shows the gender, age range, insulin pump model, and sensor band type for each data contributor, by cohort.

The dataset includes: a CGM blood glucose level every 5 minutes; blood glucose levels from periodic self-monitoring of blood glucose (finger sticks); insulin doses, both bolus and basal; self-reported meal times with carbohydrate estimates; self-reported times of exercise, sleep, work, stress, and illness; and data from the Basis Peak or Empatica Embrace band. The data for individuals who wore the Basis Peak band includes 5-minute aggregations of heart rate, galvanic skin response (GSR), skin temperature, air temperature, and step count. The data for those who wore the Empatica Embrace band includes 1-minute aggregations of GSR, skin temperature, and magnitude of acceleration. Both bands indicated the times they detected that the wearer was asleep, and this information is included when available. However, not all data contributors wore their sensor bands overnight.

Data for the first six individuals was released in 2018 for the first Blood Glucose Level Prediction (BGLP) Challenge, which was held in conjunction with the 3rd International Workshop on Knowledge Discovery in Healthcare Data, at IJCAI-ECAI 2018, in Stockholm, Sweden. Data for six additional people is being released in 2020 for the second BGLP Challenge, to be held at the 5th International Workshop on Knowledge Discovery in Healthcare Data, at ECAI 2020, in Santiago de Compostela, Spain. This paper subsumes and supersedes the paper which documented the original 2018 dataset [3]. In order to provide a unified overview of the entire dataset, this paper incorporates most of the original paper verbatim.

The following sections of this paper provide background information, detail the data format, describe the OhioT1DM Viewer visualization software, and tell how to obtain the OhioT1DM Dataset and Viewer for research purposes.

2 BACKGROUND

We have been working on intelligent systems for diabetes management since 2004 [1, 4, 5, 6, 7, 8, 10, 11]. As part of our work, we have run five clinical research studies involving subjects with type 1 diabetes on insulin pump therapy. Over 50 anonymous subjects have provided blood glucose, insulin, and life-event data so that we could develop software intended to help people with diabetes and their professional health care providers.

Our most recent study was designed so that de-identified data could be shared with the research community. All data contributors to the OhioT1DM Dataset signed informed consent documents allowing us to share their de-identified data with outside researchers. This agreement clearly delineated what types of data could be shared and with whom. The data in the dataset was fully de-identified according to the Safe Harbor method, a standard specified by the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule

[9]. To protect the data contributors and to ensure that the data is used only for research purposes, a Data Use Agreement (DUA) must be executed before a researcher can obtain the data.

3 OhioT1DM DATA FORMAT

For each data contributor, there is one XML file for training and development data and a separate XML file for testing data. This results in a total of 24 XML files, two for each of the 12 contributors. Table 2 shows the number of training and test examples for each contributor.

Table 2 also indicates the BGLP Challenge for which the data was released. For the 2018 BGLP Challenge, the number of test examples was equal to the number of data points in the XML testing file. However, for the 2020 BGLP Challenge, the first hour of data in each XML testing file is excluded from the set of points used for evaluation. This is to allow unbiased comparison of prediction models using all training data to predict each test point, as the first test points would otherwise be too close chronologically to the training data. Thus, for the 2020 BGLP Challenge, there are 12 more data points in each XML testing file than the number of test examples shown in Table 2.

Each XML file contains the following data fields:

1. <**patient**> The patient ID number and insulin type. Weight is set to 99 as a placeholder, as actual patient weights are unavailable.
2. <**glucose level**> Continuous glucose monitoring (CGM) data, recorded every 5 minutes.
3. <**finger stick**> Blood glucose values obtained through self-monitoring by the patient.
4. <**basal**> The rate at which basal insulin is continuously infused. The basal rate begins at the specified timestamp *ts*, and it continues until another basal rate is set.
5. <**temp basal**> A temporary basal insulin rate that supersedes the patient's normal basal rate. When the value is 0, this indicates that the basal insulin flow has been suspended. At the end of a temp basal, the basal rate goes back to the normal basal rate, <basal>
6. <**bolus**> Insulin delivered to the patient, typically before a meal or when the patient is hyperglycemic. The most common type of bolus, normal, delivers all insulin at once. Other bolus types can stretch out the insulin dose over the period between *ts begin* and *ts end*.
7. <**meal**> The self-reported time and type of a meal, plus the patient's carbohydrate estimate for the meal.
8. <**sleep**> The times of self-reported sleep, plus the patient's subjective assessment of sleep quality: 1 for Poor; 2 for Fair; 3 for Good.

9. <**work**> Self-reported times of going to and from work. Intensity is the patient's subjective assessment of physical exertion, on a scale of 1 to 10, with 10 the most physically active.
10. <**stressors**> Time of self-reported stress.
11. <**hypo event**> Time of self-reported hypoglycemic episode. Symptoms are not available, although there is a slot for them in the XML file.
12. <**illness**> Time of self-reported illness.
13. <**exercise**> Time and duration, in minutes, of self-reported exercise. Intensity is the patient's subjective assessment of physical exertion, on a scale of 1 to 10, with 10 the most physically active.
14. <**basis heart rate**> Heart rate, aggregated every 5 minutes. This data is only available for people who wore the Basis Peak sensor band.
15. <**basis gsr**> Galvanic skin response, also known as skin conductance or electrodermal activity. For those who wore the Basis Peak, the data was aggregated every 5 minutes. Despite this attribute's name, it is also available for those who wore the Empatica Embrace. For these individuals, the data is aggregated every 1 minute.
16. <**basis skin temperature**> Skin temperature, in degrees Fahrenheit, aggregated every 5 minutes, for those who wore the Basis Peak, and every 1 minute, for those who wore the Empatica Embrace.
17. <**basis air temperature**> Air temperature, in degrees Fahrenheit, aggregated every 5 minutes. This data is only available for people who wore the Basis Peak sensor band.
18. <**basis steps**> Step count, aggregated every 5 minutes. This data is only available for people who wore the Basis Peak sensor band.
19. <**basis sleep**> Times when the sensor band reported that the subject was asleep. For those who wore the Basis Peak, there is also a numeric estimate of sleep quality.
20. <**acceleration**> Magnitute of acceleration, aggregated every 1 minute. This data is only available for people who wore the Empatica Embrace sensor band.

Note that, in de-identifying the dataset, all dates for each individual were shifted by the same random amount of time into the future. The days of the week and the times of day were maintained in the new timeframes. However, the months were shifted, so that it is not possible to consider the effects of seasonality or of holidays.

4 THE OhioT1DM VIEWER

The OhioT1DM Viewer is a visualization tool that opens an XML file from the OhioT1DM Dataset and graphically displays the integrated data. It aids in developing intuition about the data and also in debugging. For example, if a system makes a poor blood glucose level

prediction at a particular point in time, viewing the data at that time might illuminate a cause. For example, the subject might have forgotten to report a meal or might have been feeling ill or stressed.

Figure 1 shows a screenshot from the OhioT1DM Viewer. The data is displayed one day at a time, from midnight to midnight. Controls allow the user to move from day to day and to toggle any type of data off or on for targeted viewing.

The bottom pane shows blood glucose, insulin, and self-reported life-event data. CGM data is displayed as a mostly blue curve, with green points indicating hypoglycemia. Finger sticks are displayed as red dots. Boluses are displayed along the horizontal axis as orange and yellow circles. The basal rate is indicated as a black line. Temporary basal rates appear as red lines. Self-reported sleep is indicated by blue regions. Life-event icons appear at the top of the pane as dots, squares, and triangles. The data in the bottom pane is clickable, so that additional information about any data point can be displayed. For example, clicking on a meal (a square blue icon) displays the timestamp, type of meal, and carbohydrate estimate.

The top pane displays sensor band data. Blue regions in the top pane are times the sensor band detected that the wearer was asleep. The step count is indicated by vertical blue lines. The curves show heart rate (red), galvanic skin response (green), skin temperature (gold), air temperature (cyan), and magnitude of acceleration (black).

5 OBTAINING THE DATASET AND VIEWER

The original 2018 OhioT1DM Dataset and the OhioT1DM Viewer are now available to researchers. The full 2020 OhioT1DM Dataset is currently being released to participants in the second BGLP Challenge. The second BGLP Challenge will take place June 9, 2020, in conjunction with the 5th International Workshop on Knowledge Discovery in Healthcare Data at ECAI 2020, in Santiago de Compostela, Spain.

After the completion of the BGLP Challenge, the entire dataset will be made available to other researchers. To protect the data contributors and to ensure that the data is used only for research purposes, a Data Use Agreement (DUA) is required. A DUA is a binding document signed by legal signatories of Ohio University and the researcher’s home institution. As of this writing, researchers can request a DUA at http://smarthealth.cs.ohio.edu/OhioT1DM-dataset.html. Once a DUA is executed, the OhioT1DM Dataset and Viewer will be directly released to the researcher.

6 CONCLUSION

The OhioT1DM Dataset was developed to promote and facilitate research in blood glucose level prediction. Accurate blood gluocose level predictions could positively impact the health and well-being of people with diabetes. In addition to their role in the artificial pancreas project, such predictions could also enable other beneficial applications, such as decision support for avoiding impending problems, “what if” analyses to project the effects of different lifestyle choices, and enhanced blood glucose profiles to aid in individualizing

diabetes care. It is our hope that sharing this dataset will help to advance the state of the art in blood glucose level prediction.

ACKNOWLEDGEMENTS

This work was supported by grant 1R21EB022356 from the National Institutes of Health (NIH). The OhioT1DM Viewer was originally implemented by Hannah Quillin and Charlie Murphy, and further refined by Robin Kelby and Jeremy Beauchamp. The authors gratefully acknowledge the contributions of Emeritus Professor of Endocrinology Frank Schwartz, MD, a pioneer in building intelligent systems for diabetes management. We would also like to thank our physician collaborators, Aili Guo, MD, and Amber Healy, DO, our research nurses, Cammie Starner and Lynn Petrik, and our past and present graduate and undergraduate research assistants. We are especially grateful to the 12 anonymous individuals with type 1 diabetes who shared their data, enabling the creation of this dataset.

REFERENCES

[1]. Bunescu R, Struble N, Marling C, Shubrook J, and Schwartz F, ‘Blood glucose level prediction using physiological models and support vector regression’, in Proceedings of the Twelfth International Conference on Machine Learning and Applications (ICMLA), pp. 135–140. IEEE Press, (2013).

[2]. Juvenile Diabetes Research Foundation (JDRF). Artificial Pancreas, 2019. Available at http://www.jdrf.org/impact/research/artificial-pancreas/, accessed January, 2020.

[3]. Marling C and Bunescu R, ‘The OhioT1DM dataset for blood glucose level prediction’, in The 3rd International Workshop on Knowledge Discovery in Healthcare Data, Stockholm, Sweden, (July 2018). Available at http://ceur-ws.org/Vol-2148/paper09.pdf, accessed January, 2020.

[4]. Marling C, Shubrook J, and Schwartz F, ‘Toward case-based reasoning for diabetes management: A preliminary clinical study and decision support system prototype’, Computational Intelligence, 25(3), 165–179, (2009).

[5]. Marling C, Wiley M, Bunescu R, Shubrook J, and Schwartz F, ‘Emerging applications for intelligent diabetes management’, AI Magazine, 33(2), 67–78, (2012).

[6]. Marling C, Xia L, Bunescu R, and Schwartz F, ‘Machine learning experiments with noninvasive sensors for hypoglycemia detection’, in IJCAI 2016 Workshop on Knowledge Discovery in Healthcare Data, New York, NY, (2016).

[7]. Mirshekarian S, Bunescu R, Marling C, and Schwartz F, ‘Using LSTMs to learn physiological models of blood glucose behavior’, in Proceedings of the 39th International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2017), Jeju Island, Korea, (2017).

[8]. Mirshekarian S, Shen H, Bunescu R, and Marling C, ‘LSTMs and neural attention models for blood glucose prediction: Comparative experiments on real and synthetic data’, in Proceedings of the 41st International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2019), Berlin, Germany, (2019).

[9]. Office for Civil Rights. Guidance regarding methods for deidentification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) privacy rule, 2012. Available at https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf, accessed January, 2020.

[10]. Plis K, Bunescu R, Marling C, Shubrook J, and Schwartz F, ‘A machine learning approach to predicting blood glucose levels for diabetes management’, in Modern Artificial Intelligence for Health Analytics: Papers Presented at the Twenty-Eighth AAAI Conference on Artificial Intelligence, pp. 35–39. AAAI Press, (2014).

[11]. Schwartz FL, Shubrook JH, and Marling CR, ‘Use of case-based reasoning to enhance intensive management of patients on insulin pump therapy’, Journal of Diabetes Science and Technology, 2(4), 603–611, (2008). [PubMed: 19885236]

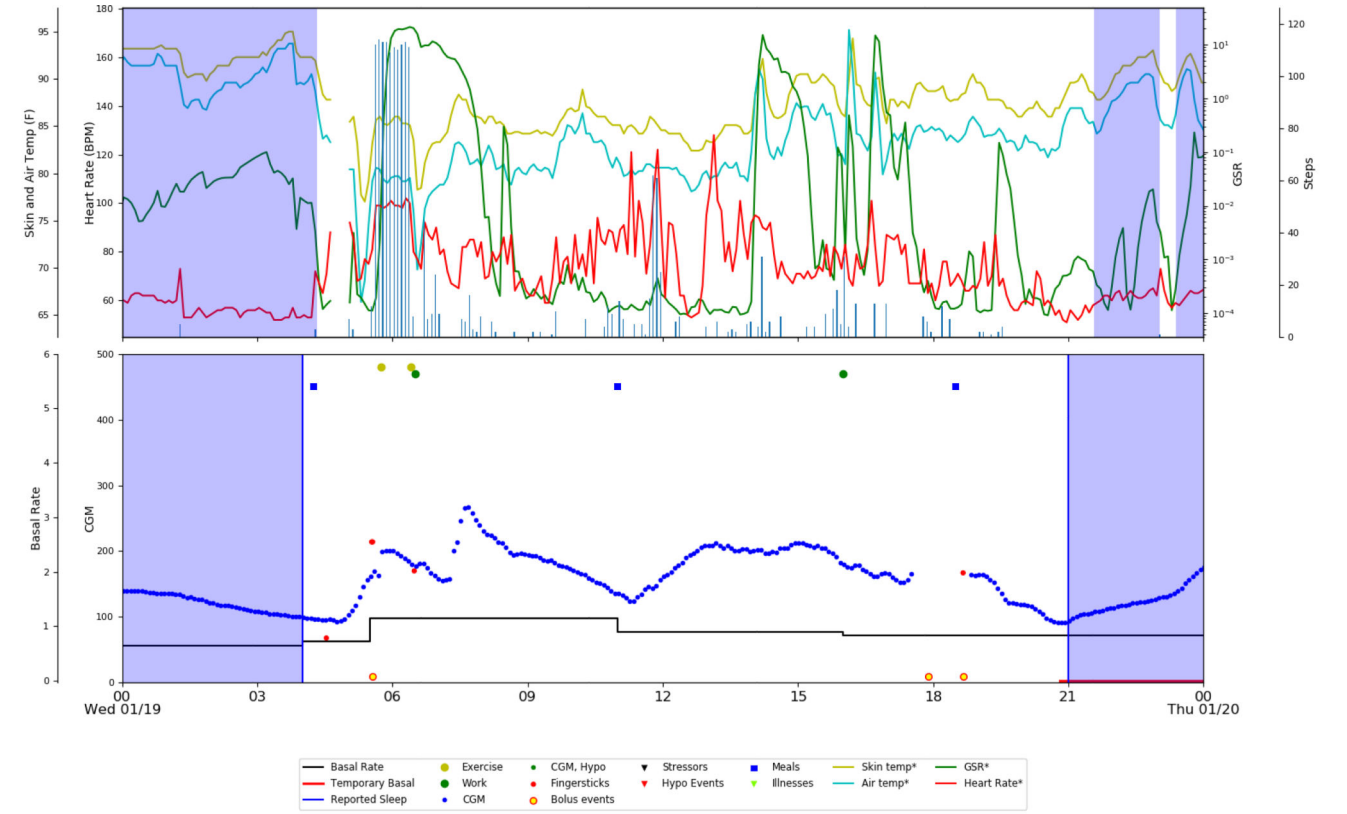

Figure 1.
Screenshot from the OhioT1DM Viewer

Table 1.

Gender, age range, insulin pump model, and sensor band type for each data contributor, by cohort

ID	Gender	Age	Pump Model	Sensor Band	Cohort
540	male	20–40	630G	Empatica	2020
544	male	40–60	530G	Empatica	2020
552	male	20–40	630G	Empatica	2020
567	female	20–40	630G	Empatica	2020
584	male	40–60	530G	Empatica	2020
596	male	60–80	530G	Empatica	2020
559	female	40–60	530G	Basis	2018
563	male	40–60	530G	Basis	2018
570	male	40–60	530G	Basis	2018
575	female	40–60	530G	Basis	2018
588	female	40–60	530G	Basis	2018
591	female	40–60	530G	Basis	2018

Table 2.

Number of training and test examples per data contributor

ID	BGLP Challenge	Training Examples	Test Examples
540	2020	11947	2884
544	2020	10623	2704
552	2020	9080	2352
567	2020	10858	2377
584	2020	12150	2653
596	2020	10877	2731
559	2018	10796	2514
563	2018	12124	2570
570	2018	10982	2745
575	2018	11866	2590
588	2018	12640	2791
591	2018	10847	2760