The OhioT1DM Dataset for Blood Glucose Level Prediction

Cindy Marling and Razvan Bunescu
School of Electrical Engineering and Computer Science
Ohio University
Athens, Ohio, 45701, USA
{marling,bunescu}@ohio.edu

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 six 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.

1 Introduction

Blood glucose level (BGL) prediction is a challenging task for AI researchers, with 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 [Juvenile Diabetes Research Foundation (JDRF), 2018] 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 six people with type 1 diabetes. All data contributors were between 40 and 60 years of age at the time of the data collection. Two were male, and four were female. All were on insulin pump therapy with continuous glucose monitoring (CGM). They wore Medtronic 530G insulin pumps and used Medtronic Enlite CGM sensors throughout the 8-week data collection period. They reported life-event data via a custom smartphone app and provided physiological data from a Basis Peak fitness band.

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 5-minute aggregations of heart rate, galvanic skin response (GSR), skin temperature, air temperature, and step count.

The rest of this paper provides background information, details the data format, describes the OhioT1DM Viewer visualization software, and tells how to obtain the OhioT1DM Dataset and Viewer for research purposes.

2 Background

We have been working on intelligent systems for diabetes management for over a decade [Schwartz *et al.*, 2008; Marling *et al.*, 2009; Marling *et al.*, 2012; Bunescu *et al.*, 2013; Plis *et al.*, 2014; Marling *et al.*, 2016; Mirshekarian *et al.*, 2017]. 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.

Throughout the years, we have received numerous requests to share the data with other researchers. 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 [Office for Civil Rights, 2012]. To protect the data and ensure that it 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

In the OhioT1DM Dataset, the data contributors are referred to by ID numbers 559, 563, 570, 575, 588 and 591. Numbers 563 and 570 were male, while numbers 559, 575, 588 and 591

were female. 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 12 XML files, two for each of the six contributors. Table 1 shows the number of training and test examples for each contributor.

Table 1: Number of Training and Test Examples per Contributor

Contributor	Training Examples	Test Examples
559	10796	2514
563	12124	2570
570	10982	2745
575	11866	2590
588	12640	2791
591	10847	2760

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 being 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 being most physically active.
14. <**basis_heart_rate**> Heart rate, aggregated every 5 minutes.
15. <**basis_gsr**> Galvanic skin response, also known as skin conductance, aggregated every 5 minutes.
16. <**basis_skin_temperature**> Skin temperature, in degrees Fahrenheit, aggregated every 5 minutes.
17. <**basis_air_temperature**> Air temperature, in degrees Fahrenheit, aggregated every 5 minutes.
18. <**basis_steps**> Step count, aggregated every 5 minutes.
19. <**basis_sleep**> Times when the Basis band reported that the subject was asleep, along with its estimate of sleep quality.

Note that, in de-identifying the dataset, all dates 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 data from the Basis Peak fitness band. Blue regions in the top pane are times that the fitness band reported that the subject was asleep. The step count is indicated by vertical blue lines. The curves show heart rate (red), galvanic skin response (green), skin temperature (gold), and air temperature (cyan).

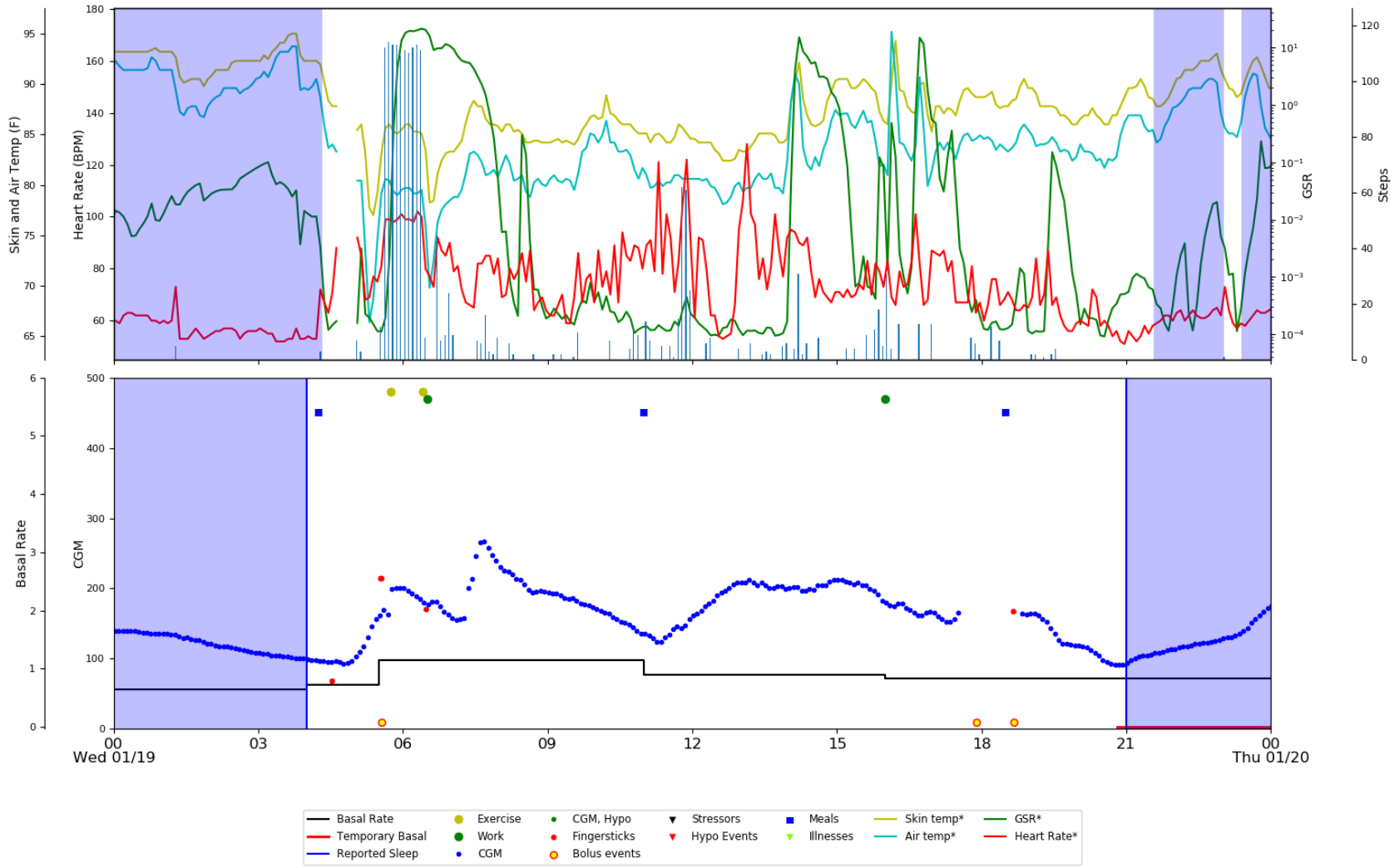

Figure 1: Screenshot from the OhioT1DM Viewer.

5 Obtaining the Dataset and Viewer

The OhioT1DM Dataset and Viewer are initially available to participants in the Blood Glucose Level Prediction (BGLP) Challenge of the Third International Workshop on Knowledge Discovery in Healthcare Data at IJCAI-ECAI 2018, in Stockholm, Sweden. After the BGLP Challenge, these resources become publicly available to other researchers. To protect the data and ensure that it 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 https://sites.google.com/view/kdhd-2018/bglp-challenge. Once a DUA is executed, the 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" analysis 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.

Acknowledgments

This work was supported by grant 1R21EB022356 from the National Institutes of Health (NIH). The OhioT1DM Viewer was implemented by Robin Kelby, based on earlier visualization software built by Hannah Quillin and Charlie Murphy. 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 anonymous individuals with type 1 diabetes who shared their data, enabling the creation of this dataset.

References

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

[Juvenile Diabetes Research Foundation (JDRF), 2018] Juvenile Diabetes Research Foundation (JDRF). Artificial Pancreas, 2018. Available at http://www.jdrf.org/research/artificial-pancreas/, accessed June, 2018.

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

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

[Marling *et al.*, 2016] C. Marling, L. Xia, R. Bunescu, and F. Schwartz. Machine learning experiments with noninvasive sensors for hypoglycemia detection. In *IJCAI 2016 Workshop on Knowledge Discovery in Healthcare Data*, New York, NY, 2016.

[Mirshekarian *et al.*, 2017] S. Mirshekarian, R. Bunescu, C. Marling, and F. Schwartz. 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'17)*, pages 2887–2891, Jeju Island, Korea, 2017.

[Office for Civil Rights, 2012] Office for Civil Rights. Guidance regarding methods for de-identification 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 June, 2018.

[Plis *et al.*, 2014] K. Plis, R. Bunescu, C. Marling, J. Shubrook, and F. Schwartz. 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*, pages 35–39. AAAI Press, 2014.

[Schwartz *et al.*, 2008] F. L. Schwartz, J. H. Shubrook, and C. R. Marling. 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.