33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Predicting Human Subcutaneous Glucose Concentration in Real Time: A Universal Data-Driven Approach

Yinghui Lu, Srinivasan Rajaraman, W. Kenneth Ward, Robert A. Vigersky, and Jaques Reifman

Abstract—**Continuous glucose monitoring (CGM) devices measure and record a patient's subcutaneous glucose concentration as frequently as every minute for up to several days. When coupled with data-driven mathematical models, CGM data can be used for short-term prediction of glucose concentrations in diabetic patients. In this study, we present a real-time implementation of a previously developed offline data-driven algorithm. The implementation consists of a Kalman filter for real-time filtering of CGM data and a data-driven autoregressive model for prediction. Results based on CGM data from 3 different studies involving 34 type 1 and 2 diabetic patients suggest that the proposed real-time approach can yield ~10-min-ahead predictions with clinically acceptable accuracy and, hence, could be useful as a tool for warning against impending glucose deregulation episodes. The results further support the feasibility of "universal" glucose prediction models, where an offline-developed model based on one individual's data can be used to predict the glucose levels of any other individual in real time.**

I. Introduction

MODERN continuous glucose monitoring (CGM) devices provide a minimally invasive mechanism for monitoring the glycemic state of a patient as frequently as every minute. However, the ability of such devices to monitor current glucose concentrations can provide alerts only when a interstitial glucose excursion is already underway (i.e., the glucose concentration may already be at an unacceptably high or low level) rather than alerting the patients of an impending glucose excursion so that a proactive therapy can be rendered.

Recently, we demonstrated the feasibility of data-driven autoregressive (AR) models to predict the near-future glucose concentrations of type 1 diabetic patients using their recorded recent-history CGM data [1], where we offline-smoothed the CGM data of a patient, developed an AR model using a portion of the offline-smoothed data, and then predicted the remaining portion of the smoothed data with the AR model. We found that, for a 30-min-ahead prediction horizon, the AR models yielded predictions with an average root mean squared error (RMSE) of 1.8 mg/dL and an average time lag of 0.2 min. More recently, we extended the aforementioned approach to type 2 diabetic patients as well [2]. Importantly, we found that a "universal" AR model could be developed based on the CGM data from one diabetic patient, and subsequently applied to predict subcutaneous glucose concentrations of other patients, without being affected by diabetes type, subject age, CGM device, and interindividual differences. However, in these two studies, the AR models were applied to predict CGM data whose entire time series had been previously (i.e., offline) smoothed. In real-time prediction, this is not possible, because the entire time-series data are not available *a priori* and only previous and current data values are known at any given time.

In this report, we describe an algorithm for predicting subcutaneous glucose concentrations in real time. The proposed real-time prediction algorithm consisted of two components: a Kalman filter and an AR model. First, we filtered the raw CGM data in real time through the Kalman filter. Then, we predicted future glucose concentrations based on the real-time filtered CGM data using the AR model. In addition, the linear AR formulation rendered analytic expressions for computing statistically based reliability measures of the predictions in the form of 95% prediction intervals (PIs).

II. Methods

A. Study Population

Table I shows the information of the three independent studies used in this investigation. Detailed information of the protocols is provided in Ref. 2. Briefly, the 3 studies included 34 subjects with either type 1 or type 2 diabetes and using three different CGM devices: iSense, Guardian RT, and DexCom. In the iSense study, subjects were included if they were between 18 and 70 yr of age, had been diagnosed with type 1 diabetes and treated with insulin for at least 12 mo, and had glycated hemoglobin (HbA1c) >6.1%. In the Guardian study, subjects were included if they were between 3 and 7 yr old or between 12 and 18 yr old, had been diagnosed with type 1 diabetes for more than 1 yr, had been using an insulin pump, and had HbA1c ≤ 10.0%. In the DexCom study, subjects were included if they were older than 18 yr of age, had been diagnosed with type 2 diabetes and treated either with oral agents, basal insulin, or both for at least 3 mo, and had HbA1c between 7% and 12%.

This work was supported in part by the U.S. Army Medical Department, Advanced Medical Technology Initiative, funded by the Telemedicine and Advanced Technology Research Center (TATRC) of the U.S. Army Medical Research and Materiel Command (USAMRMC), Fort Detrick, Maryland, and by the U.S. Air Force Diabetes Research Program.

Y. Lu is with the Bioinformatics Cell (BIC), TATRC, USAMRMC, Fort Detrick, MD 21702 USA (e-mail: ylu@bioanalysis.org).

S. Rajaraman is with the BIC, TATRC, USAMRMC, Fort Detrick, MD 21702 USA (e-mail: srini@bioanalysis.org).

W. K. Ward is with Oregon Health and Sciences University and with Legacy Health System, both in, Portland, OR 97239 USA (e-mail: kenward503@msn.com).

R. A. Vigersky is with the Diabetes Institute, Walter Reed Army Medical Center, Washington, DC 20307 USA (e-mail: robert.vigersky@amedd.army.mil).

J. Reifman is a Senior Research Scientist and Director of the BIC, TATRC, USAMRMC, ATTN: MCMR-TT, 504 Scott Street, Fort Detrick, MD 21702 (corresponding author; phone: 301-619-7915; fax: 301-619-1983; e-mail: jaques.reifman@us.army.mil).

TABLE I
SUMMARY INFORMATION OF THE THREE STUDIES

	iSense	Guardian RT	DexCom
No. of Subjects	9	18	7
Diabetes Type	1	1	2
Sampling Interval (min)	1	5	5
Collection Time (days)	5	6	56

We used the first 4,000 min of the recorded CGM data of each subject and down-sampled the data of the iSense study to 5-min sampling intervals so that every data set contained 800 data points. We used the first 400 data points (or 2,000 min) as training data for AR model development and the subsequent 2,000 min as testing data for assessing the real-time predictions.

B. Offline Smoothing of the Training Data

Because AR models capture the temporal autocorrelation in the time-series CGM data, the data need to be smoothed prior to model development. Otherwise, AR models fitted on raw CGM data yield trivial, random-walk models [1].

Here, similar to our previous work [1-3], we offline smoothed the training data of each subject using the Tikhonov regularization approach, where the regularization parameter was chosen such that signals with period less 1 h were removed from raw CGM data. These offline-smoothed training data were used as the "noise-free" CGM data for AR model development.

C. Real-time Filtering of the Testing Data

In our previous studies [1,2], we offline smoothed the entire CGM data set (training and testing data). However, for real-time prediction, offline smoothing of the testing data is not possible because at a given time only previous and current data points are available. In this report, we filtered the testing data of each subject in real time using a Kalman filter to simulate the scenario of real-time predictions. The Kalman filter was formulated in the discrete-time domain, using an AR-model state-space representation, described as:

$$\tilde{Y}(n) = A \cdot \tilde{Y}(n-1) + W(n) \tag{1}$$

$$y(n) = H \cdot \tilde{Y}(n) + v(n) \tag{2}$$

with

$$\tilde{Y}(n) = \begin{bmatrix} \tilde{y}(n) \\ \tilde{y}(n-1) \\ \vdots \\ \tilde{y}(n-m+1) \end{bmatrix} \quad A = \begin{bmatrix} b_1 & b_2 & \cdots & b_m \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 1 \end{bmatrix} \quad W(n) = \begin{bmatrix} \varepsilon(n) \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

and

$$H = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}$$

where $\tilde{y}(n)$ denotes the noise-free CGM data at time n, $\tilde{Y}(n)$ is the state vector, A is the m x m state transition matrix consisting of AR-model coefficients, b_i, $i = 1, 2, \ldots, m$, where m is the order of the AR model, $W(n)$ is the process noise having white noise $\varepsilon(n)$ with zero mean and variance σ_ε^2, $y(n)$ is the raw CGM data, H is the matrix that denotes the relationship between the state vector and the raw CGM data, and $v(n)$ denotes white measurement noise with zero mean and variance σ_v^2. In this formulation, the raw CGM data $y(n)$ are regarded as the measurements that contain noise $v(n)$, and the noise-free CGM data $\tilde{y}(n)$ are considered equivalent to the offline-smoothed CGM data, which need to be estimated by the Kalman filter in real time.

At each discrete time n, the Kalman filter yields the optimal (in the minimum mean-squared-error sense) estimation of the state vector $\tilde{Y}(n)$ through the following prediction and update formulation [4]:

$$\hat{\tilde{Y}}^-(n) = A \cdot \hat{\tilde{Y}}(n-1) \tag{3}$$

$$\hat{\tilde{Y}}(n) = \hat{\tilde{Y}}^-(n) + K \cdot [y(n) - H \cdot \hat{\tilde{Y}}^-(n)] \tag{4}$$

where $\hat{\tilde{Y}}^-(n)$ denotes the prior estimate, $\hat{\tilde{Y}}(n)$ denotes the optimal estimate of $\tilde{Y}(n)$, and K denotes the Kalman gain, which is determined by σ_ε^2 and σ_v^2. Given $\tilde{Y}(n)$, the optimal estimate of the noise-free CGM data, at time n, is computed as $\hat{\tilde{y}}(n) = H \cdot \hat{\tilde{Y}}(n)$.

In real-time prediction, outliers of the raw CGM data were corrected on the fly before being fed into the Kalman filter, i.e., if the difference between the current data value $y(n)$ and the previous one $y(n–1)$ was more than 4 mg/dL, we limited the rate of change of blood glucose to ± 4 mg/dL [5].

As represented in Eqs. (1)-(4), the proposed Kalman filter smoothing approach requires two components: an AR model that is fitted on the offline-smoothed CGM data and the corresponding Kalman filter parameters σ_ε^2 and σ_v^2. These two components were obtained after the completion of AR model development using the training data, as described below.

D. AR Modeling and Prediction

The AR model was fitted on the offline-smoothed training data. Equation (1) provides the state-space representation of the AR model, which is equivalent to the conventional AR-model expression, that is,

$$\tilde{y}(n) = \sum_{i=1}^{m} b_i \tilde{y}(n-i) + \varepsilon(n). \tag{5}$$

The AR coefficients b_i describe the temporal correlation between the current value $\tilde{y}(n)$ and each of the previous values $\tilde{y}(n-i)$, $i = 1, 2, \ldots, m$, and $\varepsilon(n)$ denotes white noise [6].

Previously, we calculated the coefficients b_i using regularized least squares to obtain a regularized AR model such that the model was able to yield stable predictions for the scenario where white noise was added to the offline smoothed testing data [1]. However, in real-time situations, such a scenario does not exist because CGM data noise is filtered out by the Kalman filter in real time before making AR-model predictions. Therefore, here we calculated the coefficients b_i using ordinary least squares.

The order of the AR model m was determined using the Bayesian Information Criterion (BIC), which balances the goodness of the model fit with model complexity [6].

Using the optimal estimates of the noise-free data and the

AR coefficients, the one-step-ahead AR prediction can be calculated as follows:

$$\hat{\tilde{y}}(n+1) = \sum_{i=0}^{m-1} b_i \hat{\tilde{y}}(n-i), \tag{6}$$

where $\hat{\tilde{y}}(n+1)$ denotes the predicted value for $\tilde{y}(n+1)$. Equation (6) can also be used to make *k*-step-ahead predictions, $\hat{\tilde{y}}(n+k)$, for $k>1$, by iteratively substituting the (k–1) predicted values for the corresponding (k–1) yet-unknown data values.

In addition, the linear AR formulation provides the following analytic expression for computing PIs at time step $n+k$ [6]:

$$PI(n+k) = \hat{\tilde{y}}(n+k) \pm 1.96 \cdot \sqrt{[(1+\sum_{j=1}^{k} \psi_j^2) \cdot \sigma_\varepsilon^2 + \sigma_v^2]}, \tag{7}$$

where ψ_j are the AR coefficients in the infinite weighted sum expression, which can be derived from b_i [6], and the coefficient 1.96 corresponds to 95% limits. The interval $PI(n+k)$ provides the statistical range within which the *k*-step-ahead raw CGM data point $y(n+k)$ shall fall 95% of the time.

The Kalman filter parameters σ_ε^2 and σ_v^2 were estimated using the offline-smoothed training data. Using the trained AR model, we set σ_ε^2 to be the variance of the one-step-ahead prediction error (applied on the offline-smoothed training data) and σ_v^2 to be the variance of the residual error between the raw and offline-smoothed training data.

III. Results

A. AR Model and Kalman Filter Parameters

We applied the BIC to the offline-smoothed training data for each of the 34 subjects and found that the optimal AR order was 6 for 20 subjects, 7 for 12 subjects, and 8 for two subjects. Therefore, we set the AR order as 6. We then fitted such models using the offline-smoothed training data, to obtain 34 AR(6) models and 34 pairs of Kalman filter parameters σ_ε^2 and σ_v^2. Table II shows the mean and standard deviation (SD) of the six model coefficients b_i, i = 1, 2,…, 6, and the ratio $\sigma_v^2/\sigma_\varepsilon^2$.

TABLE II
Mean and Standard Deviation (SD) of the Autoregressive (AR) Coefficients and the Ratio $\sigma_v^2/\sigma_\varepsilon^2$ for the 34 Continuous Glucose Monitoring (CGM) Signals

	AR Coefficients						$\sigma_v^2/\sigma_\varepsilon^2$
	b_1	b_2	b_3	b_4	b_5	b_6	
Mean	4.96	-10.63	12.62	-8.78	3.39	-0.57	2634.62
SD	0.29	1.12	2.09	1.88	0.87	0.17	455.02

B. Real-Time Prediction

We used the testing data to assess the (simulated) real-time predictive performance of the 34 AR models for 2 prediction horizons: 10- and 20-min-ahead. The model's performance was assessed based on RMSE and time lag. First, we offline smoothed the testing data using the same Tikhonov regularization technique as used to smooth the training data. Then, we used the smoothed testing data as the reference against which we computed the RMSE and time lag of the predictions, where the time lag was calculated based on the cross-correlation between the predictions and the reference. Both RMSE and time lag were computed from 2,250 to 4,000 min, where were excluded the initial (2,000-2,250 min) transient response of the Kalman filtering.

To evaluate the performance of each of the 34 individual models, we used the AR(6) model and the corresponding σ_ε^2 and σ_v^2 obtained from the training data of the corresponding subject (individualized model and parameters). Figure 1 shows the 10- and 20-min-ahead prediction results for a typical subject, Guardian *subject* #9. As shown in Fig. 1, as the prediction horizon increased, the RMSE, time lag, and PIs increased. The time lag increased to ~10 min for the 20-min-ahead prediction, indicating that the effective horizon for real-time predictions is ~10 min.

To verify the portability of the AR(6) models and as the impact of the Kalman filter parameters on the predictions, we conducted a leave-one-out (LOO) prediction procedure, where we predicted the testing data of each of the 34 subjects using the averaged AR model coefficients and the averaged Kalman filter parameters based on the training data of the other 33 subjects. Figure 2 shows the comparison of the 10-min-ahead predictions for Guardian *subject* #9 using the LOO AR model and Kalman filter parameters with those based on individualized model and parameters for that subject. We found the results to be essentially

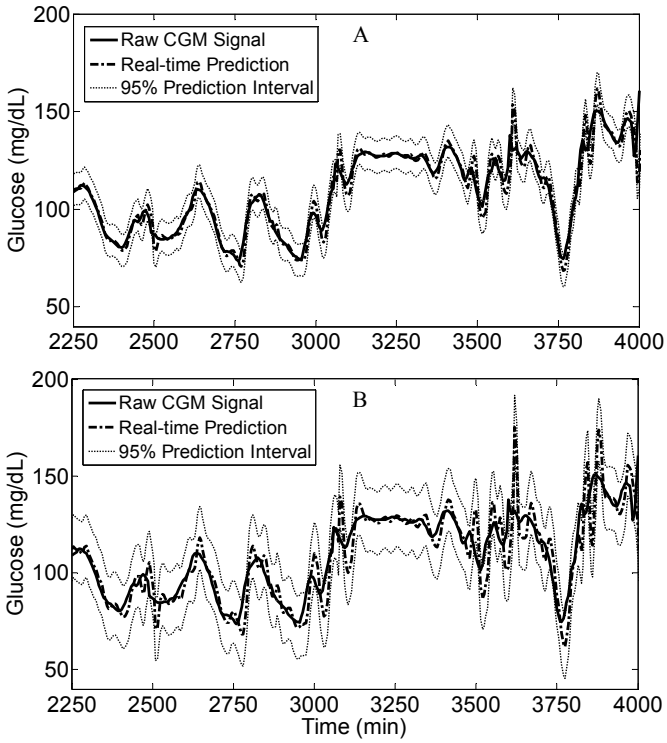

Fig. 1. Simulated real-time prediction for Guardian *subject* #9, based on the subject's individualized autoregressive (AR) model and parameters σ_ε^2 and σ_v^2. A: 10-min-ahead prediction [root mean squared error (RMSE) = 5.22 mg/dL and delay = 0 min. B: 20-min-ahead prediction (RMSE = 8.94 mg/dL and delay = 10 min). CGM, continuous glucose monitoring.

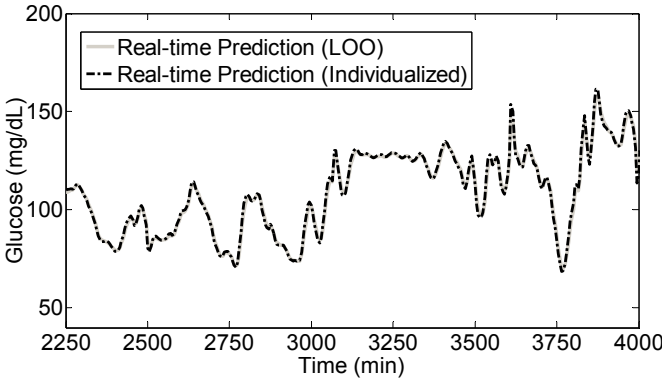

Fig. 2. 10-min-head simulated real-time predictions for *Guardian* #9. A comparison of the individualized AR model and parameters σ_ε^2 and σ_v^2 versus the leave-one-out (LOO) AR model and parameters is shown. Individual case: RMSE = 5.22 mg/dL and delay = 0 min; LOO case: RMSE = 4.70 mg/dL and delay = 0 min.

indistinguishable.

Table III shows the overall results of the 10- and 20-min-ahead predictions using individualized and LOO AR models and Kalman filter parameters for the 34 subjects. The results in Table III show that the average RMSE of the predictions using individualized and LOO AR models and Kalman filter parameters were arguably close. The entries in parentheses show the average prediction RMSE during the hypoglycemic events as defined in [2]. Furthermore, when we performed pair-wise comparisons of the RMSEs of the predictions generated by the two approaches using a paired-sample *t*-test [7], we found that, at the 5% significance level, the results were statistically indistinguishable for both the 10- and 20-min-ahead predictions.

TABLE III
PREDICTIVE PERFORMANCE (AVERAGED OVER THE 34 CGM SUBJECTS)

	Individualized AR(6) and Parameters σ_ε^2 and σ_v^2		LOO AR(6) and Parameters σ_ε^2 and σ_v^2	
Prediction Horizon (min)	RMSE (mg/dL)	Lag (min)	RMSE (mg/dL)	Lag (min)
10	8.97 (9.93)	2.50	8.97 (12.97)	1.76
20	16.06 (13.46)	9.26	15.69 (15.32)	9.56

IV. DISCUSSION AND CONCLUSIONS

Using the proposed Kalman filter/AR model approach, we were able to predict CGM data 10-min-ahead in real time with an average time lag of 2.5 min and an average RMSE of 8.97 mg/dL, which is ~6% of the mean value of a CGM data (the averaged mean value over the CGM data of the 34 subjects is 159.06 mg/dL). In addition, the proposed approach also provided statistically based PIs, which provide a measure of reliability of the model predictions. Using the Clarke error grid analysis [8] based on the predictions and the reference data, we noted that the simulated real-time predictions were clinically acceptable (98.6% in zone A and 1.4% in zone B, calculated for the predictions using individualized AR model and Kalman filter parameters; plots are not shown). However, the RMSE was ~5 times larger than those of the offline predictions achieved in our previous work [1], and the average prediction time lag increased from 0.2 to 2.5 min.

The degraded performance was caused by the so-called "end effect" observed in real-time filtering [6]. In real time, the filter can only use data up to the current discrete time *n* to filter the data at *n*, as opposed to offline, where the data beyond time *n* are available to improve the smoothing at time *n*. The end effect, which is particularly problematic in predicting oscillatory data, creates special and unique challenges for predictive algorithms, as the most recent samples, which carry the majority of the predictive information, cannot be properly filtered. The end effect may be alleviated by "borrowing" data from the near future, but that may not necessarily improve the predictive performance because borrowing data causes an effective increase in the prediction horizon.

The LOO simulation results support our previous observations regarding the universality of AR-based predictive models [2]. As in the offline case, the penalty for developing one AR model (based on one individual or averaged over a group of individuals) and applying it for the real-time prediction of glucose levels of an unseen individual is arguably very small. For long-term use of CGM devices, the variance of the measurement noise, σ_v^2, may change with time and its value may need to be updated. In this case, an adaptive Kalman filter may be used, where σ_v^2 is estimated in each iteration step of the Kalman filtering [4]. The corroboration of model universality for real-time predictions has important practical implications, as it provides direct support for the incorporation of predictive models with CGM devices in the development of a shrink-wrapped system.

DISCLAIMER

The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the U.S. Army or of the U.S. Department of Defense. This paper has been approved for public release with unlimited distribution.

REFERENCES

[1] A. Gani, A. V. Gribok, S. Rajaraman, W. K. Ward, and J. Reifman, "Predicting subcutaneous glucose concentration in humans: data-driven glucose modeling," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 2, pp. 246-254, Feb, 2009.

[2] A. Gani, A. Gribok, Y. Lu, W. K. Ward, R. A. Vigerksy, and J. Reifman, "Universal glucose models for predicting subcutaneous glucose concentration in humans," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 1, pp. 157-165, Jan, 2010.

[3] Y. Lu, A. V. Gribok, W. K. Ward, and J. Reifman, "The importance of different frequency bands in predicting subcutaneous glucose concentration in type 1 diabetic patients," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 8, pp. 1839-1846, Apr, 2010.

[4] Analytic Sciences Corporation. Technical Staff and A. Gelb, *Applied Optimal Estimation*. Cambridge, Mass. M.I.T. Press, 1974

[5] B. P. Kovatchev, W. L. Clarke, M. Breton, K. Brayman, and A. McCall, "Quantifying temporal glucose variability in diabetes via continuous glucose monitoring: Mathematical methods and clinical application," Diabetes Technol. Ther., vol. 7, pp. 849–862, 2005.

[6] C. Chatfield, *Time-Series Forecasting*, Chapman & Hall/CRC, 2001.

[7] J. H. Zar, *Biostatistical Analysis*, Upper Saddle River, NJ: Prentice-Hall, 1999.

[8] W. L. Clarke, "The original Clarke error grid analysis (EGA)," Diabetes Technol. Ther., vol. 7, pp. 776–779, 2005.