Interindividual Glucose Dynamics in Different Frequency Bands for Online Prediction of Subcutaneous Glucose Concentration in Type 1 Diabetic Subjects

Chunhui Zhao, Youxian Sun, and Luping Zhao
State Key Laboratory of Industrial Control Technology, Dept. of Control Science and Engineering, Zhejiang University, Hangzhou, 310027, China

DOI 10.1002/aic.14176
Published online July 11, 2013 in Wiley Online Library (wileyonlinelibrary.com)

This article investigates the interindividual variability of underlying glucose dynamics and the relative predictive power of exogenous inputs in different frequency bands (FBs) for online subcutaneous glucose prediction for subjects with type 1 diabetes mellitus. Auto-regressive (AR) models and AR models with exogenous inputs (ARX) are developed based on two groups of ambulatory subjects and two groups of in silico subjects using different combinations of FBs. Some important modeling parameters are studied with respect to their influences on glucose prediction. Four issues are of particular interest and discussed based on the illustration results, suggesting that: (i) In different frequency bands, the underlying glucose dynamics act differently across subjects for online glucose prediction; (ii) A global AR model can be developed for one subject in some FB and then used to make online glucose predictions for other subjects in the same FB, revealing little interindividual variability; (iii) The exogenous inputs have different influences in different FBs for prediction of future subcutaneous glucose concentration; (iv) The exogenous inputs may not excite the glucose signals in some FBs, that is, the inclusion of the exogenous inputs may not result in more accurate models in comparison with standard AR model in some FBs. © 2013 American Institute of Chemical Engineers *AIChE J*, 59: 4228–4240, 2013
Keywords: frequency bands, subcutaneous glucose prediction, continuous glucose monitor (CGM), interindividual glucose dynamics, diabetes

Introduction

Type 1 diabetes mellitus (T1DM) is a disease characterized by the inability of the body to regulate blood glucose concentration. T1DM results from autoimmune destruction of the pancreatic β-cells, which produce the hormone insulin. Without appropriate treatment with exogenous insulin, people with T1DM have difficulty maintaining their blood glucose concentration within a normal range (e.g., 70–150 mg/dL). Consequently, they can suffer from large glycemic excursions, including episodes with very low glucose (hypoglycaemia) and very high glucose (hyperglycemia); both situations are detrimental to the quality of life.[1] In general, a careful balance is required among a person's daily activities, diet, and insulin administration in order to bring the blood glucose into a normal range. However, this balancing is not an easy task because large glycemic variations often go undetected, including asymptomatic hypoglycemia.[2]

Continuous glucose monitoring (CGM) devices can measure glucose time-series data online (i.e., every 1–5 min). The real-time data provide timely and important information about the person's current glycemic state, and also reveal its direction and rate of change. CGM devices have opened new opportunities for glycemia management of subjects with T1DM. For example, it has been reported that if the recent glucose history follows previous known patterns, future blood glucose values might be anticipated from past measurements.[3] Many empirical (or "data-driven") modeling techniques have been evaluated for glucose prediction based on both in silico and clinical studies.[3–14] In general, the glucose prediction model has the form of a linear dynamic model where the future glucose concentration is predicted based on current and past glucose signals, sometimes available exogenous input signals, notably insulin delivery and meal carbohydrate (CHO) estimates. Bremer and Gough[3] first suggested that glycemic time-series data had an inherent structure that could be described by a simple linear dynamic model. The linear models that have received the most attention for T1DM applications are autoregressive models (AR) and autoregressive models with exogenous inputs (ARX) based on whether the exogenous inputs are used. For AR modeling, only CGM data are required to develop the model and to predict future glucose concentrations as a linear combination of recent measurements. The Cobelli[7,8] and Reifman research groups[9] have clinically evaluated subject-specific AR models with different model orders in order to improve management of glucose concentrations. Eren-Oruklu and coworkers[10,11] have reported subject-specific recursive AR models where the model parameters were recursively updated to reflect the recent glucose history. It is noted that without special

Correspondence concerning this article should be addressed to C. Zhao at chhzhao@zju.edu.cn.

statement, both AR and ARX models are developed based on the conventional least-squares (LS) algebra.[15] The best fit in the least-squares sense minimizes the sum of squared residuals, a residual being the difference between an observed value and the fitted value provided by a model. A new latent variable (LV)-based empirical modeling algorithm has recently been proposed[16,17] for T1DM applications where AR and ARX models are developed using LV-based multivariate statistical analysis algorithm.[18] By this method, the relationships among multiple variables are examined and modelled, which can reduce a large number of variables to a smaller, more important and informative, number of factors. The results for clinical and in silico applications have demonstrated the effectiveness of the proposed method and its improved prediction accuracy, compared with standard AR and ARX models.

In contrast to the widespread development of personalized glucose prediction models, there is a limited body of work concerning the analysis of intersubject variability and its effects on glucose prediction. In particular, Gani et al.[19] have mentioned and verified the concept of a universal higher-order prediction model for short-term (0–60 min) glucose prediction for T1DM. However, the time-series data are smoothed retrospectively[19,20] using a Tikhonov regularization method.[21] That is, future glucose values are needed to smooth the current glucose signals. Thus, the model is not suitable for online applications where future glucose values are not available, which, thus, was limited regarding its practical application. Zhao et al.[22] reported a global AR model for online glucose prediction based on frequency-band separation where glucose dynamics are separated into two parts, the low-frequency band and high-frequency band, where a threshold value of 60 min is in general chosen for frequency band separation. The low-frequency band is deemed to cover the important and subject-common glucose dynamics. A global prediction model can be developed based on the low-frequency band glucose signals for one subject and then applied to other subjects without model modification for online glucose prediction. Besides glucose prediction, various glucose control algorithms have been developed for diabetes subjects.[23–26] Van Heusden[26] designed control-relevant models for T1DM to achieve the desired control performance rather than minimize a prediction error of future glucose values where a conservative and approximate control relevant model was first designed for all subjects and then it was personalized using *a priori* patient characteristics. Good control performance was reported that hypoglycaemia was completely avoided even after large meal disturbances.

Rahaghi and Gough[27] have recently suggested that the blood glucose dynamics for subjects without diabetes can be divided into four distinct frequency ranges with different periods. Each band is driven by different physiological mechanisms of glucose regulation and intrinsic blood glucose dynamics. It is reported for example that in healthy individuals, the pulsatile insulin secreted by the pancreas is reflected in patterns of blood glucose signals oscillations with periods between 4 and 15 min[28] and the second time scale, approximately 60–120 min, reflects the intrinsic oscillatory behavior. The efficient development of glucose concentration predictive models for type 1 diabetic subjects requires a more detailed analysis of the frequency components of the CGM glucose signals[29] where the relative importance and predictive power of these four frequency bands are considered for short-term (0–60 min) glucose prediction. Subject-dependent sub-band AR models[29] were developed to model and predict the temporal dynamics of CGM signals of nine deidentified type 1 diabetic subjects. These articles provide meaningful insights into the underlying glucose dynamics with respect to glucose prediction from a frequency domain viewpoint. However, they are limited to the subject-dependent analysis of glucose dynamics. Remarkably, the glucose dynamics may vary differently across individuals in different frequency bands (FBs). Moreover, the relative importance and predictive power of exogenous inputs for forecasting of glucose concentrations in different bands are not studied by their work. Thus, the model is not suitable for use in a model-based control strategy such as model predictive control. In each FB, how the glucose dynamics evolve with time and change across individuals, and how they act under the influences of exogenous inputs, as important issues, should be addressed, which is hoped to provide important guideline for glucose modeling and online prediction.

In this article, the glucose dynamics are investigated with respect to the interindividual variability and the relationships with exogenous inputs in different FBs for online glucose prediction. The CGM glucose signals are divided into four different frequency bands as suggested by previous work.[27,29] Different AR and ARX models are developed for each subject in different FBs. They are then applied to the corresponding bands of different subjects for glucose prediction so that the interindividual variability of glucose dynamics in different FBs and the relative predictive power of exogenous inputs in each FB can be analysed. The conventional LS algebra and one popular LV based multivariate statistical analysis method, partial LS-canonical correlation analysis (PLS-CCA) method[18] which is a combination of (PLS)[30,31] and CCA algorithms,[32,33] are employed here for AR/ARX empirical modeling. Based on the above strategy, it answers whether the glucose dynamics in different FBs are common and shared among different subjects. The feasibility of developing global AR-based glucose prediction models for T1DM is also analyzed and explained from the viewpoint of frequency analysis. Moreover, it reveals how the glucose dynamics response to the excitation of exogenous inputs in different FBs. The above assumptions and thoughts are illustratively demonstrated by applying the developed methods to both ambulatory and in silico subjects. It is noted that prediction is the concerned issue in this work. Control-relevant issues regarding how to get effective control models and how to get the right sign of insulin-glucose gain, etc., are beyond the scope of this article.

Methodology

Standard AR and ARX prediction models

In this article, AR modeling techniques based on standard LS algebra[15] and a new latent variable-based statistical analysis[16,17] are used to develop empirical prediction models from glucose time series data, revealing the glucose autocorrelations. ARX models are also developed after adding the exogenous inputs to the AR models, which model the relationships between the exogenous inputs and glucose signals besides the glucose autocorrelations.

The LS-based AR and ARX models are linear dynamic models having been widely considered for prediction and

control calculations in the diabetes control literature. The general form of the LS-based ARX model used in this paper is given by Eq. 1

$$A\left(q^{-1}\right)g_t = B_{\text{ins}}\left(q^{-1}\right)u_{\text{ins},t-k_{\text{ins}}} + B_{\text{meal}}\left(q^{-1}\right)u_{\text{meal},t-k_{\text{meal}}} + \beta + \varepsilon_t \tag{1}$$

where g_t denotes glucose concentration at sampling instant t. $u_{\text{ins},t}$ and $u_{\text{meal},t}$ are the exogenous inputs, bolus insulin and meal carbohydrate content, at time t. In general, insulin dosing is divided into two regimens: basal and bolus insulin. The basal insulin is required for fasting conditions; while bolus insulin is calculated to correct for meals or hyperglycemia condition. In this work, meal and bolus insulin are used as the input while basal insulin which stays invariable is not included in the model for simplicity. β is a constant bias term, and ε_t is a zero mean, random disturbance at time t. Note that this ARX model is somewhat unusual because it is based on physical variables and a bias term, rather than deviation variables. The advantage of this approach is that it eliminates the need to specify an appropriate steady-state reference value for the glucose concentration, information that may be difficult to determine in practice due to the inherent dynamic behavior of blood glucose concentration. In Eq. 1 the input time delays, k_{ins} and k_{meal}, can be different for each subject. The time delays are expressed as integer multiples of the sampling period. In this article, the sampling period is 5 min.

In Eq. 1, $A(q^{-1})$, $B_{\text{ins}}(q^{-1})$, and $B_{\text{meal}}(q^{-1})$ denote polynomials in q^{-1}, where q^{-1} is the backward shift operator, that is, $q^{-1}g_t \equiv g_{t-1}$. For example

$$A\left(q^{-1}\right) = a_0 + a_1 q^{-1} + a_2 q^{-2} + \ldots + a_{n_{\text{A}}} q^{-n_{\text{A}}} \tag{2}$$

where n_{A} *is t*he order of the $A(q^{-1})$ polynomial. It determines the number of previous glucose measurements that are relevant for prediction. When polynomials B_{ins} and B_{meal} are set equal to zero, the ARX model in Eq. 1 reduces to an AR model.

The identification of an AR or ARX model corresponds with specified model orders and can be performed analytically using standard LS regression[15] to estimate the model coefficients (e.g., $\{a_i\}$). However, if the training data are highly correlated, an ill-conditioned or rank deficient problem can arise. To address this problem, a regularization modification to the LS calculations was used by Gani et al.,[19,20] to provide a trade-off between the fit to the training data and the smoothness of future predictions. The net effect of regularization is the introduction of a small bias to the standard LS solution.

LV-based AR and ARX models

An LV-based AR/ARX glucose prediction method that has been recently reported[16,17] is also considered in this article. The resulting models have the same structures as standard ARX/AR models. However, the model parameters are calculated using different principles and algorithms. The LV-based AR/ARX models are developed from two sets of data,[16,17] predictor variable data $\mathbf{X}(N \times J_x)$, and output variable data $\mathbf{y}(N \times 1)$ where N is the number of observations and J_x is the number of predictor variables. The predictor data matrix $\mathbf{X}$ consists of past and current glucose concentrations and the two exogenous inputs. The single output variable is the future glucose concentration. For clarity, the resulting LV-based glucose prediction model is denoted as an LVX model when the exogenous inputs (bolus insulin and meal CHO) are included in the predictor matrix $\mathbf{X}$, and as an LV model when they are not.

For T1DM, the time-series glucose measurements contain the autocorrelation structure relating past and future glucose concentration values. Also, the changes of glucose values are closely related to the exogenous inputs. This provides a good analysis platform for the LV-based modeling method. In this method, a few LVs are first calculated and extracted from the predictor matrix which can capture the important glucose dynamics as well as the glucose-inputs relationships and are linear combinations of the available measurements. In the second step, the quantitative relationship between the LVs and future glucose concentration is determined. A variety of LV-based regression methods[30,31] have been developed with the major difference being how the LVs are calculated. The details of the LV-based methods for this research are summarized in Appendix A. The LV-based glucose prediction model is briefly described in Appendix B.

Frequency-band separation based modeling

Rahaghi and Gough[27] have recently suggested that the glucose dynamics for subjects without diabetes can be divided into four distinct frequency ranges with different periods:

- Band I: 5–15 min, which is a result of pulsatile secretion of insulin and contains very little of the total energy of the system.
- Band II: 60–120 min, which reflects the intrinsic oscillatory behavior.
- Band III: 150–500 min, which contains large fraction of signal energy and is caused by external perturbations such as meals and insulin injections.
- Band IV: $\geq$700 min, which corresponds to modulation of other time scales and a circadian rhythm of the blood glucose baseline.

Clearly, each band is driven by different physiological mechanisms and intrinsic blood glucose dynamics. Lu et al.[29] considered the relative importance and predictive power of the four frequency bands using subject-dependent AR models for short-term (0–60 min) glucose prediction. That is, they tried to find out the important bands which were sufficient enough to represent the raw glucose information for model development. Their results indicated that a combination of glucose data for Band II, and either Band III or IV, sufficiently represents the underlying glucose dynamics necessary for model development. The developed model based on the two combined bands is sufficient for prediction of the full-band glucose signals. These articles provide meaningful insights into the underlying glucose dynamics from a frequency domain viewpoint. However, there is very little literature that considers some important aspects of glucose prediction: the possibility that the glucose dynamics may vary differently from subject to subject (i.e., different intersubject variability) for a given frequency band and the exogenous inputs may impose different influences on glucose dynamics in different sub-bands. Consequently, in contrast to previous T1DM papers on frequency domain analysis,[29] in this article glucose dynamics are investigated from an intersubject perspective. The objective is to (i) analyze the inter-individual variability of glucose dynamics so as to explain the reason of global model across subjects; (ii) study the

influences of exogenous inputs on glucose dynamics in different sub-bands and thus their relative roles in glucose prediction.

For frequency band separation, the glucose data are first filtered in order to divide them into four frequency bands. Four bandpass filters are used, each of which only passes certain period band[29] of the raw CGM time-series signals. For example, a first-order low-pass Butterworth filter is employed after specifying its threshold period 700 min to get the glucose signals in Band IV. The filter has the form

$$\tilde{x}(k)=\beta_1 x(k)+\beta_2 x(k-1)\text{-}\alpha\tilde{x}(k-1) \tag{3}$$

where, x is the glucose measurement,$\tilde{x}$ is the filtered value, and the constant filter parameters are α, β_1, and β_2. Thus filter output $\tilde{x}$ is a linear combination of the previous filtered value, and the previous and current measurements.

For subject i, the data in each frequency band are denoted by $\mathbf{x}_i^{\ell}(K\times 1)(\ell=\text{I}\sim\text{IV})$, where K is the number of observations. These data are generated by passing the original CGM glucose data through different filters, including low-pass filter, band-pass filter, and high-pass filter. As suggested by Lu et al.,[29] Band I is lost in the blood glucose signals of type 1 diabetic subjects because their β cells are destroyed and insulin is not produced for this population. Therefore, the lower-frequency band glucose signals with longer period than 60 min are used as the reference values without special statement. Three different single sub-bands and their different combinations are used in this study.

The empirical prediction models can then be developed based on data for either a single band or combined bands. In order to evaluate the feasibility of developing a global model, different models are considered:

1. Global single-band model (GSB): The global model is identified based on one single band glucose data $\mathbf{X}_i^{\ell}$ for one subject. This model is then used to make glucose predictions for the other subjects.
2. Subject-dependent single-band model (SSB): A separate model is identified and used for each subject in each single band.
3. Global multiband model (GMB): The global multiband model is identified based on glucose data in combined sub-bands for a single subject. This model is then used to make glucose predictions for the combined sub-bands in other subjects.
4. Subject-dependent multiband models (SMB): Individual multiband models are designed for each subject and combined sub-bands. Then glucose predictions are made by adding the individual model predictions together for this subject.
5. Global low-frequency band model (GLB): The global low-frequency band model is identified based on the low-frequency band glucose data for a single subject. This model is then used to make glucose predictions for the low-frequency band glucose of the other subjects.
6. Subject-dependent low/high-frequency band model (SLB/SHB): Individual low/high-frequency band models are designed and used for each subject and low/high-frequency band glucose.

These global models are evaluated for different subjects, where cross-subject and cross-group analyses (CS and CG) are performed, which are distinguished by whether the test subjects are from the same group as those training subjects. When the model developed for training data of one subject in one group is applied to test data of the other subjects in the same group, it is called cross-subject analysis, while when the model is applied to test data of subjects in a different group, it is called cross-group analysis.

In order to characterize the glucose prediction performance, two metrics are used:

1. Root mean-square error (RMSE [mg/dL])[19,20]

$$\text{RMSE}=\sqrt{\frac{1}{N}\sum_{i\in N}\left(y(i)-\hat{y}(i)\right)^2} \tag{4}$$

where $\hat{y}(i)$ is the predicted value, $y(i)$ is the measurement; and N is the number of samples. RMSE is a frequently used measure of the differences between values predicted by a model or an estimator and the values actually observed, which serves to aggregate the individual residuals into a single measure of predictive power.

2. The coefficient of determination (R^2 [%])[35]

$$R^2\hat{\mathbf{y}}=\left(1-\frac{\sum_{i\in N}\left(y(i)-\hat{y}(i)\right)^2}{\sum_{i\in N}\left(y(i)-\bar{\mathbf{y}}\right)^2}\right)\times 100\% \tag{5}$$

where $\bar{\mathbf{y}}$ denotes the mean value of quality measurement. $R^2(\%)$, calculated as the comparison of explained variance (variance of the model's prediction) with the total variance, provides a measure of how well future values are likely to be predicted by the model.

Based on the evaluation of two metrics, comparison is made between the global model and subject-dependent model, single-band model and multiband model, AR model, and ARX model, respectively, with respect to the prediction performance for future glucose concentrations. It checks how the glucose dynamics change across subjects in different FBs, whether a global prediction model is comparable to the subject-dependent models, which band is significantly influenced by the exogenous inputs and what is the relative predictive power of inputs for online glucose prediction.

Illustration Results and Discussion

Clinical subjects

Two groups of deidentified ambulatory clinical data from subjects with T1DM are used in this investigation. Both datasets are retrieved from the Diabetes Research in Children Network (DirecNet) website,[36] which makes continuous glucose data for six different studies. For this article, we used two different studies entitled "The Accuracy of Continuous Glucose Monitors in Children with Type 1 Diabetes" and "A Randomized Clinical Trial to Assess the Effectiveness of the GlucoWatch Biographer in the Management of Type 1 Diabetes in Children." The subjects gave their voluntary and written informed consent to participate. They were provided with different real-time CGM systems, collecting subcutaneous glucose concentrations every 5 min for several continuous days. For the analysis in this article, 12 subjects and 14 subjects are included in two different groups respectively which possess consecutive 3500-min segments (i.e., 700 data points) without data gaps. We use the first 2000 min (400 samples) of the CGM measurements as training data to identify the glucose prediction models in different frequency bands and then the remaining data (testing data) are used for

validation of model performance. Based on the two groups, the empirical models can be developed and applied across different individuals, with different CGM devices and different study objectives. Here, for clinical subjects, the feasibility of global AR model is analyzed to reveal how glucose dynamics change across subjects in different bands. So in the following illustration results, the comparison between global and subject-dependent AR models is focused on.

The AR modeling technique is evaluated as a candidate for developing FB-based models in the following manner. First, different threshold filter values are selected for different filters. Then the original CGM signals are divided into four different frequency ranges. Figure 1 illustrates the data filtering results for a typical type 1 diabetes subject in Group 1 where the glucose profiles in each sub-band and combined sub-bands are comparatively plotted. In general, it is directly observed that the glucose signals in the lower-frequency bands (covering Bands III and IV) seem to be able to capture the overall glucose trends while the high-frequency signals (especially those in Band I which are not shown here) are more oscillatory.

In a preliminary investigation of LS/LV-based AR model development,[17] it was determined that as the model order (i.e., the predictor length (PL) which reveals the number of glucose variables included in AR model) increases, glucose prediction accuracy for subject-specific models did not improve for PL>7. Therefore, a value of PL = 7 is used for model identification. Then for each of the subjects in Groups 1 and 2, training data for a single subject are used to develop different AR models, single-band and multiband models. The model coefficients are then averaged across different subjects in different bands as shown in Figure 2. Also, for comparison, the models are also developed for the glucose signals across full bands (that is, raw glucose signals without any filtering). Clearly, for single sub-band and different combinations of multiple sub-bands, the AR model coefficients show the similar profile and the standard deviation (STD) across subjects is relatively small in comparison with the absolute values of model coefficients. Comparatively, the model coefficients for full-band glucose signals are obviously different and reveal relatively large interindividual STD.

To evaluate how glucose dynamics change across subjects in each sub-band, for each subject, different models developed from the other subjects are applied for glucose prediction. Cross-subject and cross-group analyses are performed where global single-band models in cross-subject analysis and cross-group analysis are denoted as CS-GSB and CG-GSB respectively. For each subject, the prediction results in each sub-band are evaluated using RMSE index by comparing the sub-band glucose measurements and the sub-band glucose predictions. The RMSE index values from different models are then averaged for each subject where the mean value and STD of RMSE index values are calculated to reveal the prediction accuracy and its interindividual

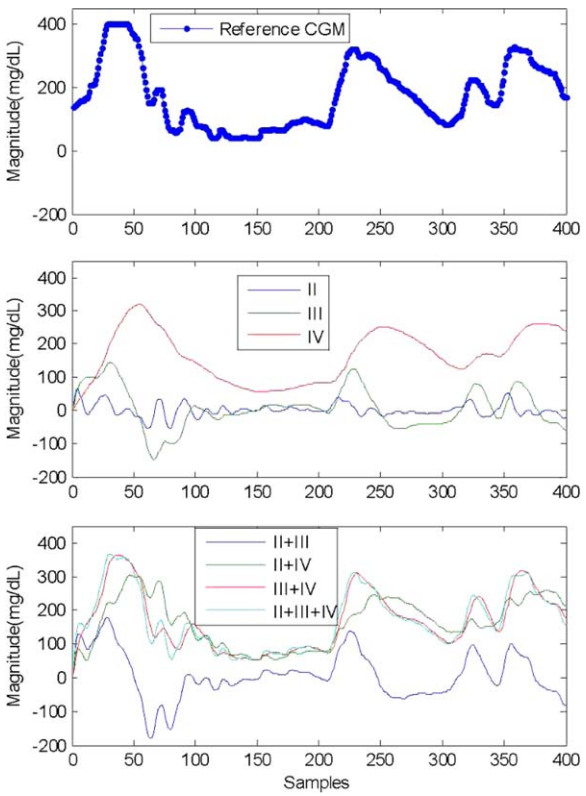

Figure 1. Subcutaneous glucose signals of a typical type 1 diabetic subject in Group 1.

(a) Raw (i.e., reference) glucose signals (b) filtered glucose signals in different single bands and (c) filtered glucose signals in different combinations of multiple bands. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com].

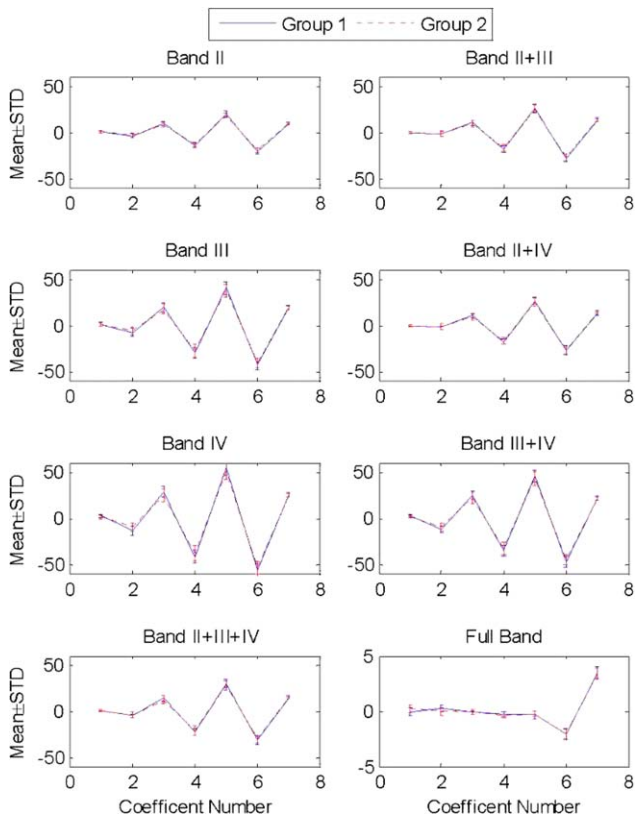

Figure 2. LS-based AR model coefficients (Mean ± STD) developed for 30-min-ahead glucose prediction and clinical subjects of Group 1 and Group 2 in different single bands and combinations of multiple bands.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com].

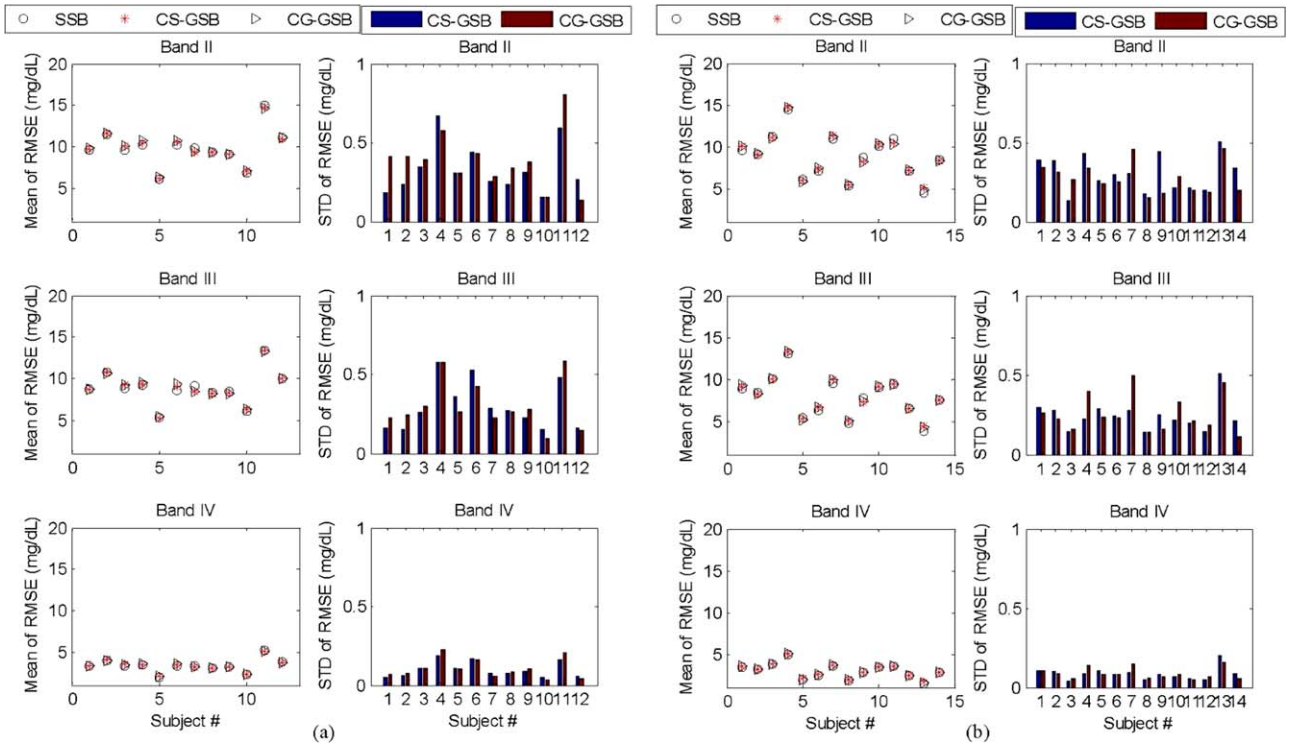

Figure 3. 30-min-ahead glucose prediction performance (RMSE [mg/dL]) using SSB model, cross-subject global single-band model (CS-GSB) and cross-group global single-band model (CG-GSB) for (a) twelve clinical subjects in Group 1 and (b) fourteen clinical subjects in Group 2 in different single bands.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com].

variability in each sub-band. As shown in Figure 3, for both groups, in general, the glucose prediction is more accurate in lower-frequency band (Band IV) than that in higher-frequency bands (Bands II and III) as evaluated by mean of RMSE index. That is, the lower the frequency band, the more accurate the prediction performance. This may result from the fact that lower-frequency signals change slower and more regularly whose time-varying correlations are thus easier to be captured. Also, in each single band, the small differences in mean prediction accuracy for the concerned three different models are not statistically significant based on a paired-t-test ($\alpha = 0.05$)[37], revealing that global model is feasible in each of single bands. For CS-GSB and CG-GSB models, the interindividual variability of glucose predictions as evaluated by STD obviously shows the smallest value in Band IV and the largest value in Band II, revealing smaller interindividual variability in lower-frequency band. Also the combination of Bands II through IV is used to develop global models and compared with subject-dependent models, revealing that the gaps between different sub-bands are not necessary for global model development. The results are not shown here for simplicity.

As shown in Figure 4, the influences of PL are studied for different single-band models regarding the 30-min-ahead glucose prediction performance. The results demonstrate that for all models, the prediction performance is not very sensitive to the choice of the PL value. Also, it illustrates that a value of $PL = 7$ is the best for AR modeling. Figure 5 presents an analogous evaluation for the second important design parameter, the prediction horizon (PH), which indicates the number of steps of ahead prediction. The choice of PH involves a trade-off. It should be large enough to ensure adequate time for a necessary intervention or corrective action, in order to avoid abnormal glycemia. On the other hand, a larger PH value may result in less accurate glucose predictions. As PH increases, the prediction accuracy of all types of models decreases, as expected.

In silico subjects

The simulated subject data are generated for a 5-min sampling period using the FDA-accepted University of Virginia/University of Padova (UVa/Padova) metabolic simulator.[38] They are used to evaluate the relative importance of exogenous inputs for glucose prediction in each sub-band. That is, how the sub-band glucose dynamics are excited by exogenous inputs. Two different groups of subjects are used, including 10 adults and 10 children. The simulations include a three-meal scenario for breakfast, lunch, and dinner taken at about 7 AM, 12 PM, and 6 PM with 40 g, 85 g, and 60 g of CHO, respectively. An optimal bolus insulin was given immediately based on the ideal insulin-to-carbohydrate ratio (I:CR). This situation is used as the nominal case for model identification. Then the meal timing and CHO meal content are varied to represent variations in daily life where a one hour shift (forward or backward) in meal timing and $\pm 75\%$ variation in CHO amount are implemented. Five-day data are simulated for each subject and used for model testing.

The training data for model identification consisted of one day of simulated data (midnight-to-midnight) for the nominal case. Considering that LV modeling method is more powerful to analyze the underlying correlations of multiple variables and can avoid the ill-condition problem in LS-based ARX modeling, it is used here to evaluate the relative importance of exogenous inputs for glucose prediction. For

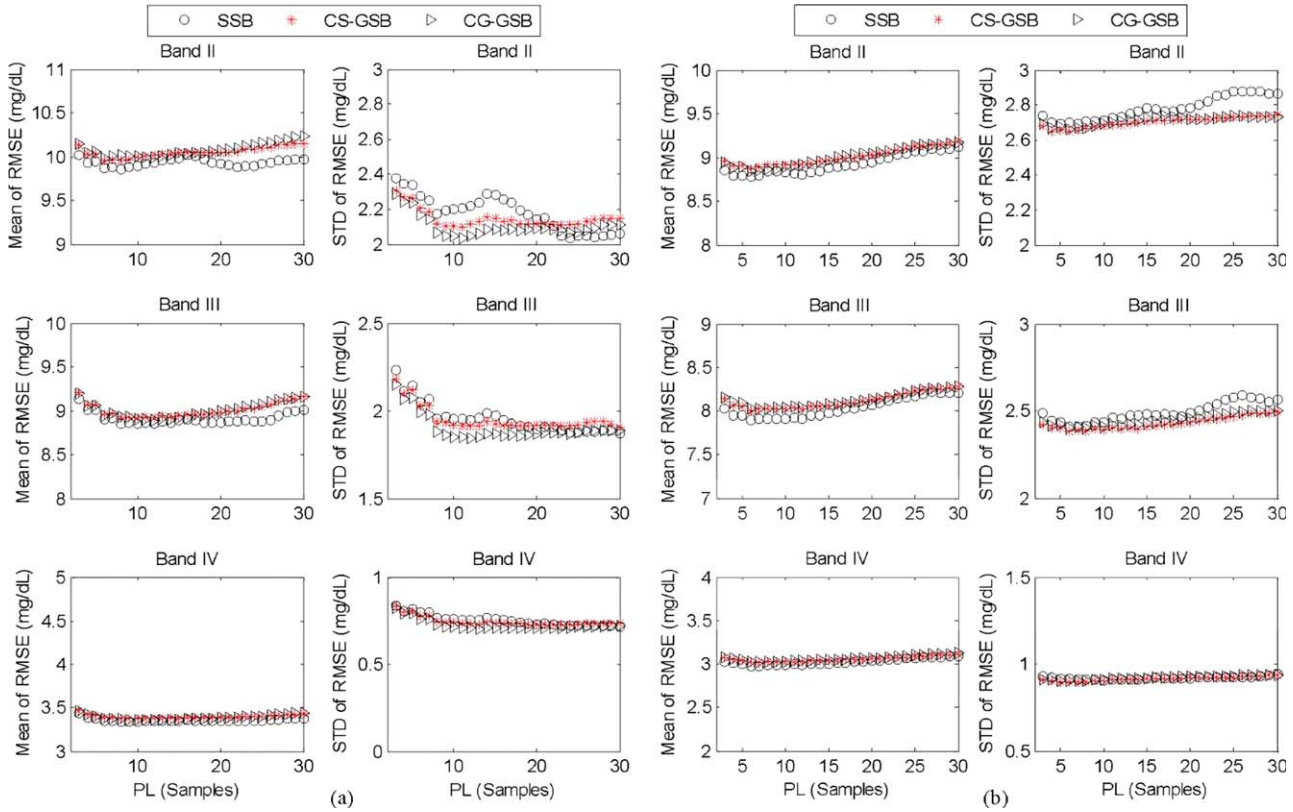

Figure 4. Effect of PL on glucose prediction performance across subjects evaluated by mean and STD of RMSE (mg/dL) for (a) 12 clinical subjects in Group 1 and (b) 14 clinical subjects in Group 2 and different models in different single bands.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com].

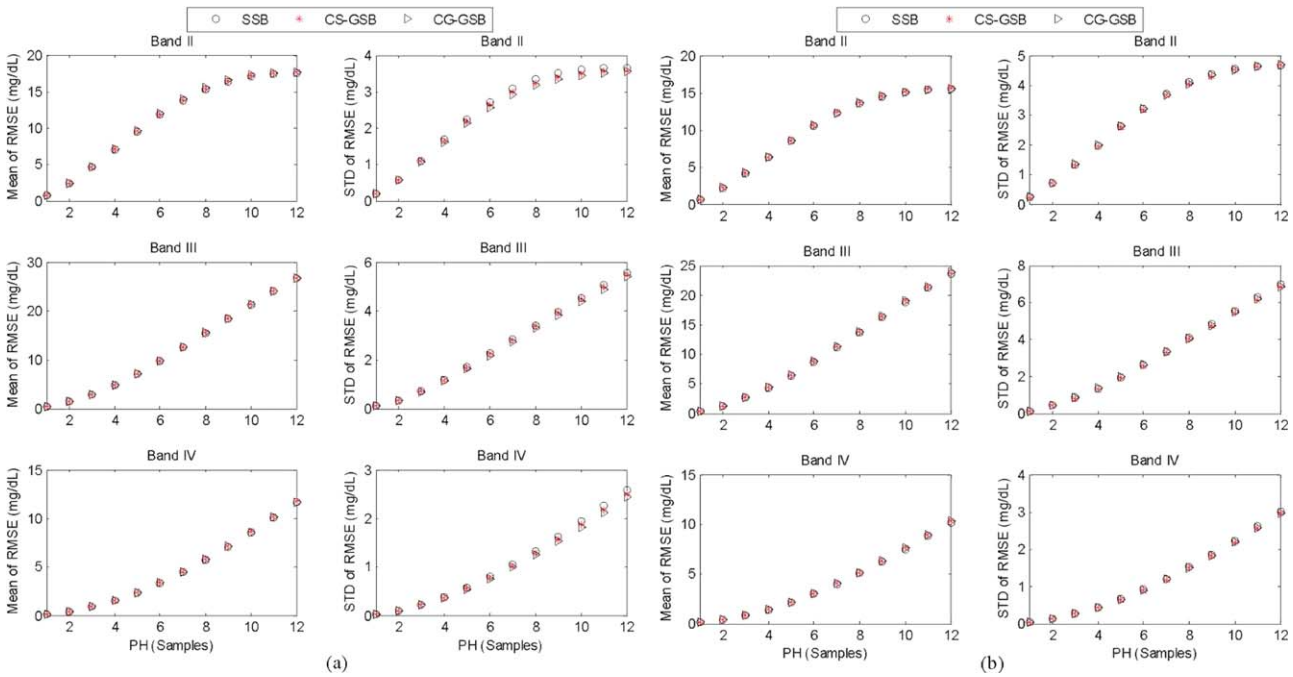

Figure 5. Effect of PH on glucose prediction performance across subjects evaluated by mean and STD of RMSE (mg/dL) for (a) 12 clinical subjects in Group 1 and (b) 14 clinical subjects in Group 2 and different models in different sub-bands.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com].

all possible predictors, two exogenous inputs and glucose signals, seven historical samples up to the current time for each predictor are used for prediction model development. In this way, the inputs with proper time delays from the current time are also included in the predictor matrix. By PLS-CCA algorithm, those predictors which are not closely related with future glucose concentration can be suppressed for the extraction of LVs and only the irrelevant predictor information is extracted by a few LVs. Therefore, LV-based ARX (termed LVX) modeling method can avoid the ill-condition problem which is in comparison with LS based ARX modeling. It is easy to understand that since different subjects have different response speeds and magnitudes to the exogenous inputs, global LVX model is not feasible. Here, for in silico subjects, only subject-dependent LVX model is used to analyze the influences of exogenous inputs. So in the following illustration results, the comparison between LV and LVX is focused on.

For different predictors, different SSB LVX models are developed. SSB-Glucose indicates the prediction model developed using only historical glucose signals. SSB-Insulin indicates the prediction model identified using only historical insulin input signals. SSB-meal indicates the prediction model identified using only historical meal input signals. SSB-G+I+M indicates the prediction model identified using the combined information of historical glucose, insulin, and meal input signals. Clearly, as shown in Figures 6a and b for subjects in both groups, without exogenous inputs, the 30-min-ahead glucose prediction accuracy using SSB-Glucose model is good enough in Band IV only using glucose information, approximately 100% as evaluated by R^2%. When exogenous inputs are included, the glucose prediction performance, in general, is not improved for all subjects by comparing SSB-Glucose model and SSB-G+I+M model in Band IV. Moreover, based on the results obtained from SSB-I and SSB-M models in Band IV, the glucose predictions based on insulin and meal information, which have been beyond the normal variation range of R^2% (0–100), are much worse than those based on only glucose information, revealing that the exogenous inputs are unable to excite the glucose information in Band IV. Comparatively, in Band II, it is clear that exogenous inputs have the most significant influences on future glucose prediction as shown by predictions of SSB-Insulin and SSB-Meal models. Also, the inclusion of insulin and meal in the prediction modeling can greatly improve the prediction performance by comparing SSB-G+I+M and SSB-Glucose models. In Band III, the predictive power of exogenous inputs is not as significant as that in Band IV. Based on the results, the relative importance of exogenous inputs for glucose prediction can not be neglected in Band II and Band III while they are meaningless and can not excite the glucose dynamics in Band IV.

Figure 7 shows the prediction accuracy for different prediction methods and PH values up to 12 samples (60 min) in Band II and Band III, respectively. The RMSE values are averaged for 10 subjects in each group and five days of testing data. As expected, in general, the prediction accuracy decreases as PH increases for SSB-Glucose and SSB-G+I+M models in both bands. However, for SSB-Insulin and SSB-Meal models in Band II, the prediction accuracy is first improved and then decreased while for the two models in Band III, the prediction accuracy is improved as PH increases. The prediction accuracy of SSB-G+I+M model, resulting from the inclusion of the exogenous inputs, is always better than that of SSB-Glucose model especially for larger PH values when exogenous inputs begin to be more influential. In general, the exogenous inputs can not influence the future glucose dynamics until 20–30 min after the current time. Moreover, it is noted that SSB-Insulin and SSB-Meal models in Band II begins to show better

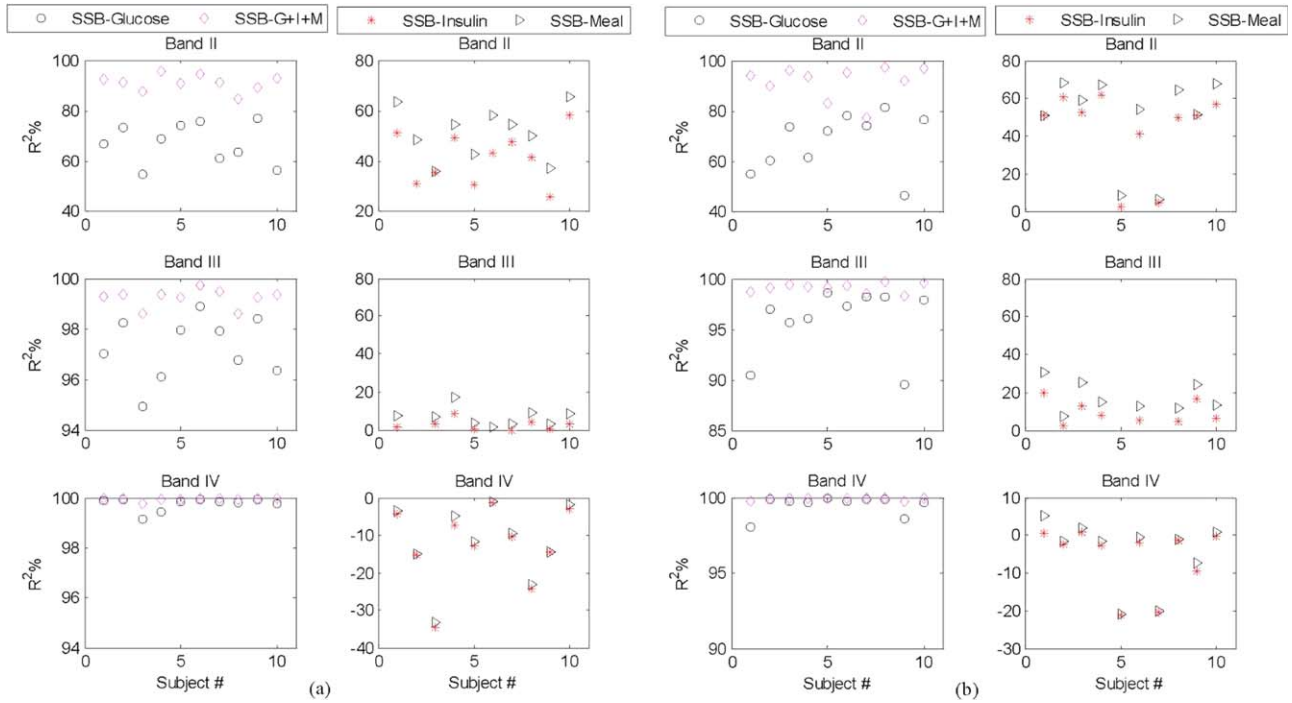

Figure 6. 30-min ahead glucose prediction performance (R^2%) using different SSB models for (a) 10 in silico adults in Group 1 and (b) 10 in silico children in Group 2 in different sub-bands.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com].

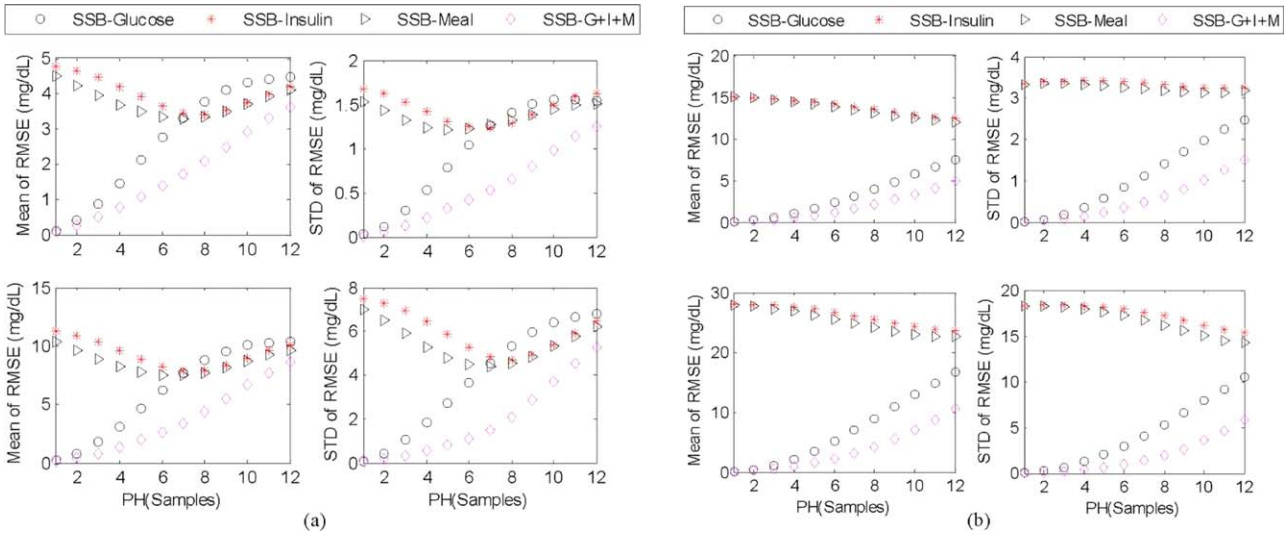

Figure 7. Effect of PH on glucose prediction performance for 10 in silico adults in Group 1 (top subplots) and 10 in silico children in Group 2 (bottom subplots) and different models in (a) Band II and (b) Band III.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com].

prediction accuracy than that of SSB-Glucose model when PH increases up to 7 samples (35 min), revealing that the glucose-inputs relationship is more influential than glucose autocorrelations from now on. Comparatively, in Band III, the predictive power of SSB-Glucose model is always better than that of SSB-Insulin and SSB-Meal models. Based on the results in Figure 7, the predictive power of input information are more significant in Band II, less important in Band III and almost inessential in Band IV.

To further check the influences of exogenous input for glucose prediction, three different subject-dependent modeling methods are compared, including the combination of LVX and LV models, single LVX model, and single LV model. Three different threshold values, 120 min, 500 min, and 700 min, directly related with Band II, Band III, and Band IV, respectively, are used for frequency band separation. Then, LV model is developed based on lower-frequency band glucose, and LVX model is developed based on higher-frequency band glucose and two exogenous inputs, termed L-LV+H-LVX model here. The combined prediction is then compared with that obtained by using single LVX model and single LV model, respectively, which are both developed based on lower-frequency band glucose with 60 min as the threshold value. In Table 1, the prediction accuracy of the three prediction models is compared for PHs of 30 and 60 min based on RMSE evaluation index. Clearly, the LVX model is more accurate than LV model for both Groups 1 and 2. Furthermore, when 500 min or 700 min is used for FB separation, the predictive importance of exogenous inputs is well modeled in the higher-frequency band so that L-LV+H-LVX model is observed to have the same degree of prediction accuracy as that of single LVX model. When 120 min is used for FB separation where only Band II is used for LVX modeling in combination with LV model in lower-frequency band, the prediction accuracy is worse than that of single LVX model. Representative comparisons of the measured and predicted glucose profiles for the three different models are shown in Figure 8 where 500 min is used for frequency band separation. The 60-min-ahead glucose prediction results are for the first subjects in Group 1 and Group 2, respectively and a representative day of test data. In general, the evolving glucose trends are captured by both L-LV+H-LVX and single LVX models and the difference in prediction accuracy is not statistically significant based on a paired t-test ($\alpha = 0.05$). Also, they both show better prediction performance than that of single LV model.

Conclusion

In this article, a combination of frequency-band separation and empirical models has been developed for investigation of interindividual glucose dynamics and the predictive power of exogenous inputs regarding the prediction of subcutaneous glucose concentrations in T1DM. The proposed assumptions and augments are illustrated in two groups of ambulatory subjects and two groups of in silico subjects. The results clearly explain the interindividual variability of glucose

Table 1. RMSE Results (mg/dL) (Mean ± MAD) for Different LVX/LV Glucose Prediction Models and Two Groups of In Silico Subjects

	PH = 6 (30 min)		PH = 12 (60 min)	
Methods	Group 1	Group 2	Group 1	Group 2
LVX	2.4 ± 0.8	4.5 ± 1.9	9.0 ± 3.0	20.0 ± 11.0
LV	5.1 ± 1.8	11.6 ± 6.6	13.1 ± 4.2	29.4 ± 18.6
L-LV+H-LVX (700 min)[a]	2.7 ± 0.8	5.0 ± 2.2	9.5 ± 3.0	21.0 ± 12.2
L-LV+H-LVX (500 min)[a]	2.7 ± 0.8	5.1 ± 2.3	9.6 ± 3.1	21.0 ± 12.2
L-LV+H-LVX (120 min)[a]	3.8 ± 1.1	7.5 ± 3.7	12.7 ± 3.8	27.2 ± 15.7

[a]The values in bracket are used for frequency band separation; three options are used in this table, 700 min, 500 min, and 120 min.

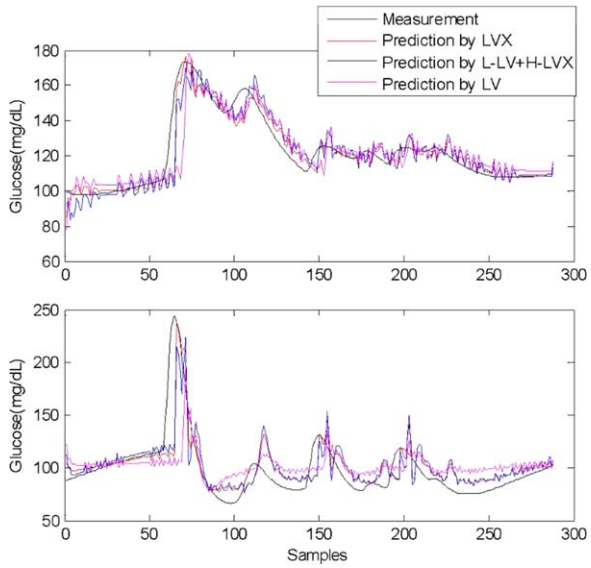

Figure 8. Comparison of representative measured and 60-min-ahead predicted glucose profiles for *in silico* Subject #1 in Group 1 (top) and in silico Subject #1 in Group 2 (bottom) based on three different models.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com].

dynamics in different FBs and the relative importance of exogenous inputs in each FB to excite the glucose signals. In general, for short-term (PH < 60 min) glucose prediction, the cross-subject and cross-study analyses demonstrate that global model developed in each of Bands II, III, and IV exhibit the same degree of prediction accuracy as that of models developed for individual subjects. Thus, these illustration results also help to explain the fact that a global AR model can be developed from one subject and then apply the models to other subjects based on low-frequency glucose information (60 min as the threshold value for frequency band separation) as shown in our previous work.[22] Furthermore, the relative importance of exogenous inputs for glucose prediction in each FB is investigated, which explains that they are influential in Band II, less important in Band III, and almost inessential in Band IV. These promising analyses results should encourage extensions of this research methodology. For example, a critical problem concerns the generation of hypoglycemic event alerts and glucose controller based on online prediction of future glucose.

Acknowledgment

This work is supported by Program for New Century Excellent Talents in University (NCET-12-0492), the National Natural Science Foundation of China (No. 61273166) and Specialized Research Fund for the Doctoral Program of Higher Education of China (20120101120182).

Literature Cited

1. Rubin RR, Peyrot M. Quality of life and diabetes. *Diabetes Metab Res Rev*. 1999;15:205–218.
2. Rabasa-Lhoret R, Garon J, Langelier H. Effects of meal carbohydrate content on insulin requirements in type 1 diabetic patients treated intensively with the basal-bolus (ultralente-regular) insulin regimen. *Diabetes Care* 1999;22:667–673.
3. Bremer T, Gough DA. Is blood glucose predictable from previous values? A solicitation for data. *Diabetes*. 1999;48:445–451.
4. Finan DA, Palerm CC, Doyle III FJ. Effect of input excitation on the quality of empirical dynamic models for type 1 diabetes. *AIChE J*. 2009;55:1135–1146.
5. Trajanoski Z, Regittnig W, Wach P. Simulation studies on neural predictive control of glucose using the subcutaneous route. *Comput Meth Prog Bio*. 1998;56:133–139.
6. Dua P, Doyle FJ III, Pistikopoulos EN. Model-based blood glucose control for type 1 diabetes via parametric programming. *IEEE Trans Biomed Eng*. 2006;53:1478–1491.
7. Sparacino G, Zanderigo F, Corazza S. Glucose concentration can be predicted ahead in time from continuous glucose monitoring sensor time-series. *IEEE Trans Biomed Eng*. 2007;54:931–937.
8. Zanderigo F, Sparacino G, Kovatchev B. Glucose prediction algorithms from continuous monitoring data: assessment of accuracy via continuous glucose error-grid analysis. *J Diabetes Sci Technol*. 2007;1:645–651.
9. Reifman J, Rajaraman S, Gribok A. Predictive monitoring for improved management of glucose levels. *J Diabetes Sci Technol*. 2007;1:478–486.
10. Eren-Oruklu M, Cinar A, Quinn L. Estimation of future glucose concentrations with subject-specific recursive linear models. *Diabetes Technol Ther*. 2009;9:438–450.
11. Eren-Oruklu M, Cinar A, Quinn L. Hypoglycemia prediction with subject-specific recursive time-series models. *J Diabetes Sci Technol*. 2010;4:25–33.
12. Finan DA, Zisser HC, Jovanovič L. Practical issues in the identification of empirical models from simulated type 1 diabetes data. *Diabetes Technol Ther*. 2007;9:438–450.
13. Finan DA, Palerm CC, Doyle III FJ. Identification of empirical dynamic models from type 1 diabetes subject data. Proceedings of American Control Conference (ACC): Seattle, Washington, USA, June 11-13, 2008, 2099–2104.
14. Palerm CC, Willis JP, Desemone J. Hypoglycemia prediction and detection using optimal estimation. *Diabetes Technol Ther*. 2005;7:3–14.
15. Ljung L. System Identification: Theory for the User. Upper Saddle River, NJ: Prentice-Hall, 1999.
16. Zhao CH, Dassau E, Harvey RA. Predictive glucose monitoring for type 1 diabetes using latent variable-based multivariate statistical analysis. *Proc IFAC*. 2011;18:7012–7017.
17. Zhao CH, Dassau E, Jovanovič L. Predicting subcutaneous glucose concentration using latent variable (LV)-based statistical analysis method for Type 1 diabetes mellitus. *J Diabetes Sci Technol*. 2012;6:617–633.
18. Yu H, MacGregor JF. Post processing methods (PLS-CCA): simple alternatives to preprocessing methods (OSC-PLS). *Chemometr Intell Lab Syst*. 2004;73:199–205.
19. Gani A, Gribok AV, Lu YH. Universal glucose models for predicting subcutaneous glucose concentration in Humans. *IEEE Tran Inf Technol Biomed*. 2010;14:157–165.
20. Gani A, Gribok AV, Rajaraman S. Predicting subcutaneous glucose concentration in humans: data-driven glucose modeling. *IEEE Trans Biomed Eng*. 2009;56:246–254.
21. Tikhonov AN, Arsenin VY. Solutions of Ill-Posed Problems. Washington, DC: Winston, 1977.
22. Zhao CH, Dassau E, Zisser H. Prediction of Short-Term Glucose Trends for Type 1 Diabetes Using Empirical Models and Frequency-Band Separation. 71st Scientific Sessions, American Diabetes Association, San Diego, CA, 2011.
23. Parker RS, Doyle FJ III, Ward JH. Robust H∞ glucose control in diabetes using a physiological model. *AIChE J*. 2000;46: 2537–2549.
24. Sánchez-Chávez IY, Martínez-Chapa SO, Peppas NA. Computer evaluation of hydrogel-based systems for diabetes closed loop treatment. *AIChE J*. 2008;54:1901–1911.
25. Dua P, Doyle FJ III, Pistikopoulos EN. Multi-objective blood glucose control for type 1 diabetes. *Med Biol Eng Comput* 2009;47: 343–352.
26. van Heusden K, Dassau E, Zisser H. Control-relevant models for glucose control using a priori patient characteristics. *IEEE Trans Biomed Eng*. 2012;59:1839–1849.
27. Rahaghi FN, Gough DA. Blood glucose dynamics, *Diabetes Technol Ther*. 2008;10:81–94.
28. Porksen N. The in vivo regulation of pulsatile insulin secretion. *Diabetologia*. 2002;45:3–20.
29. Lu YH, Gribok AV, Ward WK. The importance of different frequency bands in predicting subcutaneous glucose concentration in type 1 diabetic patients. *IEEE Trans Biomed Eng*. 2010;57:1839–1846.

30. Burnham AJ, Viveros R, MacGregor JF. Frameworks for latent variable multivariate regression. *J Chemometr*. 1996;10:31–45.
31. Lindgren F, Geladi P, Wold S. The kernel algorithm for PLS. *J Chemometr*. 1993;7:45–59.
32. Dayal B, Macgregor J. Improved PLS algorithms. *J Chemometr*. 1997;11:73–85.
33. Anderson TW. Canonical correlation analysis and reduced rank regression in autoregressive models. *Ann Stat*. 2002;30:1134–1154.
34. Canonical Correlation, a Tutorial. Magnus Borga. Available at: http://www.imt.liu.se/~magnus/cca/tutorial/tutorial.pdf, 2001.
35. Kleinbaum DG, Kupper LL, Muller KE. Applied regression analysis and other multivariable methods, 3rd ed. California: Wadsworth Publishing Co Inc, 2003.
36. Diabetes Research in Children Network (DirecNet). Available at: http://direcnet.jaeb.org/ViewPage.aspx?PageName=PreviousStudies.
37. Montgomery DC, Runger GC. Applied statistics and probability for engineers, 4th ed. New York: Wiley, 2006.
38. Kovatchev BP, Breton M, Dalla Man C. In silico preclinical trials: a proof of concept in closed-loop control of type 1 diabetes. *J Diabetes Sci Technol*. 2009;3:44–55.

Appendix A: LV-based Statistical Analysis

The LV-based models for this article are empirical, linear dynamic models that predict an output variable, future glucose concentration, from past glucose measurements, bolus insulin, and meal CHO estimates (the predictor variables). Because these predictor data tend to be highly correlated, multivariable statistical methods are a natural choice since they have the ability to analyze large amounts of highly correlated data. The underlying assumption is that the predictor data can be described by a small number of orthogonal LVs that can be directly linked to the output variable via regression analysis.

A variety of LV-based regression methods[30–34] have been developed with the chief differences being how the LVs are calculated. A general comparison of LV-based methods has been reported by Burnham et al.[31] In this research, the LV-based modeling method is a PLS-CCA approach,[18] where CCA[33,34] provides post processing of PLS[30–32] modeling results. It involves two modeling algorithms, PLS and CCA. The LV-based methods used in this paper are briefly described below.

PLS[30–32] is a common LV-based regression method. The LVs are linear combinations of the predictor variables $\mathbf{X}(N\times J_x)$ that result in maximal covariance with the output variable $\mathbf{y}(\mathrm{N}\times 1)$. Thus, the first LV (or score) $\mathbf{t}_1$ can be expressed as

$$\mathbf{t}_1=\mathbf{X}\mathbf{w}_1 \tag{A1}$$

where $\mathbf{X}(N\times J_x)$ is the predictor data matrix and the N is the number of samples and J_x is the number of predictor variables. The first weight vector $\mathbf{w}_1$ is a value of $\mathbf{w}$ that maximizes the objective function

$$\begin{aligned}&\arg\max_{\mathrm{w}}\left(\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y}\mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbf{w}\right)\\&\text{subject to } \mathbf{w}^{\mathrm{T}}\mathbf{w}=1\end{aligned} \tag{A2}$$

where vector $\mathbf{y}(N\times 1)$ denotes the output variable data. Thus $\mathbf{w}_1$ is the eigenvector that corresponds to the largest eigenvalue of matrix $\mathbf{X}^{\mathrm{T}}\,\mathbf{y}\mathbf{y}^{\mathrm{T}}\,\mathbf{X}$.

The second weight $\mathbf{w}_2$ is calculated in a similar manner after $\mathbf{X}$ and $\mathbf{y}$ have been deflated by $\mathbf{t}_1$

$$\mathbf{p}_1^{\mathrm{T}}=\left(\mathbf{t}_1^{\mathrm{T}}\mathbf{t}_1\right)^{-1}\mathbf{t}_1^{\mathrm{T}}\mathbf{x} \tag{A3}$$

$$\mathbf{E}_1=\mathbf{X}-\mathbf{t}_1\mathbf{p}_1^{\mathrm{T}} \tag{A4}$$

$$q_1=\left(\mathbf{t}_1^{\mathrm{T}}\mathrm{t}_1\right)^{-1}\mathbf{t}_1^{\mathrm{T}}\mathbf{y} \tag{A5}$$

$$\mathbf{f}_1=\mathbf{y}-\mathbf{t}_1q_1 \tag{A6}$$

where $\mathbf{p}_1$ and q_1 are the PLS loadings for the predictor variables and the output variable, respectively. In order to calculate $\mathbf{w}_2$, the calculation in (A2) is repeated with $\mathbf{E}_1$ and $\mathbf{f}_1$ replacing $\mathbf{X}$ and $\mathbf{y}$, respectively. The remaining weight vectors are also calculated using this iterative procedure. The number of LVs in the PLS model, R_{LV}, is an important design parameter that can be as large as the rank of $\mathbf{X}$. In this article, an appropriate value of R_{LV} was determined by cross validation.

Next, the weight vectors $\mathbf{w}_1$, $\mathbf{w}_2$,… are collected in a weight matrix $\mathbf{W}$ while the loadings for the predictor variables and the output variable are collected in matrix $\mathbf{P}$ and vector $\mathbf{q}$, respectively. Finally, the score vectors $\mathbf{t}_1$, $\mathbf{t}_2$,… are arranged as the columns of matrix $\mathbf{T}$. Then the output variable prediction is given by

$$\hat{\mathbf{y}}=\mathbf{X}\mathbf{R}\mathbf{q}=\mathbf{T}\mathbf{q} \tag{A7}$$

where,

$$\mathbf{R}=\mathbf{W}\left(\mathbf{P}^{\mathrm{T}}\mathbf{W}\right)^{-1} \tag{A8}$$

A classical PLS calculation procedure is described by Lindgren et al.[31]

One problem associated with the PLS method requires special attention. The PLS objective is to model the variations in $\mathbf{X}$ and maximize their covariance with $\mathbf{y}$. But large covariance does not necessarily mean strong correlation. When the predictor matrix $\mathbf{X}$ contains a considerable amount of process variations that are uncorrelated with $\mathbf{y}$, it is possible that the PLS LVs may capture the major systematic variations in the predictors $\mathbf{X}$ but only have relatively weak correlation with $\mathbf{y}$. This situation leads to a complex model structure and an over-fitting problem.

Unlike PLS, canonical correlation analysis, CCA,[33,34] inherently ignores the variations in $\mathbf{X}$ that are uncorrelated with $\mathbf{y}$ and directly maximizes the variations that are correlated with $\mathbf{y}$. The CCA objective function is to determine the weight vectors $\mathbf{w}$ so that

$$\begin{aligned}&\arg\max_{\mathrm{w}}\left(\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y}\right)\\&\text{subject to } \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w}=1\end{aligned} \tag{A9}$$

The first weight vector $\mathbf{w}_1$ is the eigenvector corresponding to the largest eigenvalue of matrix $\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}(\mathbf{y}^{\mathrm{T}}\mathbf{y})^{-1}\mathbf{y}^{\mathrm{T}}\mathbf{X}$ The maximum number of CCA LVs can be up to $L_{\mathrm{cca}}=\min(J_x, J_y)$, where J_x and J_y are the numbers of predictor and output variables, respectively. In this paper, a single output variable is considered and thus $L_{\mathrm{cca}}=1$. A comparison of Eqs. A2 and A9 indicates that the CCA objective is to maximize correlation while the PLS objective is to maximize covariance.

For CCA the single LV and loading for the output variable are calculated as

$$\mathbf{t}=\mathbf{X}\mathbf{w} \tag{A10}$$

$$q=\left(\mathbf{t}^{\mathrm{T}}\mathbf{t}\right)^{-1}\mathbf{t}^{\mathrm{T}}\mathbf{y}=\mathbf{t}^{\mathrm{T}}\mathbf{y} \tag{A11}$$

Finally, the output prediction is given by

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}q = \mathbf{t}q \tag{A12}$$

Unfortunately, directly applying CCA to the $\{\mathbf{X}, \mathbf{y}\}$ data can lead to ill-conditioned problems resulting from the $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ term in the calculations. To avoid this problem, Yu and MacGregor[18] have suggested a two-step LV-based modeling algorithm (PLS-CCA), where CCA is used as a post-processing technique to further improve the PLS LVs. In this way, a parsimonious regression model with the same prediction ability as the standard PLS model can be obtained. Based on these considerations, their PLS-CCA algorithm is employed in this paper to develop the empirical model for glucose concentration prediction. Compared with the conventional AR and ARX modeling methods, the y-related variability in the predictor data is modeled by only a few LVs, which are calculated in order based on their relationships with future glucose values.

Appendix B: LV/LVX Prediction Model

In order to apply the LV-based modeling technique, the predictor and output datasets must be organized in an appropriate manner. For both the simulation and clinical studies, the predictor data were available for multiple days using a 5-min sampling period. The CGM glucose data can be represented as a data vector $\mathbf{g}(K\times 1)$ where K is the number of samples. The recorded bolus insulin and estimated meal CHOs are denoted by $\mathbf{u}_{\mathrm{I}}(K\times 1)$ and $\mathbf{u}_{\mathrm{M}}(K\times 1)$, respectively. It is desired to predict the glucose concentration PH time steps ahead where PH is the PH. These predictions are based on recent values of the predictor variables. A key question is how many past values of glucose and the two exogenous inputs should be included in the model? Let L_{G}, L_{I}, and L_{M} denote the numbers of past samples for glucose, insulin and meal CHO, respectively, that are used to make the predictions. These parameters are referred to as the PLs which indicate how much historical information on which the future output depends.

For the model development, the training data is organized as follows. The predictor matrix is defined as

$$\mathbf{X}(N\times J_x) = [\mathbf{G}(N\times L_{\mathrm{G}}), \mathbf{U}_{\mathrm{I}}(N\times L_{\mathrm{I}}), \mathbf{U}_{\mathrm{M}}(N\times L_{\mathrm{M}})] \tag{B1}$$

Where N is the number of glucose measurements to be predicted, $\mathrm{N}=\mathrm{K}-\mathrm{L}-\mathrm{PH}+1$, $L=\max(L_{\mathrm{G}}, L_{\mathrm{I}}+D_{\mathrm{I}}-1, L_{\mathrm{M}}+D_{\mathrm{M}}-1)$, and D_{I} and D_{M} are the input time delays for the bolus insulin and the meal CHO, respectively. Note that $N<K$ due to the initialization period required to acquire the past data for the first glucose prediction. Model parameter $J_x = L_{\mathrm{G}}+L_{\mathrm{I}}+L_{\mathrm{M}}$ is the number of variables in the arranged data matrix, that is, the model order.

The bolus insulin predictor data are arranged as

$$\mathbf{U}_{\mathrm{I}}(N\times L_{\mathrm{I}}) = \begin{bmatrix} \mathbf{u}_{\mathrm{I},1}{}^{\mathrm{T}}(1\times L_{\mathrm{I}}) \\ \mathbf{u}_{\mathrm{I},2}{}^{\mathrm{T}}(1\times L_{\mathrm{I}}) \\ \vdots \\ \mathbf{u}_{\mathrm{I,K\text{-}L\text{-}PH+1}}{}^{\mathrm{T}}(1\times L_{\mathrm{I}}) \end{bmatrix} \tag{B2}$$

Each row vector $\mathbf{u}_{I,i}{}^{\mathrm{T}}(1\times L_I)$ $(i=1,2,\ldots,N)$ contains the bolus insulin information from time $i+L-L_{\mathrm{I}}$ to $i+L-1$. Similarly, the analogous predictor matrices for the other two predictor variables are denoted by $\mathrm{U}_{\mathrm{M}}(N\times L_{\mathrm{M}})$ and $\mathbf{G}(N\times L_{\mathrm{G}})$, respectively.

The model output data are arranged as

$$\mathbf{y}(N\times 1) = \begin{bmatrix} g_{\mathrm{L+PH}} \\ g_{\mathrm{L+PH+1}} \\ \vdots \\ g_{\mathrm{K}} \end{bmatrix} \tag{B3}$$

where $g_{\mathrm{L+PH}}$ is the glucose measurement at time L+PH.

Based on the above data arrangement, the regression dataset is available, $\{\mathbf{X}, \mathbf{y}\}$. The training data $\{\mathbf{X}, \mathbf{y}\}$ are normalized to have zero mean and unit variance, respectively, which reduces the data nonlinearity to some extent. After applying the PLS-CCA approach[18] to the normalized data,

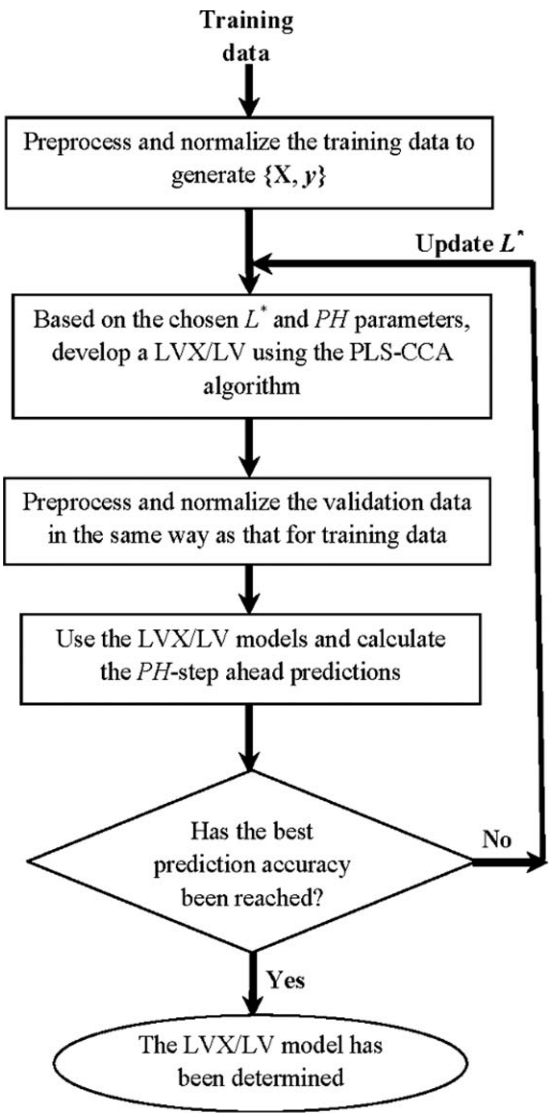

Figure B1. A schematic representation of the LVX/LV modeling method (L^*: For LVX modeling, it includes the values of L_{G}, L_1 and L_{M}; as well as the input time delay D_1 and D_{M}; while for LV modeling, it only includes the value of L_{G}).

the final LV-based regression model can be readily calculated

$$t_c = \mathbf{X}\mathbf{w}_c \tag{B4}$$

$$q_c = \left(t_c{}^T t_c\right)^{-1} t_c{}^T y \tag{B5}$$

$$\hat{\mathbf{y}}_c = \mathbf{t}_c q_c \tag{B6}$$

where $\mathbf{w}_c(J_x \times 1)$ is the single PLS-CCA weight vector. LV $\mathbf{t}_c$ is a linear combination of the predictor variables and weight vector $\mathbf{w}_c(J_x \times 1)$, indicating the systematic variations in predictor variables that are closely related to the output variable. The vector of prediction errors $\mathbf{f}$ is defined

$$\mathbf{f} = \mathbf{y} - \hat{\mathbf{y}}_c \tag{B7}$$

The two-step modeling method can be summarized as follows. First, the PLS LVs are calculated and then CCA is used to further process them and to calculate the final predictions in Eq. B6. In this article, only one PLS-CCA LV can be used due to the single output variable and CCA algorithm itself, regardless of the number of PLS LVs. Thus, only the underlying systematic glycemic variability that is closely related to the output variable (predicted glucose concentration) is captured by the single PLS-CCA LV. The loading coefficient for the output variable, scalar q_c, is obtained by regressing $\mathbf{y}$ on $\mathbf{t}_c$ which indicates how much $\mathbf{t}_c$ contributes to $\mathbf{y}$. Then the future glucose prediction $\hat{\mathbf{y}}_c$ and the prediction error $\mathbf{f}(N \times 1)$ can be calculated after.

A flowchart for the proposed modeling strategy is given in Figure B1. The developed LVX/LV model is then used for online application to new data. During the online application, the newly available predictor vector $\mathbf{x}_{new}{}^T(1 \times J_x)$ at each sampling instant can be expressed as $[\mathbf{g}_{new}{}^T(1 \times L_G), \mathbf{u}_{I,new}{}^T(1 \times L_I), \mathbf{u}_{M,new}{}^T(1 \times L_M)]$ where L_G is the number of current and past glucose measurements and L_I and L_M are the corresponding values for the bolus insulin and meal CHO estimates, respectively. The data normalization of $\mathbf{x}_{new}{}^T$ is based on information obtained from training data; then the normalized $\mathbf{x}_{new}{}^T$ is projected onto the LVX/LV model in order to make the PH-step-ahead prediction, $\hat{y}_{new}$. The prediction error f_{new} can be calculated after PH samples when the new measurement y_{new} becomes available

$$t_{new} = \mathbf{x}_{new}{}^T \mathbf{w}_c \tag{B8}$$

$$\hat{y}_{new} = t_{new} q_c \tag{B9}$$

$$f_{new} = y_{new} - \hat{y}_{new} \tag{B10}$$

Manuscript received Dec. 3, 2012, and revision received Jun. 5, 2013.