Transactions on Materials, Biotechnology and Life Sciences
ISSN: 2960-1932, eISSN: 2960-2319 | Volume 7
MISBP 2024

BiGRU-Att: Blood Glucose Prediction with Bidirectional Recurrent Neural Networks and Attention Mechanisms

Wenshan Zhang*

School of Statistical and Actuarial Sciences, Western University, London, Ontario N6A 3K7, Canada

* Corresponding Author Email: wzhan779@uwo.ca

Abstract. This paper presents an advanced deep learning model that combines a bidirectional gated recurrent unit (BiGRU) and an attention mechanism for predicting blood glucose levels. The innovation of the model lies in its unique structural design, in which the BiGRU is able to capture the backward and forward dependencies in time series data, while the attention mechanism further enhances the sensitivity of the model to critical time steps, thus improving the accuracy of the prediction. This combination not only exploits the power of BiGRU in processing serial data, but also enables adaptive weighting of important features in the input data through the attention mechanism. We validate the effectiveness of the model through experiments on 10 in silicodatasets generated by the UVA/Padova T1D simulator. On these datasets, the model demonstrated excellent performance with a root mean square error (RMSE) of only 0.0719, a metric that is significantly lower than that of existing techniques, proving the superiority and innovativeness of our model for blood glucose prediction tasks.

Keywords: BiGRU; Attention; Residual Blocks; Diabetes Management.

1. Introduction

Blood glucose (BG) is a key indicator for the diagnosis of diabetes, especially for patients with type 1 diabetes, who require lifelong glucose management[1].Type 1 diabetes is usually caused by the destruction of pancreatic β-cells, which results in insufficient insulin secretion. With the rapid development of artificial intelligence technologies, machine learning algorithms such as support vector regression (SVR)[2] and artificial neural networks (ANN)[3] have been used to assist in the prediction of blood glucose levels in patients with type 1 diabetes. However, these traditional methods suffer from performance bottlenecks when dealing with large-scale training datasets, especially when dealing with complex datasets. Therefore, it becomes crucial to explore deeper network structures to improve prediction accuracy and model complexity.

In recent years, researchers have begun to explore the use of Recurrent Neural Networks (RNN)[4] and their variants such as Long Short-Term Memory Networks (LSTM)[5] and Gated Recurrent Units (GRU)[6]. These models are valued for their ability to capture temporal dependencies when processing sequential data. However, LSTM and GRU models still face challenges in some cases, including the problem of vanishing or exploding gradients, and efficiency issues when dealing with very long sequences. In addition, these models may struggle to capture the full range of complex features and long-term patterns in the data, limiting their performance in glycaemic prediction tasks.

To address the limitations of existing models, this study proposes an innovative deep learning model that combines a bi-directional gated recurrent unit (BiGRU) and an attention mechanism. The BiGRU enhances the model's ability to understand sequential data by simultaneously processing past and future information, while the attention mechanism[7] allows the model to adaptively focus on key parts of the input sequence. In addition, our model employs dilated causal convolution and residual concatenation to improve the ability to capture complex patterns and solve the gradient propagation problem in deep networks.

The paper is structured as follows: Section 2 will introduce related research and existing techniques; Section 3 describes in detail the proposed model architecture and methodology; Section 4 demonstrates the experimental setup, dataset, and evaluation metrics; Section 5 provides the experimental results and analyses; and, finally, Section 6 summarizes the full paper and discusses future research directions.

2. Related Works

2.1. GRU

The Gated Recurrent Unit (GRU) is a variant of Recurrent Neural Networks (RNN) widely used in sequence modelling. GRU solves the problem of vanishing or exploding gradients encountered by traditional RNNs when dealing with long sequences by introducing update gates and reset gates. The update gate controls the flow of information from the previous state to the current state, while the reset gate allows the model to ignore or include previous state information. These features of GRU make it perform well in natural language processing, speech recognition, and other sequence prediction tasks. For example, Ioannou et al used GRUs to model sequential data from medical records to predict a patient's risk of readmission [8].

The principle behind GRU is very similar to that of LSTM, i.e., a gating mechanism is used to control the input, memory and other information to make a prediction at the current time step, and LSTM is briefly introduced next. LSTM, short for Long Short-Term Memory, is a type of Recurrent Neural Network (RNN)[9]. Compared to the basic RNN, LSTM performs better in addressing the issues of gradient vanishing and exploding in long sentences [10]. This is primarily due to LSTM's "gating mechanisms", which include the input gate, forget gate, output gate, and the memory cell.

$$i_t = \sigma(w_t \times [h_{t-1}, x_t] + b_i) \tag{1}$$

$$f_t = \sigma(w_f \times [h_{t-1}, x_t] + b_f) \tag{2}$$

$$o_t = \sigma(w_o \times [h_{t-1}, x_t] + b_o) \tag{3}$$

$$c_t = f_t \times c_{t-1} + i_t \times \tanh(W_c \times [h_{t-1}, x_t] + b_c) \tag{4}$$

$$h_t = o_t \times \tanh(c_t) \tag{5}$$

The core of LSTM lies in its cell state, which has the ability to remember and utilize information over long periods. The structure of an LSTM network includes an input gate, a forget gate, an output gate, and a cell state. These gating mechanisms work together to determine what information should be retained, what should be forgotten, and how to update and output information. The input gate i_t decides which information needs to be remembered. Based on the current input and the hidden state from the previous time step, it outputs a value between 0 and 1 using the sigmoid function [11], indicating the retention level of the corresponding information. The forget gate $f_t$determines what information should be forgotten. Similarly, it takes the current input and the previous hidden state as inputs and outputs a forget coefficient through the sigmoid function, controlling the retention level of the cell state from the previous time step. The output gate o_t decides the output of the cell state. It considers the current input, the previous hidden state, and the updated cell state, and jointly determines the final output value through the sigmoid and tanh functions. The cell state h_t is responsible for storing and updating information in the sequence. It updates the cell state based on the calculations from the input gate, forget gate, and the update cell, for subsequent use.

GRU has two has two gates, a reset gate and an update gate. Intuitively, the reset gate determines how the new input information is combined with the previous memory, and the update gate defines the amount of the previous memory that is saved to the current time step. If we set the reset gate to 1

and the update gate to 0, we will once again have the standard RNN model. The basic idea of using a gating mechanism to learn long-term dependencies is the same as for LSTM [12].

2.2. Attention Mechanism

Attention mechanisms [13] were initially widely used in the field of neural machine translation, where the core idea is that the model is able to focus on the most relevant parts of the sequence data at the current moment when processing it. This mechanism works by calculating the weights of each element in the input sequence and then aggregating the weighted elements to generate the model's output. The introduction of the attention mechanism has significantly improved the ability of the model to capture key information in the input data, leading to breakthroughs in a variety of tasks. Wang et al applied the attention mechanism in the field of blood glucose prediction and improved the prediction accuracy by enabling the model to focus on the key time points of blood glucose changes.

$$Q = XW_q \tag{6}$$

$$K = XW_k \tag{7}$$

$$V = XW_v \tag{8}$$

$$\text{attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{9}$$

Where W_q, $W_k$and W_v are linear change matrices; Q, K and V are query vector, key vector and value vector respectively. d_k is the dimension of the query vector and key vector. Finally, the attention weight and value vector are weighted and summed to get the output of the self-attention mechanism.

Combining the use of GRU with the attention mechanism can further improve the performance of the model in sequence prediction tasks. The combination of the long and short-term memory capabilities of GRU with the focusing capabilities of the attention mechanism allows the model to not only capture long-term dependencies in a time series, but also to identify and highlight the pieces of information that are most important for the current prediction. show in their work how this combination how it can improve the performance of language models.

3. Methods

3.1. Model Overview

The BiGRU-ATT model proposed in this paper is a state-of-the-art deep learning architecture designed to improve the accuracy of blood glucose level prediction by combining bidirectional gated recurrent units (BiGRUs) and attention mechanisms. The model is particularly suitable for analyzing and predicting time series data, capturing long-term dependencies in the series and enhancing the prediction by adaptively focusing on key information. The overall design of BiGRU-Att model is shown in Figure 1.

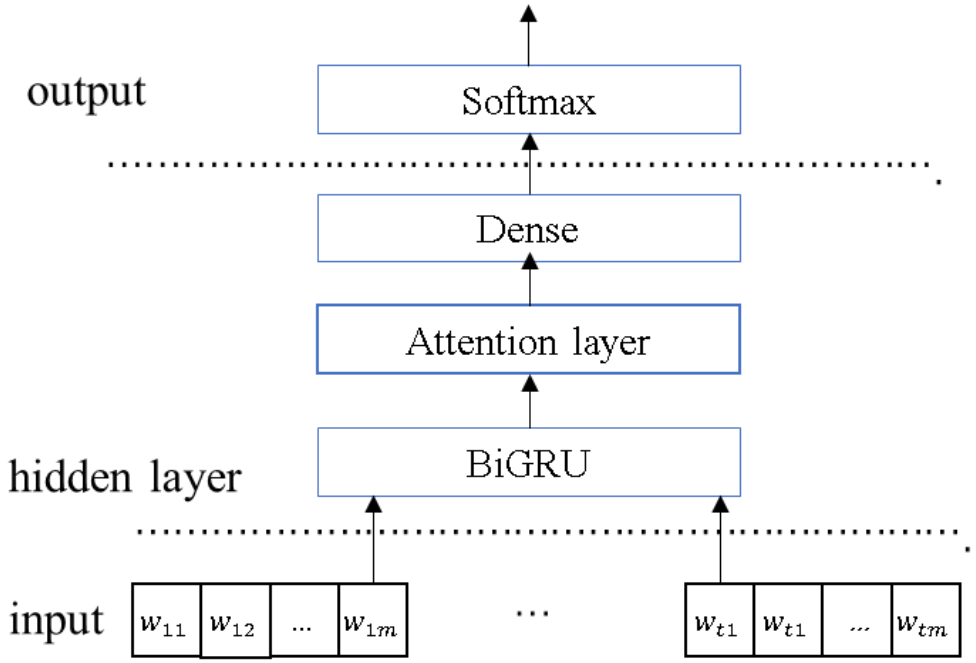

Figure 1. Performance of BiGRU-Att

3.2. Model Architecture

The bidirectional GRU layer is one of the core components of the model. Compared with the traditional unidirectional GRU, the bidirectional GRU is able to consider both forward and backward information of sequences, thus capturing the contextual dependencies of sequences more comprehensively. In this model, the bi-directional GRU layer has an input dimension of input size and a hidden layer size of hidden size and is set to batch first=True to accommodate the format of batch data.

The attention layer is used to weight the output of the bidirectional GRU layer. The layer obtains the attention weights for each time step by matrix multiplication of a trainable weight vector self.attention_weights (of the shape (hidden_size * 2, 1), since the output dimension of the bidirectional GRU is twice that of the unidirectional one) with the hidden states of the bidirectional GRU. Subsequently, these weights were normalized to a probability distribution using the SoftMax function to represent the importance of different time steps. Finally, the normalized attention weights were multiplied with the hidden states of the bidirectional GRU to achieve the weighting operation.

The fully connected layer (also known as the output layer) receives the attention-weighted sequence representation (obtained by time-step summation) and maps it to the final output space. In this model, the input dimension of the fully connected layer is hidden_size * 2 (the same as the output dimension of the bi-directional GRU), and the output dimension is output_size, which represents the number of target variables predicted by the model.

In terms of methodology, a backpropagation algorithm [14] is usually used to train the network. The loss function can be chosen according to the specific prediction task, e.g., mean square error (MSE) for regression problems and cross-entropy loss for classification problems. Optimizers[15] can be chosen such as Adam, RMSprop, etc. to adjust the weights of the model during the training process.

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{10}$$

4. Experimental

4.1. Data Set

This study used a dataset generated by the UVA/Padova T1D simulator[16], which is a US Food and Drug Administration (FDA)-approved tool for glucose level simulation. The dataset covers the in silico simulation of 10 virtual adult type 1 diabetic patients with a simulation period of 360 days, containing data from three meals per day. To ensure high temporal resolution of the data, data points were collected every 5 minutes, yielding a total of 288 data observations per day. The dataset contains six key variables: blood glucose level, insulin dose, carbohydrate intake, time stamp, glucose emergence, and plasma insulin level. Together, these variables provide a comprehensive view of the glycaemic dynamics of patients with simulated type 1 diabetes. In particular, plasma insulin level and glucose emergence rate are key predictors of blood glucose levels. Insulin dosing in the dataset mimics daily injection patterns that coincide with mealtimes, reflecting real-world dietary and therapeutic habits.

4.2. Experimental Setup

The data was loaded from a .mat file named "adult360_1.mat". The file contains a dataset called history. We extracted the following three variables: blood glucose (CGM), insulin (u6), and meal size (u1), and normalized them: normalized blood glucose to 100 mg per deciliter, converted insulin to units based on body weight (BW), and converted meal size to time steps (Ts).We also generated an array representing the index of days and integrated all the processed data into a complete dataset stored as a pandas Data Frame. In order to convert the time series data into inputs and outputs suitable for model training, we defined a function to generate the dataset. The function uses a sliding window approach to extract fixed-length input sequences and corresponding prediction targets from the complete dataset. Specifically, we use data from the first 24-time steps to predict the blood glucose value at the 30th minute.

In this study, we adopted a systematic dataset division strategy to ensure the scientificity and effectiveness of model training and evaluation. The dataset was divided into training, validation and test sets, with a specific ratio of 70% for training, 15% for validation and the remaining 15% as the test set. This division is designed to ensure that the model is trained on a sufficient amount of data, while retaining enough data for independent validation and testing of model performance. Model training was performed using the Adam optimizer, which is widely used in deep learning due to its adaptive learning rate. We set the initial learning rate to 0.001, which is a commonly used starting point to achieve faster convergence in the early stages of training. The loss function was chosen to be the mean square error (MSE), which is a common loss function used in regression problems to measure the difference between predicted and actual values. In order to adequately train the model and avoid premature convergence to a local optimum, we set a number of 50 training rounds. In each round of training, the model iterates over the entire training set, constantly updating the network weights through a backpropagation algorithm. We also set a random seed of 12345 to ensure the reproducibility of the experiment. During the training process, we closely monitored the training loss and validation loss to evaluate the performance of the model on both the training and validation sets. With an early stopping strategy, we stopped training to prevent overfitting when the validation loss did not decrease significantly over multiple consecutive epochs.

5. Result

5.1. Model performance indicators

This section will show the performance comparison of the BiGRU-ATT model with two commonly used recurrent neural network models, LSTM and GRU. All models were trained and tested on the same dataset to ensure a fair comparison. Several metrics were used for the performance evaluation,

including loss to test (MSE), mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and coefficient of determination (R^2).

This section will show the performance comparison of the BiGRU-ATT model with two commonly used recurrent neural network models, LSTM and GRU. All models were trained and tested on the same dataset to ensure a fair comparison. Several metrics were used for the performance evaluation, including loss to test (MSE), mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and coefficient of determination (R^2).

5.2. Comparison of results

Firstly, the results, as shown in Table 1, show that in terms of the loss on test (MSE), the BiGRU-ATT model outperforms the LSTM's 0.0089 and the GRU's 0.0089by 0.0052. This result indicates that the BiGRU-ATT model has the smallest average prediction error over the entire test set, reflecting the model's overall higher prediction accuracy. Further observing the RMSE metric, the BiGRU-ATT model also exhibits the lowest error value of 0.0719, compared to the RMSE of 0.0946 and 0.0942 for LSTM and GRU, respectively. As an important measure of the model's prediction accuracy, lower values of RMSE imply less deviation between the model's predicted value and the actual value, thus verifying that the BiGRU- ATT model's advantage in prediction accuracy. On the MAE indicator, the BiGRU-ATT model's 0.0546 further proves its superiority in the average of absolute values of prediction errors. Meanwhile, the MAPE metric also showed the superiority of the BiGRU-ATT model in terms of relative error, with its MAPE of 4.4038% lower than that of the LSTM of 5.6229% and that of the GRU of 5.4688%, which indicated that the BiGRU-ATT model had less relative error in predicting changes in glycaemic levels. The R^2 value, which measured the model's ability to account for the variability in the data, the R^2 value of the BiGRU-ATT model was 0.9605, which was higher than the LSTM's 0.9316 and GRU's 0.9322. This result suggests that the BiGRU-ATT model is better at explaining data variability and better at capturing the underlying patterns in the data.

Table 1. Result Comparison

Indicators	LSTM	GRU	BiGru-Att
MSE	0.0089	0.0089	0.0052
RMSE	0.0946	0.0942	0.0719
MAE	0.0714	0.0707	0.0546
MAPE	5.6229	5.4688	4.4038
R^2	0.9316	0.9322	0.9605

Considering all the performance metrics together, the BiGRU-ATT model demonstrates significant advantages in terms of prediction accuracy, error distribution, and data interpretation ability (as shown in Figure 2). We believe that this superiority mainly stems from the innovative design of the model structure: the two-way GRU layer is able to capture both forward and backward dependencies of the time series data, while the attention mechanism enables the model to focus on the time steps that are most critical for prediction. Moreover, the balanced performance of the BiGRU-ATT model on all evaluation metrics reflects its high robustness in different aspects. This is particularly important in real-world applications, where blood glucose prediction models need to provide stable and reliable predictions in all situations.

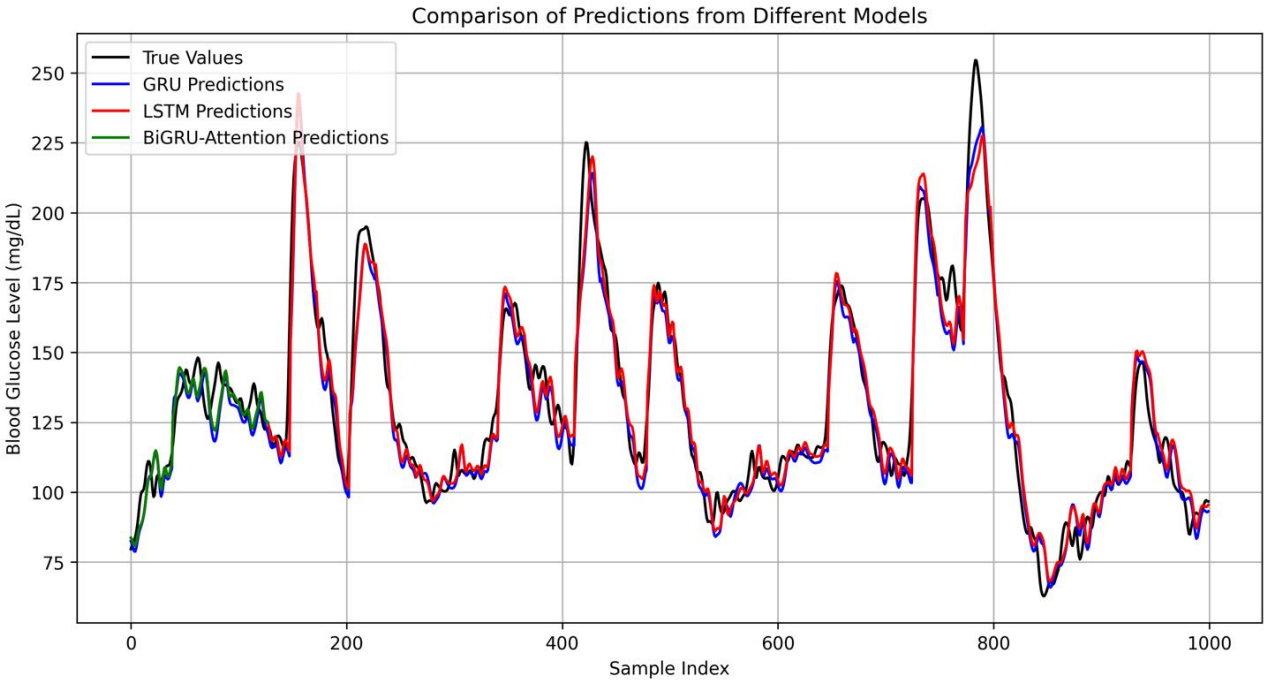

Figure 2. Comparison of Predictions from Different Models

6. Conclusions

In this study, we successfully proposed a deep learning model, BiGRU-ATT, combining bidirectional gated recurrent unit (BiGRU) and attention mechanism for predicting blood glucose levels in type 1 diabetes patients. Through a series of experiments, we verified the significant advantages of the model in terms of prediction accuracy. The BiGRU-ATT model demonstrated lower root mean square error (RMSE) and higher coefficient of determination (R^2) on the test set, and performed outstandingly compared to other recurrent neural network models.

The main contribution of the BiGRU-ATT model is its innovative structural design, which enables the model to process time series data more efficiently and capture the dynamics of key time steps. The introduction of the bi-directional GRU layer allows the model to consider both past and future information of the sequence data, while the attention mechanism further enhances the model's ability to recognize important features. Together, these features contribute to the model's high accuracy in blood glucose prediction tasks.

In summary, the BiGRU-ATT model provides a new deep learning approach for glycaemic prediction in patients with type 1 diabetes. The superior performance of the model is not only reflected in the experimental results, but also in the fact that it provides a more accurate and reliable glucose management tool for patients. With the continuous advancement of technology and the increasing abundance of medical data, we believe that the BiGRU-ATT model will play an important role in diabetes management in the future.

References

[1] M. A. Atkinson, G. S. Eisenbarth, A. W. Michels, “Type 1 diabetes,”The Lancet, vol. 383, issue 9911, pp. 69-82, ISSN 0140-6736, 2014.

[2] E. I. Georga, V. C. Protopappas, D. Ardigo, M. Marina, I. Zavaroni,D. Polyzos, and D. I. Fotiadis, “Multivariate prediction of subcutaneous glucose concentration in type 1 diabetes patients based on support vectorregression,” IEEE Journal of Biomedical and Health Informatics, vol.17, no. 1, pp. 71–81, Jan 2013.

[3] Agatonovic-Kustrin S, Beresford R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research[J]. Journal of pharmaceutical and biomedical analysis, 2000, 22(5): 717-727.

[4] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.

[5] Yu Y, Si X, Hu C, et al. A review of recurrent neural networks: LSTM cells and network architectures[J]. Neural computation, 2019, 31(7): 1235-1270..

[6] Taylor, Sean J, and Benjamin Letham. 2018. “Forecasting at scale.” The American Statistician72 (1): 37–45.

[7] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[8] Chung, J., Gülçehre, Ç., Cho, K., & Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. ArXiv, abs/1412.3555.

[9] Li, F., Zhang, M., Fu, G., Qian, T., & Ji, D. (2016). A Bi-LSTM-RNN Model for Relation Classification Using Low-Cost Sequence Features. ArXiv, abs/1608.07720.

[10] Kumar, J., Goomer, R., & Singh, A.K. (2018). Long Short Term Memory Recurrent Neural Network (LSTM-RNN) Based Workload Forecasting Model For Cloud Datacenters. Procedia Computer Science, 125, 676-682.

[11] Han J, Moraga C. The influence of the sigmoid function parameters on the speed of backpropagation learning[C]//International workshop on artificial neural networks. Berlin, Heidelberg: Springer Berlin Heidelberg, 1995: 195-201.

[12] Jamadar, D.J., Pittala, R.B., Silpa, D.G., Arunkumar, P., Ahmad, A., & Prasad, D.G. (2024). GRU-RNN Model to Analyze and Predict the Inflation by Consumer Price Index. Journal of Electrical Systems.

[13] Zhang, J., Wang, P., & Gao, R.X. (2020). Attention Mechanism-Incorporated Deep Learning for AM Part Quality Prediction. Procedia CIRP, 93, 96-101.

[14] Rojas R, Rojas R. The backpropagation algorithm[J]. Neural networks: a systematic introduction, 1996: 149-182.

[15] Le Q V, Ngiam J, Coates A, et al. On optimization methods for deep learning[C]//Proceedings of the 28th international conference on international conference on machine learning. 2011: 265-272.

[16] Pennant M E, Bluck L J C, Marcovecchio M L, et al. Insulin administration and rate of glucose appearance in people with type 1 diabetes[J]. Diabetes Care, 2008, 31(11): 2183-2187.