Prediction Model for Glucose Metabolism Based on Lipid Metabolism*

Y. Hatakeyama1; H. Kataoka1; N. Nakajima1; T. Watabe1; S. Fujimoto2; Y. Okuhara1

1Center of Medical Information Science, Kochi University Medical School, Kochi, Japan;
2Department of Endocrinology, Metabolism and Nephrology, Kochi University Medical School, Kochi, Japan

Keywords
Prediction model, diabetes mellitus, probability model, particle filter

Summary
Objectives: We developed a robust, long-term clinical prediction model to predict conditions leading to early diabetes using laboratory values other than blood glucose and insulin levels. Our model protects against missing data and noise that occur during long-term analysis.

Methods: Results of a 75-g oral glucose tolerance test (OGTT) were divided into three groups: diabetes, impaired glucose tolerance (IGT), and normal (n = 114, 235, and 325, respectively). For glucose metabolic and lipid metabolic parameters, near 30-day mean values and 10-year integrated values were compared. The relation between high-density lipoprotein cholesterol (HDL-C) and variations in HbA1c was analyzed in 158 patients. We also constructed a state space model consisting of an observation model (HDL-C and HbA1c) and an internal model (disorders of lipid metabolism and glucose metabolism) and applied this model to 116 cases.

Results: The root mean square error between the observed HbA1c and predicted HbA1c was 0.25.

Conclusions: In the observation model, HDL-C levels were useful for prediction of increases in HbA1c. Even with numerous missing values over time, as occurs in clinical practice, clinically valid predictions can be made using this state space model.

Correspondence to:
Yutaka Hatakeyama
Center of Medical Information Science
Kochi University Medical School
Oko-cho Kohasu, Nankoku
Kochi, Kochi 783-8505
Japan
E-mail: hatake@kochi-u.ac.jp

1. Introduction

Lifestyle-related diseases associated with aging and daily living conditions are increasing and have become a serious problem in many countries [1], as they can lead to cerebrovascular and cardiovascular complications. Lifestyle-related diseases develop as a result of both congenital factors and various acquired and environmental factors. If efforts are made to eliminate major risk factors related to lifestyle, these diseases can be prevented or their outcomes improved [2, 3].

Predictive medicine is important to prevent disease onset and progression. In patients with lifestyle-related diseases, early warnings about clinical conditions are known to improve outcomes [4, 5]. Predictive medicine requires that changes in a target disease be understood and that methods be developed to describe and predict these changes. These methods should be based on data easily obtained in daily clinical practice and routine medical checkups. If such methods are used, physicians will be able to predict future changes and offer treatment and advice to patients. They may also be able to motivate patients to improve negative lifestyle habits.

Type 2 diabetes, a lifestyle-related disease that has increased in prevalence in recent years, can be prevented or delayed in many patients [1]. Many mathematical models of glucose metabolism in diabetes have been proposed [6–10]. Typical models include glucose-insulin models [11–16] and mathematical models combined with pancreatic β-cell models [17, 18]. These models are generally limited to specific clinical situations, such as short-term intensive care unit patients, where periodic laboratory data can be obtained. In addition, these models predict target changes for a few hours, or at most, several months. These mathematical models provide accurate predictions due to the fixed nature of the environment where data are collected and input into the model.

However, when targeting patients with early diabetes in outpatient settings or routine medical checkups, prediction of changes over years is necessary. However, various problems can occur when using the above mathematical models in this situation. First, insulin testing is not performed in routine clinical practice; thus, data do not exist and cannot be input into the model. Second, data on whether blood glucose measurements represent fasting or post-prandial levels do not exist and cannot be input into the model. Third, frequent data acquisition is not necessarily performed on a regular basis and missing values must therefore be addressed, making model implementation
difficult. Finally, the accumulation of measurement-related and daily variation-related errors make long-term predictions difficult. If markers to represent early diabetes other than blood-glucose or insulin-related data can be identified from tests performed in routine clinical practice, a robust model can be constructed that protects against missing data and noise that occurs with long-term analysis.

2. Objectives

This paper reports on the construction of a prediction model based on data at KMS Hospital that includes HbA1c data and data linked to other laboratory values. Previous studies [19, 20] suggest that after adjustment for several metabolic syndrome components, lower high-density lipoprotein cholesterol (HDL-C) levels were independently related to a risk of future type 2 diabetes in a predominantly Caucasian population [19] and in Korean subjects [20]. Moreover, the analysis results [19] demonstrated that low-onset type 2 diabetes mellitus was related not only to lower HDL-C levels but was also strongly related to lower HDL-C-to-apolipoprotein A1 and HDL-C-to-apolipoprotein A2 ratios. This paper defines a prediction model based on a hypothesis where glucose metabolism is linked to risk factors evaluated by HDL-C levels according to these previous studies. Therefore, we analyzed changes in HDL-C levels before diabetes onset.

For a long-term prediction model, a model indirectly describing the prediction ability of lipid metabolism would be more suitable than a direct relational model, such as a glucose-insulin model, for minimizing the influence of noise. Moreover, testing for data needed in a glucose-insulin model is not routinely performed in clinical practice, so this model cannot be used. In this paper, a state space model [21] was constructed using test data reflecting an internal observation model. More specifically, we used HDL-C levels, which reflect disordered lipid metabolism and the internal state of these levels was considered a risk of future type 2 diabetes. We also used HbA1c levels, which reflect the internal state of disordered glucose metabolism. The model was applied to data gathered by KMS Hospital to examine its validity.

3. Methods

The prediction model for HbA1c of outpatients was constructed using outpatient data stored at KMS Hospital. The prediction model was defined as a probability model where HbA1c is more likely to increase in the case of lower HDL-C levels based on suggestions in previous studies. This paper conducted two analyses to evaluate the hypothesis and the prediction model. The first analysis tested the hypothesis that disorders of HDL-C values occurred before disorders of glucose metabolism assessed by a 75-g oral glucose tolerance test (OGGT) data. This paper introduces the notion of a “state space model” [21] into the prediction model and confirms the validity of the model. The second analysis evaluated the prediction results applied to 116 cases based on root mean square error between observed HbA1c and predicted HbA1c.

KMS Hospital has a long-term, large-scale database with accumulated data from 280,000 patients since 1981, including 32,000 patients with data for ≥10 years. The data are anonymous and stored in a data warehouse for research and educational use. Analysis was performed based on test data adjusted by measured conversion factors and whose accuracy is guaranteed by the internal and external quality control of the clinical laboratory department at KMS Hospital.

3.1 Feature Extraction by OGGT Data and Observation Model Selection

The first analysis tested that a disorder of HDL-C values occurred before type 2 diabetes onset. Moreover, the validity of HDL-C levels as a predictive factor was compared with triglyceride (TG) and low-density lipoprotein cholesterol (LDL-C) levels to include other disorders of lipid metabolism. In patients who underwent a 75-g OGGT, data used for analysis included measured baseline blood glucose, 2-h blood glucose, baseline insulin, and 1-h insulin. Patient data were classified as described below into three groups: diabetes (DM), impaired glucose tolerance (IGT), and normal.

DM was defined as baseline blood glucose (mg/dl) ≥ 126, 2-h blood glucose (mg/dl) ≥ 200, and 1-h insulin (µU/ml) > baseline insulin (µU/ml). IGT was defined as baseline blood glucose (mg/dl) 110 to < 126, 2-h blood glucose (mg/dl) 140 to < 200, 1-h insulin (µU/ml) > baseline insulin (µU/ml), and homeostatic model assessment of insulin resistance (HOMA-IR) < 2. Normal was defined as baseline blood glucose (mg/dl) < 110, 2-h blood glucose (mg/dl) < 140, 1-h insulin (µU/ml) > baseline insulin (µU/ml), and HOMA-IR < 2.

The number of data points in each group was normal, 325 (mean age, 44.28 ± 17.18 years; males, 71); IGT, 235 (mean age, 50.66 ± 15.51 years; males, 68); and DM, 114 (mean age, 57.89 ± 14.47 years; males, 54). For each of the three groups, data were totaled by two methods based on the day when OGGT was performed: 1) near data, which represented mean values within 30 days from day of OGGT; and 2) 10-year data, which represented data over 10 years, defined as: Value = [All data integrated values within 10 years from day of OGGT]/[Number of days].

Significant differences between groups in total values for each of these laboratory values were analyzed, and the validity of selecting HbA1c and HDL-C levels for the observation model was examined. Comparison of OGGT results between the three groups was examined by the Kruskal-Wallis test. Then, for each pair, multiple comparisons were performed by Mann-Whitney U-test with Bonferroni-adjusted p values. Statistical analysis was performed using R (version 2.13.0; R Foundation for Statistical Computing, Vienna, Austria). The level of statistical significance was set at p = 0.05.

3.2 Prediction Model

It is difficult to define the relation between HbA1c and HDL-C levels directly. However, these values can be regarded as a dynamic system (internal model) that determines time changes in metabolic conditions. If results of directly observable tests...
change with these metabolic capacities, an internal model is identified based on the test results. These observable models are defined as observation models. A pair of these two models is constructed as a "state space model". Figure 1 shows an overview of the constructed state-space model. The internal model of disordered lipid metabolism only affects the internal model of disordered glucose metabolism based on the hypothesis that disordered lipid metabolism was considered a risk factor for future type 2 diabetes. Each internal model corresponds to HDL-C and HbA1c in the observation model. The values of the "observation" model in the prediction model are not actual measurement data but predicted data and are the output of the prediction model.

When applying clinical data to this model, the occurrence of noise related to observations (e.g., daily variations and observation errors) cannot be avoided. Therefore, this effect is defined by a stochastic model where noise is added. In this study, the state space model at time t is defined as follows:

\[
L(t) = L(t-1) + \omega \cdots (1)
\]

\[
G(t) = G(t-1) + f(G(t-1), L(t-1)) + \omega \cdots (2)
\]

\[
\text{HDL-C}(t) = 95 - L(t) + \omega \cdots (3)
\]

\[
\text{HbA1c}(t) = 4.07 + 0.92 \exp\left(\frac{(G(t)-44.9)}{28.9}\right) + \omega \cdots (4)
\]

In Equations 1 and 2, L and G are disordered lipid metabolism and disordered glucose metabolism in the internal model and defined as higher values indicating worsening conditions. \( \omega \) is a noise term, and Gaussian noise is assumed in this study. The function \( f \) in Equation 2 is a function of L and G and is defined by fuzzy inference based on the stored data. In Equations 3 and 4, HDL-C and HbA1c are the HDL-C and HbA1c values calculated from each metabolic abnormality and indicate the observation model. In other words, because the internal model values cannot actually be defined, they are definitively defined according to the observation model. The time interval in the constructed model should ideally be small, but with clinical data, obtaining continuous clinical data is difficult. However, even if the time interval is somewhat large, the observation data can be allowed if the amount of variation can be followed in the constructed model. In this paper, considering the intervals when HbA1c levels are obtained in clinical practice, this is defined as \( dt = 90 \) days.

![Figure 1](image1)  
State space model for predicting glucose metabolism

![Figure 2](image2)  
HbA1c distribution vs. number of days. Test day when HbA1c was first \( \geq 6.5 \) is day 0.
The state space model consists of the internal model (disordered lipid metabolism, disordered glucose metabolism) and the observation model (HDL-C, HbA1c). The observation model is defined from the internal model and represents predicted values, not actual measurement data. The hypothesis that lower HDL-C levels are a risk factor for future type 2 diabetes is simply applied to the relation between these internal models.

In the observation model (Equations 3, 4) and in Equation 2, function $f$ is defined based on the data accumulated to date at KMS Hospital.

Figure 2 shows data for patients with multiple HbA1c tests, demonstrating the distribution of HbA1c values and number of elapsed days, with the day of testing when the value was first $\geq 6.5$ as baseline ($t = 0$). This value used as the baseline value is defined based on diagnostic criteria for diabetes. As shown by the graph distribution, the distribution of HbA1c changes differed before and after baseline. In other words, in the clinical data, HbA1c values over time showed nonlinear changes rather than linear changes. A nonlinear distribution was also reported in a large-scale cohort study [22].

This paper assumes that glucose metabolism worsens with time in diabetic patients, and that this is reflected by nonlinearity of HbA1c values, while Equation 4 is defined as an exponential model based on a time series distribution. The parameters in the definitional equation are set such that they fit with the data shown in Figure 2. The observation model for HDL-C is implemented as a model that decreases if there is disordered lipid metabolism. The model of disordered lipid metabolism signifies that changes in a 90-day period are small. When applied to actual data, this corresponds to time series changes according to the noise term. In the distribution of test values at KMS Hospital.

Figure 3 shows the outputs of function $f$ based on this inference. The graph values show function $f$; the red regions show plus variation, and blue regions show minus variation for each output. The horizontal axis and vertical axis show G and L, respectively, which are the model input values. Disordered glucose metabolism at the following time of Equation 2 is determined by the variation and noise term.

Internal models $G$ and $L$ and observation models HbA1c and HDL-C are shown as probability distributions rather than fixed values, but when dealing with clinical data, the influence of noise due to various factors must be considered. Therefore, the probability distribution cannot be assumed to be a parametric model, and formulation of each model distribution is difficult. For this reason, the constructed model is implemented using a particle filter [23, 24]. As mentioned previously, the noise term $\omega$ in Equations 1–4 is assumed to be a Gaussian distribution.

A particle filter is an appropriate filter for state prediction by repetitive resampling, prediction, and weighting. One particle in our constructed model consisted of the internal model parameters and estimated observed values. The constructed model was implemented with a particle number of 20,000. As likelihood in the resampling step in particle filtering, normal distribution values based on the Mahalanobis distance is used.
nobis distance were used. The clinical data for defining the Mahalanobis distance used data pairs for the respective HDL-C and HbA1c test days of < 14 days.

Actual clinical data from 116 patients that fulfilled the three following conditions were applied, and the root mean square error (RMSE) between the actual HbA1c and predicted HbA1c by the observation model was evaluated: 1) for data from patients prescribed oral hypoglycemic agents, insulin and hyperlipidemia drugs, only data before the first prescription was used; 2) only data from patients with a history of testing HbA1c and HDL-C ≥ 3 times was used; and (3) only data from patients with an initial HbA1c ≤ 5.9 was used.

The analysis data used to construct function $f$ shown in Figure 3 and the data used for model application were different patient data. Because the predicted values are expressed as a probability distribution, values at the center of the probability distribution at the time of model evaluation were used as predicted values. This was done to reflect features of the distribution because the probability distribution was not a normal distribution. Evaluation was performed by significance testing using the RMSE and Pearson product-moment correlation coefficient. The RMSE was calculated between HbA1c values of the observation model and actual measurement data.

4. Results

Table 1 shows the HDL-C values for “near data” in the DM, IGT, and normal groups. The $p$ values are shown for multiple comparisons by Bonferroni-adjusted Mann-Whitney U-test. Significant differences were seen by Kruskal-Wallis analysis. Comparisons between the IGT and DM groups showed significant differences.

Figure 4 shows an example of the predicted results. The horizontal axis is the time course, and one unit is 90 days. The initial test day is shown as Day 0. The left vertical axis shows HbA1c (%), and the right vertical axis shows HDL-C (mg/dl). The squares show the actual HbA1c (observed value), and the circles show the predicted HbA1c by the constructed model (predicted value). The predicted values are shown at the center of the probability distributions. The triangles show the actual HDL-C values. The RMSE for the patient data was 0.25.

Predictions were made for data from 116 patients, and for the predicted data other than the first test day (285 instances), comparisons of the predicted HbA1c and observed HbA1c showed an RMSE of 0.25. The mean period for calculation of the predicted data was 265.6 ± 406.4 days. The maximum and minimum absolute errors were 1.21 and 0.00, respectively. Figure 5 shows the relationship between predicted HbA1c and observed HbA1c. The horizontal axis shows the predicted values, and the vertical axis shows the observed values. A test for significance of the Pearson product-moment correlation coefficient showed significant differences ($r = 0.95$).

![Figure 4](image_url) Results of applying the prediction model to patient data. Circles, predicted HbA1c; squares, observed HbA1c; triangles, observed HDL-C

### Table 1 “Near data” laboratory values

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>IGT</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDL-C (mg/dl)</td>
<td>51.0 ± 17.8</td>
<td>60.5 ± 16.7</td>
<td>66.0 ± 18.7</td>
</tr>
<tr>
<td>n</td>
<td>74</td>
<td>158</td>
<td>183</td>
</tr>
</tbody>
</table>

5. Discussion

HDL-C was significantly different between the DM group and the IGT group. Because the model parameters require a variability of test values for long periods and may not be detected in the “near data”, this shows the need for analysis of distribution of values over a long period. When laboratory values differ over long periods in comparison with a normal group, a prediction model can be defined based on those values. Therefore, “long-term data” must be analyzed. ▶ Table 2 shows these results. HDL-C between the IGT and normal group showed a trend toward a difference, but the p value was 0.052; thus, based on the significance level of 0.05, the difference was not significant. This is probably because the number of data points acquired for HDL-C was insufficient. The individual differences for HDL-C were relatively small. Therefore, the use of HDL-C as an observation model was reasonable as a prediction model in this study.

As an evaluation of model validity, errors in results of the second analysis when applied to actual data were assessed. Considering that the input data were time series data that had missing values, and considering the influence of noise, the RMSE obtained was sufficient to demonstrate validity. This also shows that the correlation between predicted values and observed values was very high, and that predictions can be made with the constructed model.

Because of implementation using a particle filter, processing could be continued even with missing values. Based on operating characteristics of the particle filter, this means that the resampling step cannot be performed when there is a missing value. Namely, continued missing values means that the probability distribution width will become wider with the noise term. The RMSE was calculated with the values at the center of each probability distribution and the observed values. Because each probability distribution was not normal, values at the center change and errors increase. The results seen in ▶ Figure 4 also show that missing values continued for a long period. In particular, because HDL-C values greatly decreased when test data were missing from periods 4 through 18 in ▶ Figure 4, deviations occurred in the central predicted value and observed value.

When the missing period is long and HDL-C variations are large, such results are more likely, as can be seen from the input data. Because these results show the center of the distribution, the errors appear large, but particles containing observed values and equal predicted values are included in these results. This is because instances when HDL-C changed greatly were also included in the probability distribution, and data predictions after resampling correspond to actual data. In other words, because probability distributions that do not assume a normal distribution are used, various state changes and data adjustments can be dealt with, and new input data can be handled using the particle filter.

The input data of these prediction results were extracted from patient data in which some disease history was available. However, the data in the prediction input were omitted after intervention for disorders of glucose metabolism and lipid metabolism. Moreover, data from patients without a history of diabetes were included in the input. Therefore, the prediction

<table>
<thead>
<tr>
<th>Table 2</th>
<th>“Ten-year data” laboratory values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DM</td>
</tr>
<tr>
<td>HDL-C (mg/dl)</td>
<td>49.5 ± 12.5</td>
</tr>
<tr>
<td>n</td>
<td>50</td>
</tr>
</tbody>
</table>

Figure 5 Correlation between observed HbA1c and predicted HbA1c by the prediction model. Test for significance of Pearson product-moment correlation coefficient: r = 0.95, p < 0.01

6. Conclusion

A probability model to predict increases in HbA1C was useful when HDL-C was used as an observation model parameter. Even with more than 10 years of time series data, with numerous missing values, clinically valid predictions can be made by appropriately defining and applying a state space model.

Acknowledgments

This study was supported by a Grant-in-Aid for Scientific Research (B) (24590615) and a Grant-in-Aid for Young Scientists (B) (22790498).

References

5. de Groot M, Doyle T, Kushnick M, Shubrook J, Topp B, Promislow K, deVries G, Miura RM, Fineman DJ, Barhak J, Ye W. Indirect estimation of a model with numerous missing values, our model is useful for the application of clinical data with missing values.