Extracting Temporal Knowledge.

Christopher Cimino, M.D.[1] and Sheila A. Turken, M.D.[2]

[1]Office of Computer Based Education and Department of Neurology
Albert Einstein College of Medicine

[2]Family Medical Care, Inc.

We have attempted to examine the value of temporal knowledge for two physicians making predictions about blood glucose levels in diabetic patients. Our hypothesis is that physicians evaluating sequential data about a patient derive an abstract description of the patient that is not temporally related (atemporal knowledge). For example, a patient might be identified as a "brittle diabetic". Atemporal knowledge should allow a physician to make isolated predictions about a patient that are better than random chance. A combination of temporal and atemporal knowledge should allow a physician to make better predictions than atemporal knowledge alone. We used the AIM-94 diabetes datasets supplied by Dr. Michael Kahn. These datasets consist predominately of blood glucose measurements and insulin dosages.We failed to show any different in predictive ability between temporal and atemporal data, between the two physicians, and between the individual physicians and three different controls. We did show that by grouping all 20 physician cases (temporal and atemporal) their predictions had an average error of 69 mg/dl of blood glucose. These predictions were better than the average of one of the controls by 30 mg/dl (standard deviation of 6.5). While this is statistically significant, we do not feel it is clinically significant. It is our conclusion that physicians may not make accurate predictions based solely on insulin dosages and glucose measurements.

Introduction

Close control of blood glucose levels is important in avoiding complications of diabetes mellitus. Diabetics attempt to achieve this control by making multiple measurements of their blood glucose level, and use this information to adjust their insulin dosage. Insulin pumps have been developed in an attempt to automate this process. Algorithms have been developed to analyze glucose data and predict insulin needs (Fisher) in an attempt to further automate the process.

It is not clear to what extent detailed temporal data is useful as compared to abstract knowledge about a patient. For example, a patient might be identified as a "brittle diabetic". This knowledge along with a single day's data can used to make predictions that are better than random chance. It seems likely that adding temporal information, in the form of several days worth of data, will improve the prediction even more. We would like to try to quantify the relative worth of temporal and atemporal knowledge.

Dataset

The 70 case AIM-94 diabetes dataset supplied by Dr. Michael Kahn was used. In the diabetes data, there are events at discrete times that can be viewed as treatment events (insulin injections, meals, and activity levels) and diagnostic events (symptoms and glucose measurements). The vast majority of the events in the cases are insulin dosages and glucose measurements. Events describing exercise, hypoglycemic symptoms, and meal intakes are rare. Our primary assumption is that clinicians use their diagnostic assessment of the patient to make treatment decisions and that their diagnostic assessment is based on readily observable data.

In a normal treatment situation, the clinician would alter treatment based on his or her impression of the patient's condition.The datasets do not permit any alterations in treatment so we will ask the expert clinician to give a diagnostic assessment in the form of a prediction about the patient's glucose measurement. We therefore need to make a secondary assumption that diagnostic knowledge is closely related to therapeutic knowledge in the diabetes domain; in other words, knowledge used to choose appropriate insulin dosage is closely related to knowledge used to predict glucose levels. Five cases were selected at random from the dataset of 70 cases. Coded data was converted into English phrases such as "Pre-supper Glucose Measurement" or "NPH Insulin Dosage".

Subjects

The subjects were two physicians. One physician is board certified in Neurology and the other is board certified in Endocrinology.

Methods

Each physician was presented with data from the beginning of the case up until the first glucose measurement. The physician was told the time of this glucose measurement and asked to predict the value. This was repeated for the first twenty glucose measurements. The physician was allowed to review all previously presented data. The first ten measurements were considered a learning period. The second ten measurements were considered predictions made with temporal knowledge. The data typically covered the first 10 days of the case. The physician was permitted to make notes about the patient during this period.

Next the physician was presented with samples of data taken from random places within the case. Each random sample included exactly one glucose measurement and would end with the next glucose measurement. The physician would be asked to predict the second measurement. No previous data was available and no feedback was given concerning the accuracy of the predictions. This was repeated until 10 predictions were collected. These predictions were considered to have been made without temporal knowledge. This procedure was repeated with each of the five cases for both physicians.

Three sets of control predictions were developed. One control used no knowledge and predicted a value of 100 md/dl for all the measurements. The second control predicted a value of 100 mg/dl for the first measurement and predicted all subsequent measurements would be the same as the previous measured value. The third control predicted the value would be a random number between 80 and 300.

Each set of ten predictions were scored by averaging the absolute difference between the actual glucose measurement and the predicted glucose measurement. A lower score means a better average prediction.

Results

We collected a total of 60 samples: 30 physician samples and 30 control samples. The average score was 89.3. The best score by a physician was 34.4. The best control score was 59.4. Seven physician scores were better than all control scores. The worst score by a physician was 117.8. The worst control score was 146.6. Four control scores were worse than all physician scores.

Learning, temporal, and atemporal scores were compared for physicians individually and combined on a case-matched basis. There was a trend toward improved scores between the learning phase and the temporal phase. This trend was not statistically significant. Assuming the trend continued, we would need to collect another 20 cases for each physician to achieve significance. There was a smaller trend for improved performance on the atemporal scores as compared to the temporal scores. Assuming this trend continued, we would need to collect at least 50 more cases for each physician to achieve significance. There was a small trend for the endocrinologist's temporal scores to be better than the neurologist's and for the neurologist's atemporal scores to be better than the endocrinologist's. We estimate at least 50 cases for these differences to reach significance.

Of the controls, the average score for the "no knowledge" control was the best (98.6). The average score for the random control was second best (105.3). The average score for the "previous value" control was the worst (107.2). None of these differences were significant. The largest trends were between the physicians and the "previous value" control. When combined, the physicians temporal scores averaged 28.8 mg/dl better (s.d. 27.9). The physicians atemporal scores averaged 32.4 mg/dl better (s.d. 25.8). We estimate 5 to 10 additional cases would be required for these differences to reach significance. When grouping the physician temporal and atemporal scores they were significantly better than the "previous value" scores by an average of 30.6 mg/dl (s.d. 6.5).

Discussion

There are three possible explanations for the lack of difference between the samples in this study. The most obvious explanation is that more cases should have been used to reach statistical significance. While it is probably true that we could show a statistically difference between physicians and the controls, it is not likely that this difference would be clinically significant. At best, its seems the physicians would have made errors on the order of 50-80 mg/dl of blood glucose while the controls would make errors of 80-110 mg/dl. If a physician can not easily and clearly outperform a simple algorithm on five cases, then there is something wrong that goes beyond just the number of samples collected.

A second explanation is that while physicians may be very good at choosing appropriate therapy, they are very poor at predicting blood glucose levels. There may be no correlation between these two

abilities. Their poor performance goes undocumented because they are never called on to predict glucose levels. This explanation seems unlikely to us.

The third explanation is that the available data is insufficient for physicians to accurately predict blood glucose levels. The vast majority of the data consisted of insulin dosages and glucose measurements. There were only rare pieces of information concerning exercise and meals.

Diabetes mellitus, in its most reductionist definition, is a disorder characterized by a deficiency of endogenous insulin action. As such, it can be treated by insulin administration. The two parameters of insulin dose and resultant glucose level are obviously two key descriptive characteristics of a patient at any given time. On the other hand, diabetes management can be complex and ultimately unpredictable. There are a multitude of other informational factors that are essential in determining a blood sugar level in an individual. Some factors in a patient are knowable, such as the age, weight, height, additional medical history, concurrent medications, activity level, and diet — other factors are not so discrete and quantifiable, such as the complex interaction of insulin action with that of the counterregulatory hormone cascade on any given occasion.

It is interesting to note that the atemporal predictions were slightly better than the temporal predictions. This seems to suggest that temporal knowledge plays little role in these predictions. We cannot make this conclusion in the face of the poor predictions overall. It is our conclusion that physicians can not make accurate predictions based solely on insulin dosages and glucose measurements.

References

Fisher, ME. A semi-closed loop algorithm for control of blood glucose levels in diabetics. IEEE Transactions on biomedical engineering. 38(1):57-61, 1991.

Correspondence

Please address correspondence to:
Christopher Cimino, M.D.
Albert Einstein College of Medicine
1300 Morris Park Avenue
Bronx, NY 10461
e-mail: cimino@aecom.yu.edu