Case-Based Reasoning as a Prelude to Big Data Analysis: A Case Study

Cindy Marling[1], Razvan Bunescu[1],
Babak Baradar-Bokaie[2] and Frank Schwartz[2]

[1] School of Electrical Engineering and Computer Science
Russ College of Engineering and Technology
Ohio University, Athens, Ohio 45701, USA
marling@ohio.edu, bunescu@ohio.edu

[2] Department of Specialty Medicine
Heritage College of Osteopathic Medicine
Ohio University, Athens, Ohio 45701, USA
babak.bokaie@gmail.com, schwartf@ohio.edu

Abstract. The 4 Diabetes Support System (4DSS) is a prototypical hybrid case-based reasoning (CBR) system that aims to help patients with type 1 diabetes on insulin pump therapy achieve and maintain good blood glucose control. The CBR cycle revolves around treating blood glucose control problems by retrieving and reusing therapeutic adjustments that have been effectively used to treat similar problems in the past. Other artificial intelligence (AI) approaches have been integrated primarily to aid in situation assessment: knowing when a patient has a blood glucose control problem and characterizing the type of problem that the patient has. Over the course of ten years, emphasis has shifted toward situation assessment and machine learning approaches for predicting blood glucose levels, as that is the area of greatest patient need. The goal has been to make large volumes of raw insulin, blood glucose and life-event data actionable. During the past year, newly available fitness bands have provided a potentially valuable source of additional data for controlling diabetes. Because it was initially unclear whether or how this new data might be leveraged, a case study was conducted, and CBR was once again called into play. This paper describes the case study and discusses the potential of CBR to serve as a prelude to big data analysis.

1 Introduction

The World Health Organization estimates that there are 347 million people living with diabetes [11]. From 5 to 10% of them have type 1 diabetes (T1D), the most severe form, in which the pancreas fails to produce insulin. T1D is neither curable nor preventable; however, it can be treated with insulin and effectively managed through blood glucose (BG) control. Good BG control helps to delay or prevent long-term diabetic complications, including blindness, amputations,

kidney failure, strokes, and heart attacks [4]. Therefore, it is important to rapidly identify and correct BG control problems.

The 4 Diabetes Support System (4DSS) is a prototypical hybrid case-based reasoning (CBR) system that detects and predicts BG control problems and suggests personalized therapeutic adjustments to correct them. The 4DSS has been extensively described in the literature [6, 7, 9]; a brief system overview is presented in the next section. A critical research thrust that grew out of work on the 4DSS is how to continuously, in real-time, predict that a BG control problem is about to occur. The key is to be able to accurately predict what the BG level will be in the next 30 to 60 minutes, which would allow enough time to intervene and prevent predicted problems. Large volumes of raw insulin and BG data are available for analysis. Machine learning algorithms for time series prediction have been efficaciously applied. Studies conducted on retrospective data show that our system predicts BG levels comparably to physicians specializing in diabetes care, but not yet well enough for use by patients in the real world [2, 8].

Recently, commercially available fitness bands and smart watches, such as the Basis Peak, Nike Fuelband, Fitbit, and Apple Watch, have made it practical to inexpensively and unobtrusively collect large quantities of physiological data. As this data may be indicative of patient activity impacting BG levels, it could potentially be used to improve BG level prediction. However, due to the complicated nature of the problem, it was not initially clear whether or how this data could be leveraged. Therefore, a case study was conducted in which a patient with T1D wore a fitness band in addition to his usual medical devices for two months. The data was consolidated and displayed via custom visualization software. The patient, his physician, and artificial intelligence (AI) researchers met weekly to review and interpret the data, using a protocol like that employed to build the 4DSS. This case-based focus shed light on how the new data could be integrated into machine learning models and leveraged to improve BG prediction. We posit that a case-based approach is especially useful in dealing with new data sources, new patients, and new medical conditions, and that early lessons learned through the CBR process can aid in later big data analysis.

2 Background

This section briefly describes the 4DSS and the work on machine learning models for blood glucose prediction prior to the new case study.

2.1 The 4 Diabetes Support System

A graphical overview of the prototypical hybrid CBR system is shown in Figure 1. The patient provides BG, insulin and life-event data to the system. BG and insulin data are uploaded from the patient's prescribed medical devices. The patient enters data about life events that impact BG levels, such as food, exercise, sleep, work, stress and illness, using a smart phone. The data is scanned by

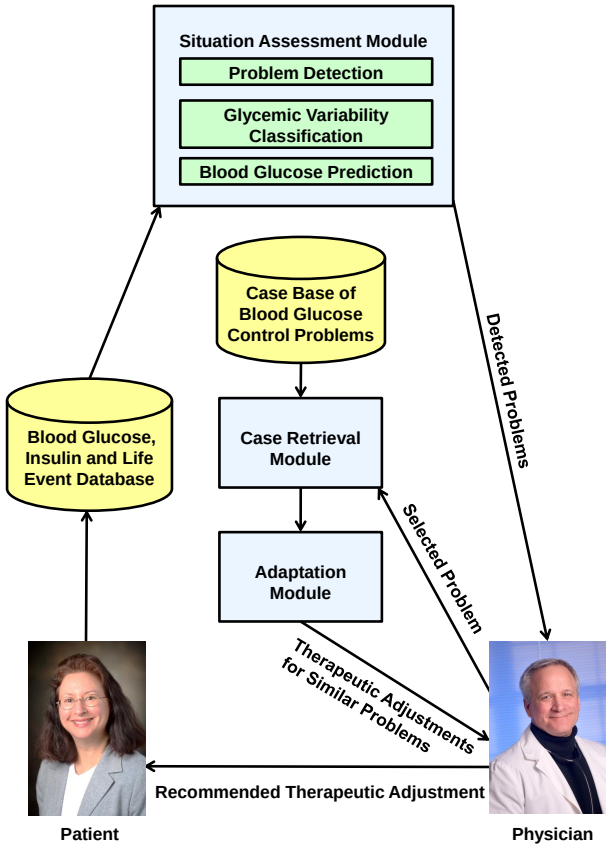

Fig. 1. Overview of the 4 Diabetes Support System, reproduced from [6]

the situation assessment module, which detects and predicts BG control problems. The most critical types of problems are: hyperglycemia, or high BG, which contributes to long-term diabetic complications; and hypoglycemia, or low BG, which may result in severe immediate reactions, including weakness, dizziness, seizure or coma. The situation assessment module reports detected problems to the physician. The physician selects a problem of interest, which is then used by the case retrieval module to obtain the most similar case or cases from the case base. Each retrieved case contains a specific BG control problem experienced by a T1D patient, a physician's recommended therapeutic adjustment, and the clinical outcome for the patient after making the therapeutic adjustment. Retrieved cases go to the adaptation module, which personalizes a retrieved solution to fit the specific needs of the current patient. A solution is a therapeutic adjustment comprising one or more actions that a patient can take. Adapted therapeutic adjustments are displayed to the physician as decision support. The physician

decides whether or not to recommend the therapeutic adjustments to the patient. Directly providing suggestions to the patient, while a long-term goal, would require regulatory approval.

2.2 Machine Learning Models for Blood Glucose Prediction

Originally conceived of as important for situation assessment, BG prediction has multiple potential applications that could enhance safety and quality of life for people with T1D if incorporated into medical devices. These applications include: alerts to warn of imminent problems; decision support for taking actions to prevent impending problems; "what if" analysis to project the effects of lifestyle choices on BG levels; and integration with closed-loop control algorithms for insulin pumps (aka the "artificial pancreas"). Predicting hypoglycemia is especially important, both for patient safety and because hypoglycemia is a limiting factor for intensive insulin therapy [3].

In our BG prediction approach, a generic physiological model is used to generate informative features for a Support Vector Regression (SVR) model that is trained on patient specific data. The physiological model characterizes the overall dynamics into three compartments: meal absorption, insulin dynamics, and glucose dynamics. The parameters of the physiological model are tuned to match published data and feedback from physicians. To account for the noise inherent in the data, the state transition equations underlying the continuous dynamic model are incorporated in an extended Kalman filter.

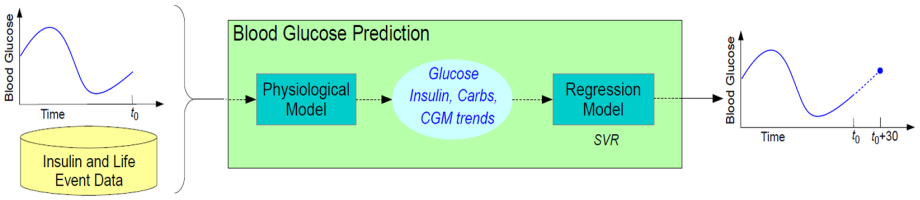

Fig. 2. Overview of the Blood Glucose Level Prediction Process

Figure 2 shows the overall BG level prediction process. A continuous dynamical system implementing the set of physiological equations is run in prediction mode for 30 and 60 minutes. Physiological model predictions are then used as features for an SVR model that is trained on the two weeks of data preceding the test point. Furthermore, an AutoRegressive Integrated Moving Average (ARIMA) model is trained on the same data and its predictions are used as additional features. The models are trained to minimize Root Mean Square Error (RMSE). SVR predictions are made at 30 and 60 minute intervals and compared to BG levels at prediction time (t_0), ARIMA predictions, and predictions made by physicians specializing in diabetes care. Results are shown in Figure 3.

Horizon	t_0	ARIMA	Physician	SVR
30 Min	27.5	22.9	**19.8**	**19.5**
60 Min	43.8	42.2	**38.4**	**35.7**

Fig. 3. RMSE of the Best SVR Model vs. the t_0, ARIMA, and Physician Predictions

3 The Case Study

The recent proliferation of commercially available fitness bands provides an opportunity to exploit new data from inexpensive, unobtrusive, portable physiological sensors. These sensors provide signals indicative of activities that are known to impact BG levels, including sleep, exercise and stress. The hope is that, by incorporating these signals, we can obtain a more accurate picture of patient activity, while reducing or eliminating the need for the patient to self-report life events. We conducted an N-of-1 study in order to learn whether or how this influx of new data could be leveraged to advantage.

The subject was a middle-aged physician who has had T1D since childhood. For two months, he wore a fitness band along with his regularly prescribed medical devices and entered life-event data via his smart phone. The fitness band, a Basis Peak, provided data for galvanic skin response (GSR), heart rate, and skin and air temperatures. The medical devices, a Medtronic insulin pump and a Dexcom continuous glucose monitoring (CGM) system, provided insulin and BG data. All of this data was consolidated in the 4DSS database.

Once a week, the subject met with his physician and AI researchers to review and analyze the data. The consolidated data was displayed via custom-built visualization software called PhysioGraph. A screen shot from PhysioGraph, showing the different types of data, is shown in Figure 4. BG control problems identified by the 4DSS software, the subject, and/or his physician, were visualized and discussed. We looked for visual patterns in the fitness band data during the times when the problems occurred.

Preliminary findings based on these visualizations were encouraging. While even subtle patterns may be detected by machine learning algorithms, we were able to detect some marked patterns as humans. The most pronounced pattern was a rise in GSR with severe hypoglycemia. The most interesting pattern revolved around shoveling snow. The study was conducted during an unusually harsh winter in which the patient (and most of the rest of us) had to frequently shovel heavy snow. Shoveling snow is strenuous exercise, and exercise is known to lower BG levels. After shoveling for extended periods, the subject sometimes experienced hypoglycemia. This is a problem we would like to predict, because, if alerted, the patient could take action to prevent it. There was a discernable pattern in the fitness band data surrounding this problem. GSR and heart rate rose, while skin and air temperature dropped. While we do not yet know if this

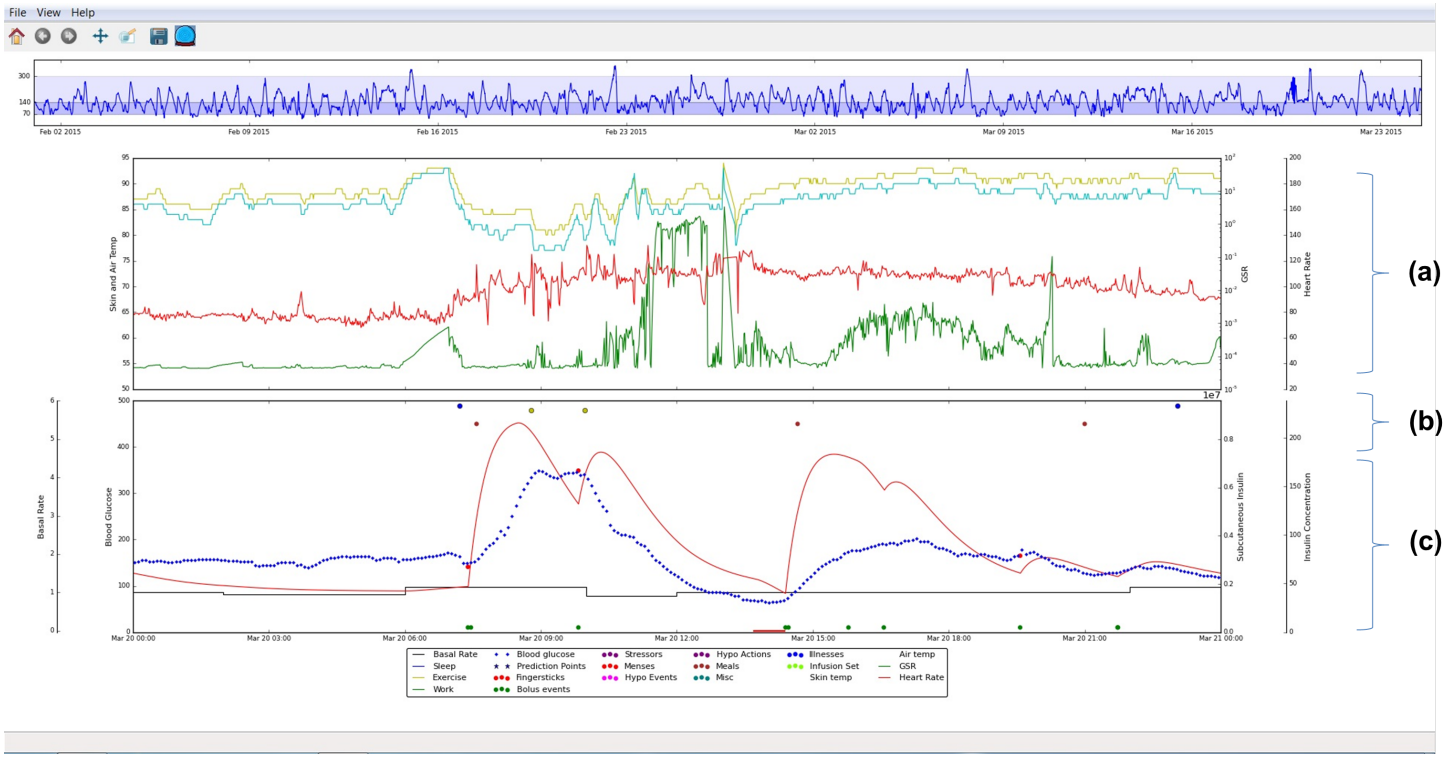

Fig. 4. A Screen Shot from PhysioGraph. Data from the fitness band is displayed in section (a). This includes GSR (green), heart rate (red), skin temperature (gold) and air temperature (cyan). Icons indicating life events appear in section (b). These icons are clickable in PhysioGraph to view event details. BG data and insulin data are shown in section (c). The dotted curve (blue) represents CGM data, while the solid curve (red) models the total insulin in the patient's system.

combination of signals will allow us to predict hypoglycemia or only detect it, we now have leads to follow.

4 Discussion

When work began on the 4DSS in 2004, we had little data and no structured cases. As we built our case base and collected data from over 50 T1D patients, we developed a database that enabled us to build machine learning models for purposes we did not envision in 2004. With more data, it becomes possible to leverage more AI approaches for more purposes. When data or knowledge is limited, however, CBR can be an enabling approach. As non-health-related examples, we can think of leveraging all of the information possible from a single oil spill, or of basing product recommendations for a customer with no purchase history on what similar customers have bought. In the medical arena, CBR has also proven useful for dealing with new situations. For example, CARE-PARTNER used CBR to determine appropriate follow-up care for the earliest stem cell transplant patients [1]. Once many patients had undergone stem cell transplantation and received follow-up care, their collective experiences were distilled into clinical practice guidelines that were used in lieu of CBR. Today, nearly 20 years later, big data tools like IBM Watson Health [5] may allow us to further evaluate, refine, and personalize treatment for these patients.

In the diabetes domain, big data is not yet publicly available; however, we anticipate its near-term future availability. Three non-technical factors contribute to this lack of data: (1) most diabetes patients do not yet wear devices or use systems that continuously collect data; (2) medical device manufacturers do not yet allow access to raw data in real-time, but require the use of their own proprietary software; and (3) patient privacy concerns inhibit data sharing, even when data exists. There has been a recent drive to collect and consolidate data from all T1D patients and all types of (currently incompatible) medical devices, spearheaded by the non-profit organization Tidepool [10]. A goal is to be able to analyze and leverage continuous data from hundreds of thousands of patients to improve diabetes care and outcomes for individuals. Our case engineering, based on the limited data we have already collected, could serve to jump start such efforts.

At the heart of any CBR system is the case. The case is a knowledge structure that, especially in complex medical domains, may embody more than a collection of readily available feature-value pairs. The design of a case for a CBR system begins with the analysis of real-world cases to identify problems, solutions and outcomes. It is necessary to understand and define the features that make cases similar to each other, reusable in different circumstances, and adaptable to the case at hand. Features engineered for cases may provide machine learning algorithms with better inputs than raw data or surface features.

In our case study, the fitness band provided physiological signals that were not a part of our original case design. There were 20 times as many data points per patient per day as we had previously collected. We did not know how the

new data could be used to anticipate or detect blood glucose control problems or, more fundamentally, how it related to BG levels in people with T1D. The inputs to our SVR models are not raw data points, but features that have been carefully engineered from the raw data based, in part, on insights gained from cases. Sometimes, simple data combinations suffice; for example, we could see during the case study that the difference between air and skin temperature was more relevant than either individual measurement. Other times, we have had to employ complex systems of equations; for example, a complex physiological model is needed to characterize the impact of insulin on BG levels. Even as we move toward more automated means of feature engineering and more reliance on machine learning techniques, early use of CBR can help to provide insight and intuition that may guide big data exploration and interpretation.

5 Summary and Conclusion

The 4DSS is a prototypical hybrid CBR system that aims to help T1D patients achieve and maintain good BG control. As cases and data have accumulated over ten years, the research emphasis has shifted toward using the accumulated data to build machine learning models for BG prediction. While these models have applicability to situation assessment within the 4DSS, their greater potential is in facilitating a wide range of practical applications that could enhance safety and quality of life for T1D patients. The recent proliferation of commercially available fitness bands has presented the opportunity to incorporate new types of data indicative of patient activity into the models to improve prediction accuracy. However, when it was initially unclear whether or how this new data might be leveraged, a case study was conducted, calling CBR back into play.

In the N-of-1 study, a T1D patient on insulin pump therapy with continuous glucose monitoring wore a fitness band and entered life-event data into the 4DSS database for two months. The aggregated data was displayed via custom-built visualization software and reviewed at weekly intervals by the patient, his physician, and AI researchers. BG control problems were analyzed with a focus on identifying patterns in the new data at the time the problems occurred. Some promising patterns could be visualized, including a marked rise in GSR with severe hypoglycemia. This case-based focus provided insight and intuition about how the new data relates to BG levels. Work on integrating the new data into our BG prediction models is currently underway.

6 Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1117489. Additional support was provided by Medtronic and Ohio University. We gratefully acknowledge the contributions of diabetologist Jay Shubrook, DO, graduate research assistants Kevin Plis and Samson Xia, and Research Experience for Undergraduates (REU) participants Hannah Quillin and Charlie Murphy.

References

1. Bichindaritz, I., Kansu, E., Sullivan, K.M.: Case-based reasoning in CAREPARTNER: Gathering evidence for evidence-based medical practice. In: Smyth, B., Cunningham, P. (eds.) Advances in Case-Based Reasoning: 4th European Workshop, Proceedings EWCBR-98. pp. 334–345. Springer-Verlag, Berlin (1998)
2. Bunescu, R., Struble, N., Marling, C., Shubrook, J., 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)
3. Cryer, P.E.: Hypoglycemia: Still the limiting factor in the glycemic management of diabetes. Endocrine Practice 14(6), 750–756 (2008)
4. Diabetes Control and Complications Trial Research Group: The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. New England Journal of Medicine 329(14), 977–986 (1993)
5. IBM: IBM Watson Health (2015), http://www.ibm.com/smarterplanet/us/en/ibmwatson/health/, accessed July, 2015
6. Marling, C., Bunescu, R., Shubrook, J., Schwartz, F.: System overview: The 4 Diabetes Support System. In: Lamontagne, L., Recio-García, J.A. (eds.) Workshop Proceedings of the Twentieth International Conference on Case-Based Reasoning. pp. 81–86. Lyon, France (2012)
7. Marling, C., Wiley, M., Bunescu, R., Shubrook, J., Schwartz, F.: Emerging applications for intelligent diabetes management. AI Magazine 33(2), 67–78 (2012)
8. Plis, K., Bunescu, R., Marling, C., Shubrook, J., 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)
9. Schwartz, F.L., Shubrook, J.H., Marling, C.R.: 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)
10. Tidepool: The Tidepool Platform: A home for diabetes data (2015), http://tidepool.org/platform/, accessed August, 2015
11. World Health Organization: 10 facts about diabetes (2014), http://www.who.int/features/factfiles/diabetes/facts/en/, accessed July, 2015