# A Method for Identifying Temporal Progress of Chronic Disease Using Chronological Clustering

Sangjin Jeong[12], Chan-Hyun Youn[3], and Yong-Woon Kim[1]
[1]Protocol Engineering Center, ETRI, Daejeon, Korea
[2]Department of Information and Communications Engineering, KAIST, Daejeon, Korea
[3]Department of Electrical Engineering, KAIST, Daejeon, Korea
sjjeong@etri.re.kr, chyoun@kaist.ac.kr, qkim@etri.re.kr

***Abstract*—The development of an integrated and personalized healthcare system is becoming an important issue in the modern healthcare industry. One of main objectives of integrated healthcare system is to effectively manage patients having chronic disease. Different from acute disease, chronic disease requires long term care and its temporal information plays an important role to manage the status of disease. Thus, a patient having chronic disease needs to visit the hospital periodically, which generates large volume of medical data. Among the various chronic diseases, metabolic syndrome has become a major public healthcare issue in many countries. There have been efforts to develop a metabolic syndrome risk quantification and prediction model and to integrate them into personalized healthcare system, so as to predict the risk of having metabolic syndrome in the future. However, the development of methods for temporal progress management of metabolic syndrome has not been widely investigated. In this paper, we propose a method for identifying a temporal progress and patient's status of metabolic syndrome. Further, the effectiveness of the proposed method is evaluated using a sample patient data while emphasizing the capability to identify chronological changes of metabolic syndrome status.**

***Keywords—healthcare, chronic disease, metabolic syndrome, decision support system***

## I. INTRODUCTION

Different from acute disease, chronic disease requires long term care and its temporal information plays an important role to manage the status of disease. Several issues and problems of conventional healthcare system have been reported in [1]. Among them, difficulty in identifying important information from large datasets containing diverse types of information in a laboratory or clinical situation becomes an problem among physicians and clinicians. It is also important to find out whether a patient's disease status is under control and within the normal range. To resolve the limitation, various types of health information systems have been developed to support the physicians for caring patients in terms of temporal disease progress management. However, most of the health information systems are focused on visualizing temporal data in a timeline, while displaying specific events from the patient data. The overall goal of these health information systems is to present physicians temporal information contained in a record, improving their ability to recognize patterns for knowledge discovery and following treatment. They introduce simple visualization tools, but some include automated computational enhancements supporting it. However, due to the complexity of the underlying data, a lot of further work is required to address these intricacies rather than using simplistic approaches [2-4]. This implies that much information is not always feasible.

In this paper, we propose a method for identifying temporal progress of chronic disease, particularly metabolic syndrome (MS). The proposed method supports temporal progress care of MS using a novel similarity degree-based chronological distance analysis method [5]. The proposed method can be used to advanced chronic disease care services in which it combines the chronic disease diagnosis services and temporal progress care services.

In the following section, a description of the proposed method for temporal progress care of chronic disease is presented. Then, the effectiveness of the proposed method is experimentally evaluated using chronological medical examination results of a sample patient while emphasizing the capability to identify chronological changes of MS status. The experimental evaluation results show that our proposed method is effective for visualizing patient's temporal disease progress and managing a patient's disease status within control range.

## II. A METHOD FOR IDENTIFYING TEMPORAL VARIATION OF CHRONIC DISEASE BASED ON CHRONOLOGICAL CLUSTERING

Chronological clustering is a punctuation equilibribum model that is learning the sequence within biological communities. It is used to investigate the temporal variation of observations or data by analyzing the similarity between them. There are two required elements to perform chronological clustering method, namely connectedness and the fusion level. Connectedness means that the data or observations of data are related to each other as well as they are distributed over time domain. Sensitivity level denoted by $\alpha$, is the clustering tuning indicator. This indicator is the primary element in defining the similarity between adjacent observations. The indicator $\alpha$ is the tolerance ratio of similarity or corelation between observations [7-8].

The chronological clustering method has been applied to investigate the temporal progress of chronic diseases [9]. A chronic disease patient requires regular medical examination even though no noticeable symptom is observed. Thus, if two consecutive examination results show a slight variation

978-1-4673-5801-9/13/$26.00 ©2013 IEEE

between them, a physician may ignore the second examination results as it is a repetition of the previous one. Also, in that case, the patient may skip regular visits to a hospital to see a physician, which can save time and reduce the cost to the patient. To support this scenario, Hazzemi et al. [9] developed a distance function and a sensitivity of change function as follows:

**Definition 1.** *A medical report is a set of medical examination results containing one or more examination criteria represented as scalar values. A normalized report value indicates that a single integrated number of one or more examination results consisting the medical report, and the range of the single integrated number is from 0 to 1.Previous report and current reprot are chronologically ordered medical reports examined at $t_1$ and $t_2$, where $t_1 < t_2$ holds.*

**Definition 2.** *Given two consecutve medical report values containg N examination criteria, the distance function $d_N(2)$ is defined as:*

$$d_N(2) = \frac{\sum_{i=1}^{N} |\text{Prev. normalized report value - Curr. normalized report value}|}{\text{Number of examination criteria}}$$

Let the previous normalized report value and the current normalized report value be $VP_i$ and $VC_i$, respectively and $N$ denotes the number of examination criteria contained in a medical report. The ranges of $VP_i$ and $VC_i$ are from 0 to 1. For example, if a medical report contains two examination criteria such as blood pressure and weight, $N$ becomes 2. Then $d_N(2)$ can be expressed as

$$d_N(2) = \frac{\sum_{i=1}^{2} |VC_i - VP_i|}{2}. \quad (1)$$

**Definition 3.** *Given the sensitivity level denoted by α and distance function $d_N$, the importance of information (IoF) for two normalized medical reports containing N examination criteria each is defined as:*

$$IoF = \begin{cases} 0, if\ d_N(2) \le 1-\alpha. \text{ NOT noticeable change} \xrightarrow{\text{yields}} \text{Discard it} \\ 1, if\ d_N(2) > 1-\alpha. \text{ Noticeable change} \xrightarrow{\text{yields}} \text{Visualize it} \end{cases} \quad (2)$$

The value of $d_N(2)$ in Eq. (2) is related to the indicator $\alpha$ that is the sensitivity level of the of change function. When *IoF* is zero, which means that $d_N(2)$ is less than or equal to (1-$\alpha$), the changes in $VC_i$ in comparison with $VP_i$ are small and the patient's status has not noticeablely changed. Thus, the current report may not need to be checked by the physician. If $d_N(2)$is larger than (1-$\alpha$), it means that the $VC_i$ has changed noticeably and the current report must be investigated by the physician. Since *IoF* relies on the value of $\alpha$, it is important to decide the value of $\alpha$ that is clinically meaningful. The range of $\alpha$ is from 0 to 1, where the correlation between the current and previous report becomes weaker as $\alpha$ becomes closer to 0. The correlation becomes stronger as $\alpha$ becomes closer to 1. A weak correlation, that is, when $\alpha$ is close to 0, causes the threshold for noticeable change to be high. Thus, the current report is discarded unless it shows significant difference from the previous one. For example, if $\alpha$ is 0.1, then the reports are considered similar whenever there is more than 10% similarity between the two reports. Also, if $\alpha$ is 0.8, then the reports considered to be similar whenever there is more than 80% similarity between them [9]. Therefore, the effectiveness of the method is largely dependent on the determination of $\alpha$.

However, the authors of [9] did not provide any further information on determining the value of $\alpha$ or quantitatively integrating the $N$ examination results contained in a medical report into single integrated quantitative value, which are the critial factors to utilize the chronological clustering method. Since the statuses of diseases are generally described by multiple examination criteria, it is essential to map the examination results of criteria in a medical report on scalar scale between 0 and 1.

To resolve the limitations of the above approach, we propose an *areal similarity degree* (ASD) based chronological clustering method to identify the temporal progress of chronic disease, especially metabolic syndrome. ASD is a similariy analysis model between two weighted radar charts comprising MS diagnostic criteria and examination results of risk factors, which was proposed by Jeong et al. [5-6]. The authors utilized ASD to establish risk quantification model for MS, which quantifies the disease status as a number between 0 and 1, and they determined that the risk model could effectively represent the disease status. The model resolves the well-known issues of MS diagnostic criteria as follows [10]:

- Equally weighted MS diagnostic criteria
- No explanation for disease status and changes of status
- No information about risk quantification by comparing examination results with diagnostic criteria

Therefore, the ASD-based risk model can be applied to identify the temporal change of chronic disease, particularly MS.

Let us consider the partial weighted radar chart shown in Fig. 1.

**Definition 4**. *According to [5], the ASD of a partial weighted radar chart comprising two polygons $A_{ij}$, i.e., $\Delta OA_iA_j$, and $B_{ij}$, i.e., $\Delta OB_iB_j$, is defined as*

$$ASD = S(A_{ij} \mid B_{ij}) = \frac{\text{Area of intersection of polygon } A_{ij} \text{ and } B_{ij}}{\text{Area of polygon } B_{ij}}. \quad (3)$$

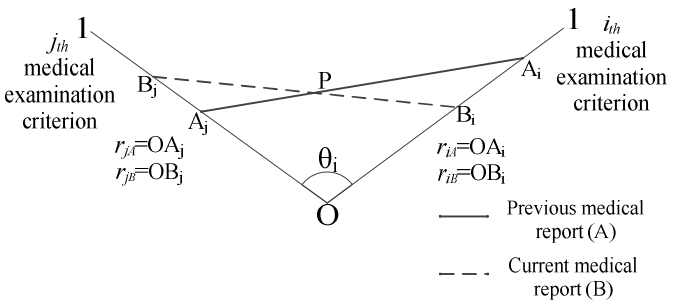


Fig. 1. A partial radar chart consisting of two polygons. The figure has been adopted from [5] and modified.

**Theorem 1.** *The distance between two chronologically ordered medical reports expressed by a partial weighted radar chart containing two examination criteria in each report can be calculated as:*

$$d_2(2) = 2 - ASD(A_{ij} \mid B_{ij}) - ASD(B_{ij} \mid A_{ij})$$

*, where $A_{ij}$ and $B_{ij}$ are partial weighted radar charts constructed using the normalized values of two medical examination criteria, namely i and j in each medical report and j=(i+1)* mod *2.*

*Proof.*

Let us assume that two medical reports $A_{ij}$ and $B_{ij}$ are results of the examination cirteria of the current and previous medical reports expressed by a partial weighted radar chart. Then according to Eq. (1), $d_2(2)$ becomes

$$d_2(2) = \frac{\sum_{i=1}^{2} \left| A_{ij} - B_{ij} \right|}{2} = \frac{\left| A_{1,2} - B_{1,2} \mid + \mid A_{2,1} - B_{2,1} \right|}{2} = \frac{\left| A_{ij} - B_{ij} \right| + \left| A_{ji} - B_{ji} \right|}{2}. \quad (4)$$

According to Fig. 1, $\left| A_{ij} - B_{ij} \right|$ equals to $\left| A_{ji} - B_{ji} \right|$. Thus, $d_2(2)$ becomes $\left| A_{ij} - B_{ij} \right|$. Since $\left| A_{ij} - B_{ij} \right|$ indicates the normalized difference between $A_{ij}$ and $B_{ij}$, and $A_{ij}$ and $B_{ij}$ can be mapped into a scalar value by means of their areas. Therefore, it becomes

$$\left| A_{ij} - B_{ij} \right| = \frac{\text{Area of } A_{ij} \text{ - Intersection of } A_{ij} \cap B_{ij}}{\text{Area of } A_{ij}} + \frac{\text{Area of } B_{ij} \text{ - Intersection of } A_{ij} \cap B_{ij}}{\text{Area of } B_{ij}} \quad (5)$$

According to Definition 4, since the ASD of a partial radar chart is defined as the ratio of the intersection of $A_{ij}$ and $B_{ij}$ over $B_{ij}$, Eq. (5) becomes

$$\left| A_{ij} - B_{ij} \right| = \{1 - S(A_{ij} \left| B_{ij} \right.)\} + \{1 - S(B_{ij} \left| A_{ij} \right.)\} = 2 - S(A_{ij} \left| B_{ij} \right.) - S(B_{ij} \left| A_{ij} \right.)$$

Therefore, we can obtain

$$d_2(2) = 2 - ASD(A_{ij} \mid B_{ij}) - ASD(B_{ij} \mid A_{ij}). \quad (6)$$

(*q.e.d*)

The conventional MS diagnosis criteria only determine whether a subject is diagnosed as having MS. However, chronic diseases such as MS that are characterized by temporal progress and periodic examinations need a method to easily determine whether the subject's disease status has changed since the previous examinations. The distance in Theorem 1 makes it easy to test whether the examined subject's disease status has changed over the patient specific sensitivity level described in Definitions 2 and 3. Further, the distance function can be calculated using the ASD of the subject's examination results and the thresholds of MS diagnosis criteria; thus, the potential risk of having MS can be easily quantified [5]. Therefore, Theorem 1 can be utilized as a preliminary diagnosis method in clinical decision support system for chronic disease.

**Corollary 1.** *Let $d_N(2)$ be the distance of two chronologically ordered medical reports containing N examination criteria in each. According to Theorem 1, $d_N(2)$ can be calculated as:*

$$d_N(2) = 2 - ASD(R_A \mid R_B) - ASD(R_B \mid R_A) \quad (7)$$

*, where $R_A$ and $R_B$ are weighted radar charts constructed using N disjoint partial weighted radar charts, $R_A = \{A_{12}, A_{23}, \ldots, A_{n1}\}$ and $R_B = \{B_{12}, B_{23}, \ldots, B_{n1}\}$. Each weighted radar chart represents a medical report with N medical examination criteria.*

Since a weighted radar chart is a sum of disjoint partial weighted radar charts and Theorem 1 provides the calculation of distance function for two partial weighted radar charts, Corollary 1 is obvious.

According to Eq. (7), we can calculate the distance function $d_5(2)$ for two weighted radar charts containing 5 examination criteria shown in Fig. 2 as 0.31.

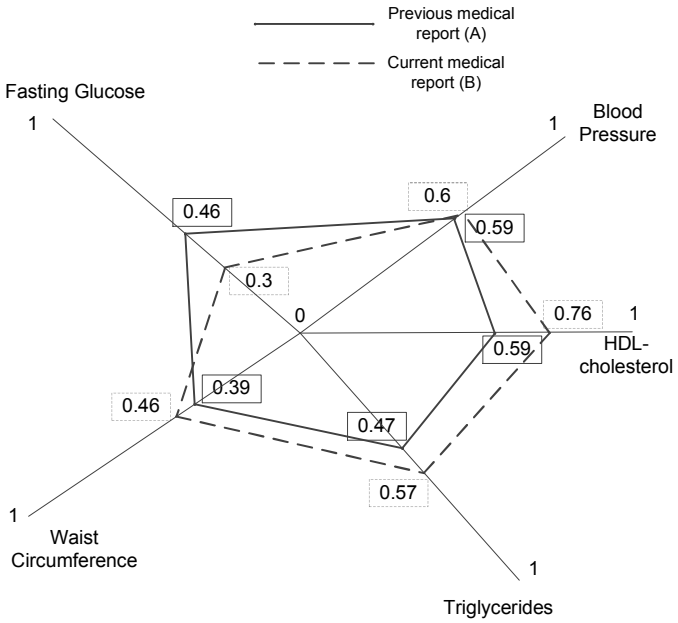


Fig. 2. Two weighted radar charts constructed from two data sets A and B. The figure has been adopted from [5] and modified.

## III. Evaluation and Discussion

This section presents a method to determine a patient's temporal disease progress based on chronological distance described in Corollary 1. Since it had been investigated that ASD can be used to quantifying and identifying risk of MS [5], we can assume that the variation of ASD over time, which is defined as a chronological distance in this paper, indicates the change of the risk status of MS. Therefore, we propose a method for identifying the temporal disease status change by analyzing the chronological distances.

**Hypothesis 1.** *If ASD quantifies the risk of having a chronic disease, particularly metabolic syndrome, we can claim that the chronological distance, i.e., $d_N(2)$, describing the difference between ASD value and linear regression value of ASD values over time can be used to identify the disease status change of a patient.*

To verify the hypothesis, we perform linear regression analysis of ASD values over time. Let $y(T)$ be a linear regression of ASD values from $t_0$ to $t_i$. Then, $y(T)$ can be expressed as follows:

$$y(T) = aT + b \quad (9)$$

Let $y(t_i)$ and $y(t_{i+1})$ be regression result at time $t_i$ and $t_{i+1}$ respectively. Also, let $y'(t_{i+1})$ be examined ASD value at time $t_{i+1}$. Then, we can define the chronological distance variance as follows:

**Definition 5**. *Chronological distance variance over $t_i$ and $t_{i+1}$ is defined as* $|y(t_i) - y'(t_{i+1})|$ . (10)

Let $\delta$ be the acceptable range of variation, then we claim that the chronological distance variance is within an control range if $|y(t_i) - y'(t_{i+1})| < y(t_{i+1}) * \delta$ holds. (11)

In other words, if the examined chronological distance at time $t_{i+1}$ is within $\delta$ % range from the regression value at time $t_{i+1}$, the disease status change at from $t_i$ to $t_{i+1}$ is within the control range. If the status change exceeds the control range, the patient will be instructed to be diagnosed in detail.

By considering Eq. (9) and Eq. (10), we can obtain the following observations.

In Eq. (9), the slope $a$ implies the temporal trend of subject's disease status. The values of a is defined as follows:

- $a \approx 0$ implies that the subject's disease status does not changed since last medical examination.
- $a > 0$ implies that the subject's examination results for criteria is in increasing trend.
- $a < 0$ implies that the subject's examination results for criteria is in decreasing trend.

In Eq. (10), the chronological distance variance quantifies the integrated change of examination results for disease criteria.

- $|y(t_i) - y'(t_{i+1})| < y(t_{i+1}) * \delta$ implies the subject's disease change is within a control range.
- $|y(t_i) - y'(t_{i+1})| \geq y(t_{i+1}) * \delta$ implies the subject's disease change is outside a control range.

To verify the proposed Hypothesis 1, we performed an evaluation with a sample patient's medical examination results. Table I shows the characteristics of the study subject over time.

TABLE I
A Sample Patient's Medical Examination Results for Metabolic Syndrome Risk Factors

| Exam. Time | T0 | T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 |
|---|---|---|---|---|---|---|---|---|---|
| Age, years | 35 | 35 | 35 | 35 | 36 | 36 | 36 | 36 | 37 |
| BMI ($kg/m^2$) | 23.6 | 23.5 | 22.4 | 22.6 | 21.7 | 22.6 | 21.7 | 21.4 | 21.9 |
| Fasting glucose (mg/dl) | 92 | 91 | 96 | 89 | 89 | 87 | 94 | 85 | 91 |
| Waist circumference(cm) | 80 | 77 | 82 | 83 | 78 | 77 | 79 | 82 | 77 |
| HDL-cholesterol (mg/dl) | 58 | 49 | 61 | 47 | 58 | 42 | 61 | 63 | 57 |
| Triglycerides(mg/dl) | 120 | 97 | 84 | 91 | 70 | 87 | 83 | 78 | 94 |
| Systolic BP (mm Hg) | 116 | 120 | 110 | 110 | 110 | 100 | 120 | 110 | 114 |
| Diastolic BP (mm Hg) | 64 | 80 | 70 | 70 | 70 | 60 | 60 | 70 | 75 |

BMI, body mass index; BP, blood pressure

Then, we obtained the linear regression result as shown in Fig. 3. The regression analysis resulted in the following statistics. The *R-squared* value is 0.3654 and the analysis showed 95% confidence level with 0.0598 error. Also, the top 5% and the bottom 5% percentile showed 0.6350 and 0.4296, respectively.

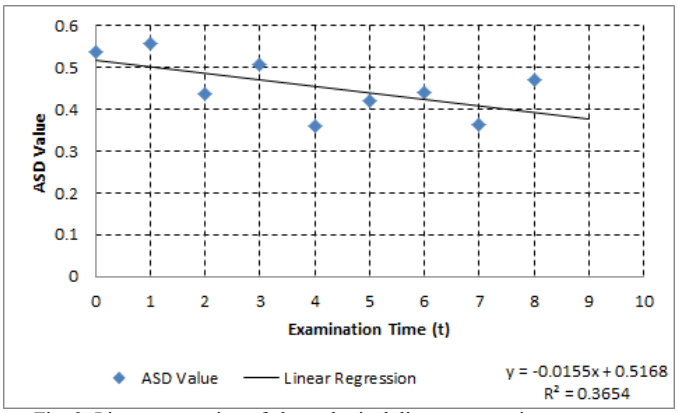


Fig. 3. Linear regression of chronological distance over time.

The calculated chronological distances using Table I and Eq. (7) are shown in Table II.

TABLE II
Calculation of Chronological Distance of a Sample Patient

| Time | $t_0$-$t_1$ | $t_1$-$t_2$ | $t_2$-$t_3$ | $t_3$-$t_4$ | $t_4$-$t_5$ | $t_5$-$t_6$ | $t_6$-$t_7$ | $t_7$-$t_8$ |
|---|---|---|---|---|---|---|---|---|
| Chronological Distance | 0.042 | 0.064 | 0.021 | 0.110 | 0.034 | 0.001 | 0.062 | 0.063 |

Let the control range at time $t_{i+1}$ be $A(t_{i+1})$. Also, Let us assume that the control range parameter δ be 0.25 (or 25%), then we can get the control range of chronological distance at time $t_{i+1}$, as follows.

$$0.2946 < A(t_{i+1}) < 0.4910 \quad (12)$$

Since ASD value at time $t_{i+1}$ is 0.4708 and within the control range $A(t_{i+1})$, we can decide that the sample patient's disease status is within the control range.

Therefore, when the examined chronological distance at time $t_{i+1}$ is outside the range shown in Eq. (12), the subject is likely to have major change at disease status and needs to consult physician. The analysis in this section has been performed with very limited sample examination results. Thus, more extended clinical evaluation with a large number of patient examination data is necessary in order to establish clinically meaningful value of $\delta$. Also, longitudinal study using cohort is need to extensively evaluate the clinical effectiveness of the proposed temporal progress identification method.

## IV. CONCLUSION

In this paper, we proposed a method for identifying the temporal progress of metabolic syndrome using chronological clustering. By using the proposed method, we can design new application services, such as analyzing long-term trends of patient's disease status, a knowledge-based decision support tool for metabolic syndrome. The clinical effectiveness of the proposed model was evaluated using a sample patient data and the evaluation results showed that the proposed model can be used to analyze the temporal progress of chronic disease, especially the risk of metabolic syndrome. Using the proposed model, we can effectively manage the patients having metabolic syndrome using the control range of patient status.

## ACKNOWLEDGMENT

This research was supported by the ICT Standardization program of MKE (The Ministry of Knowledge Economy).

## REFERENCES

[1] S. Jeong, C.-H. Youn, et a., "An Integrated Healthcare System for Personalized Chronic Disease Care in Home-Hospital Environments," *IEEE Transactions on Information Technology in Biomedicine*, vol.16, no.4, pp. 572-585, July 2012.

[2] A. Moxey, et al. "Computerized clinical decision support for prescribing: provision does not guarantee uptake," *Journal of the American Medical Informatics Association*, vol.17, no. 1, pp. 25-33, 2010.

[3] D. Kim, Y. Seo, and C.-H. Youn, "Prediction of Atrial Fibrillation Onsets by Multiple Formulas based on Long Term Variation of Heart Rate Variability, " in *Proc. Healthcom*, 2008, pp.5482-5485.

[4] Jason A Lyman, Wendy F Cohn, Meryl Bloomrosen, Don E Detmer, "Clinical decision support: progress and opportunities," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 487-492, 2010.

[5] S. Jeong, Y. M. Jo, S.-O. Shim, Y.-J. Choi, C.-H. Youn, "A Novel Model for Metabolic Syndrome Risk Quantification Based on Areal Similarity Degree," *IEEE Transactions on Biomedical Engineering*, 2013, doi:10.1109/TBME.2013.2286197.

[6] S. Jeong, C.-H. Youn, "A Personalized Disease Identification Scheme using Analytic Hierarchy Process for u-Healthcare System," in *Proc. International Exposition Yeosu Korea Conference on Information Technology*, 2012, pp.110-116.

[7] Michael A. Bell and Pierre Legendre, Multicharacter Chronological Clustering in a Sequence of Fossil Sticklebacks, *Systematic Zoology*, vol. 36, no. 1, pp. 52-61, Mar., 1987.

[8] P. Legendre, S. Dallot and L. Legendre, Succession of Species within a Community: Chronological Clustering, with Applications toMarine and Freshwater Zooplankton, *The American Naturalist*, vol. 125, no. 2, pp. 257-288, Feb. 1985.

[9] Fawaz AL Hazemi, Chan Hyun Youn, Khalid A. Al-Rubeaan, "Grid-based Interactive Diabetes System," in *Proc. First IEEE International Conference on Healthcare Informatics, Imaging and Systems Biology*, 2011, pp. 258-263.

[10] R. Kahn, J. Buse, E. Ferrannini, M. Stern, "The Metabolic Syndrome: Time for a Critical Appraisal," *Diabetes Care*, vol. 28, no. 9, pp. 2289-2304, Sept. 2005.