ABSTRACT

Title of Dissertation: NOVEL STATISTICAL METHODS IN HEALTHCARE ANALYTICS FOR PEOPLE WITH DIABETES AND PREDIABETES

Shiping Liu
Doctor of Philosophy, 2021

Dissertation Directed by: Professor Guodong (Gordon) Gao
Robert H. Smith School of Business
Professor Paul Smith
Statistics Program, Department of Mathematics

A great amount of statistical tools and methods have been applied in health care analytics to assist decision making and improve the quality of diabetes related health care services. However, limitations of existing methods, new types of data, and specific demands in different areas are challenging current statistical tools. These challenges further encourage developing new statistical methods or extending existing methods to better fit different demands and improve performance of models and methods in practice. In this dissertation, we developed, applied, and extended many innovative statistical models and methods to address practical issues in health care of diabetes related population. Firstly, we developed a novel automated event detection method for univariate time series, Continuous Glucose Monitoring (CGM) data, from diabetic patients. Secondly, we invented a low-dimensional framework to classify and track longitudinal glucose status of CGM users based on within-subject analysis and unsupervised variable selection methods. Thirdly, we investigated the

influence of daily activities on glucose series by applying a nonparamentric multivariate two sample test with independence assumption relaxed. Moreover, besides focusing on diabetic population, we also developed predictive models to access the risk of diabetes for population with prediabetes in later two chapters. Two types of response variables, binary indicator and HbA1c values, were used to aid different demands in practical healthcare services.

NOVEL STATISTICAL METHODS IN HEALTHCARE ANALYTICS FOR PEOPLE WITH DIABETES AND PREDIABETES

by

Shiping Liu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2021

Advisory Committee:
Professor Paul Smith, Co-Chair
Professor Guodong (Gordon) Gao, Co-Chair
Professor Benjamin Kedem
Professor Margret Vilborg Bjarnadottir
Professor Ping Wang

Acknowledgments

Throughout the writing of this dissertation, I have received a great deal of support and help. I owe my gratitude to all the people who have made this possible.

I would like to express my deepest gratitude to my advisor, Professor Gordon Gao, for giving me invaluable opportunities to work on challenging and interesting projects over the past three years. He has always helped and supported me whenever I've encountered difficulties. It is an great honor to work with and learn from Professor Gao.

I would also like to thank my co-advisor, Professor Paul J. Smith. His extraordinary guidance and advice pushed me to think harder and brought this dissertation to a higher level. I would like to extend my sincere thanks to Professor Margret Bjarnadottir, Professor Benjamin Kedem, and Professor Ping Wang for their willingness to serve on my thesis committee and for their invaluable time reviewing the dissertation.

I would like to acknowledge the help and effort of my team members at University of Maryland, Center for Health Information and Decision Systems, and my collaborators at Welldoc and Medstar. Michelle Dugas and Kenyon Crowley helped me start-off and provided valuable contribution in many projects. Dr. Mansur Shomali shared his professional thoughts and suggestions as diabetes expert, and

made the studies more practical to real life applications. Di Hu, Abhimanyu Kumbara, Weibo Chen, Lijun Yang, and Junjie Luo provided unwavering assistance and shared great ideas in different projects. Without their help, we would not have finished these projects and studies.

Lastly, I would like to thank my parents and friends for their constant and unconditional love and support. They are always there for me, and always supportive of my new ideas and adventures. Their smiles always remind me of what is important in my life and give me strength and courage to try and forge ahead.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AIC	Akaike information criterion
ANCOVA	Analysis of covariance
ANN	Artificial neural network
ANOVA	Aanalysis of variance
APP	Application
ARMA	Auto regressive moving average
AUC	Area under the ROC curve
BC_a	bias-corrected and accelerated
BGM	Blood glucose monitoring
BIC	Bayesian informaiton criterion
BMI	Body mass index
BP	Blood pressure
CC	Cross-correlation
CDC	Centers for Disease Control and Prevention
CDM	Cross-distance matrix
CGM	Continuous glucose monitoring
CPU	Central processing unit
CV	Coefficient of variation
DBA	DTW Barycenter Averaging
DTW	Dynamic time warping
EHR	Electronic health record
EM	Expectation-Maximization
FCM	Fuzzy c–means
FF	Forward filtering
FFBS	Forward filtering backward smoothing
GL	Glucose
GLMM	Generalized mixed-effect model
GMI	Glucose management indicator
HbA1c	Hemoglobin A1c
HDL	High density lipoproteins
IDF	International Diabetes Federation
IQR	Interquartile range
KNN	K nearest neighbor
LASSO	Least absolute shrinkage and selection operator

LDL	Low density lipoproteins
LR	Logistic regression
MAE	Mean absolute error
Mgl	Mean glucose
MLE	Maximum likelihood estimator
NCC	Cross-correlation with normalization
NLP	Nature language processing
NN	Nearest neighbor
OGTT	Oral glucose tolerance test
PC	Principal component
PCA	Principal component analysis
RF	Random forest
ROC	Receiver operator curve
SD-gl	Standard deviation of glucose
SGPT	Serum glutamic pyruvic transaminase
SSD	Sum of the squared distance
SVM	Support vector machine
T2D	Type 2 diabetes
TAR	Time above range
TBR	Time below range
TIR	Time in range

Chapter 1: Introduction

1.1 Background

Diabetes mellitus is one of the most common chronic diseases that describes metabolic disorders characterised by high blood glucose. Hyperglycemia can cause serious health problems and increase the risk of developing cardiovascular disease, retinopathy, neuropathy, and kidney disease [CDC, 2020a]. The global prevalence of diabetes and impaired glucose tolerance in adults has been increasing over recent decades [Cho et al., 2018]. The International Diabetes Federation (IDF) estimated the global prevalence to be 151 million in 2000, 246 million in 2006, 366 million in 2011, and 451 million in 2017 [Cho et al., 2018]. In US, according to the Centers for Disease Control and Prevention (CDC), there are 34.2 million people living with diabetes which is 10.5% of the US population. Besides the high prevalence rate, the cost of managing T2D is also increasing. Total costs of the treatment of diagnosed diabetes have risen to $327 billion in 2017 from $245 billion in 2012, which represents a 26% increase over a five-year period [ADA, 2020]. In efforts to prevent and manage diabetes, more than 1,500 CDC-recognized organizations offer different programs in T2D prevention, and costs of such programs are now covered by Medicare as well as approximately 40 commercial health plans [CDC, 2020a].

Various statistical tools and methods have been widely applied in diabetes related studies to aid the management and prevention of diabetes. Fundamental statistical tests are often applied for hypothesis testing, such as one-sample Z-/t-test to identify significant predictors, two-sample t-test for two-group comparison and analysis of variance (ANOVA) for multiple group comparisons, Hosmer-Lemeshow test for goodness of fit [Adane et al., 2020, Chutani and Pande, 2017, Yokota et al., 2017], etc. Analysis on correlation, such as Pearson's and Spearman correlation, is popular to investigate relationship between variables [Adane et al., 2020, Wilson et al., 2007, Yokota et al., 2017]. Moreover, linear regression and generalized linear models, such as logistic regression for binary outcome variables, are widely used in future value or risk predicting studies [Wilson et al., 2007, Schmidt et al., 2005, Yokota et al., 2017]. These parametric regression models are also commonly applied to investigate the associations between outcome variables and risk factors [Yeh et al., 2010, Schulze et al., 2007, Hodge et al., 2006, Christine et al., 2015] in multivariate analysis. For more complex settings, such as repeated measures or time variant analysis, mixture models have been introduced to related topics [Luo et al., 2018, Ngufor et al., 2019, Christine et al., 2015]. In addition to parametric methods, non-parametric algorithms are developed and applied in the diabetes context and performed well in many studies [Nguyen et al., 2019, López et al., 2018, Bradley, 1997, Karpati et al., 2018, Luo et al., 2018]. Tree-based methods, such as decision tree, random forest, boosted trees, are widely applied in both classification and regression tasks. Some unsupervised methods, such as clustering and k-Nearest neighbor methods, are also applied to this area in many data mining problems and incorporated with other statistical methods to build more complicated approaches. With the improvement of computation power, an increasing number of studies applied deep learning algorithms and introduced vector embedding techniques to many applications [Swapna et al., 2018, Naz and Ahuja, 2020, Ye et al., 2020].

1.2 Chapter Plan

In this dissertation, two scenarios are discussed and various methods are applied and developed to address different issues in both scenarios. The reminder of this dissertation is organized as follows.

The first scenario is mainly focused on data mining tasks on Continuous Glucose Monitoring (CGM) time series data. CGM is widely used to adjust insulin therapy for patients with Type 1 Diabetes. However, in recent years, more CGM therapy started to help monitor lifestyle intervention for patients with type 2 diabetes. CGM device records an average glucose value every five minutes for each user, which creates tons of data need to be further analyzed in order to support diabetes experts or users for glycemic control and therapy adjustments. In chapter 2, we developed an automated algorithm for CGM event detection to identify and classify significant CGM segements from long streaming CGM data. In chapter 3, we invented a low-dimensional framework to classify and track users' glucose status based on within-subject correlation study and unsupervised variable selection. In chapter 4, we investigated the influence of real life activities on the glucose time series by extending the use of multivariate two-sample test.

The second scenario is regarding to prevention of diabetes among prediabetic population. The management of prediabetes is a crucial step in curbing the growth of diabetes. There exists substantial variation in the risk of progression to diabetes from prediabetes. Thus, addressing the risk of developing diabetes among patients with prediabetes would aid decision making in health care of prediabetes and optimize the distribution of medical resources among patients with different level of risks. In chapter 5, we developed a predictive model for binary outcome, whether patients develop diabetes in 3 years after their diagnose of prediabetes or not. Besides the binary outcome, in chapter 6, we developed a dynamic linear model to predict the future hemoglobin A1c (HbA1c) values, which is a major measure of diagnosing diabetes.

Chapter 2: Analysis of Continuous Glucose Monitoring Data: An Automated Event Detection Method on Time Series Data

2.1 Introduction

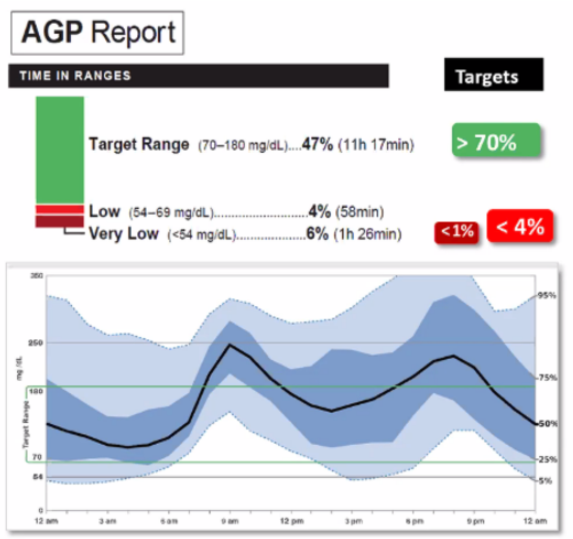
Regular glucose monitoring is the most important thing to manage diabetes. It can help patients and their health care team to identify influence of different activities and make decisions about a best diabetes care plan. These decisions can help delay or prevent diabetes complications. In addition to the traditional measure method, finger stick-based blood glucose monitoring (BGM), continuous glucose monitoring (CGM) has been invented and widely used to help in diabetes management.

CGM measures interstitial fluid every 10 seconds and records an average glucose value every five minutes. CGM is widely used to adjust insulin therapy for patients with Type 1 Diabetes, and in recent years, more CGM therapy started to help monitor lifestyle intervention for patients with Type 2 diabetes. Plenty of applications (APP) on smart electronic devices, for example smart phones, can sync CGM records and combine with other real life activities to generate individualized reports and provide advises accordingly.

CGM provides more accurate records of daily glucose fluctuations, and helps to monitor the effects of foods, physical activities, and medications on glucose. Monitoring, recording, and presenting glucose continuously can help users improve lifestyle management and optimize combinations of different types of interventions for diabetic patients. Moreover, continuous available glucose monitoring can discover and avoid unexpected hypoglycemia, which may be fatal for patients with diabetes [Cryer et al., 2003]. Based on the benefits of CGM, statistical analysis of CGM data could summarize a patient's status; exhibit longitudinal changes efficiently and visually; and recognize potential user patterns to further assist physicians on decision making and patient training.

Current analysis on CGM data are limited and mostly restricted to basic descriptive summaries for CGM users. Figure 2.1 shows an example of current daily CGM report. A report over a longer period, for example three month, is similar [Battelino et al., 2019]. We propose to design more complete and accurate reports for users and physicians to help design and adjust the patient's therapy and training.

Our collaborator Welldoc Inc., a digital health care company, provided 1) a large amount of CGM data from de-identified users; 2) self-reported activity entries, such as foods, physical exercises, medications, etc. from their APP users, including CGM users; 3) users' profiles, including demographics, medical information, etc. The real data analyses in Chapter 2–4 use different types or combinations of these



AGP Report

TIME IN RANGES

Targets

Target Range (70–180 mg/dL)....47% (11h 17min) > 70%

Low (54–69 mg/dL)..........................4% (58min)

Very Low (<54 mg/dL)..................6% (1h 26min) <1% < 4%

Figure 2.1: CGM daily report example

data to address a variety of problems. More details will be introduced in the each real data study.

2.2 Motivations for event detection

With the development of CGM devices, CGM has become an indispensable tool for more and more patients with diabetes. Compared to the development of the hardware, most CGM reports provide limited information to users and physicians. Commonly, the report only provides users' glucose information based on high-level summary statistics of different metrics. However, physicians are eager to know how their patient's glucose behaves at a more detailed level, for example, event level, besides a high-level overview. CGM devices, recording users' full glucose history, provide opportunities to capture and view important glucose segments, and further help their glycemic control, patient training, medication adjustment, etc.

Based on the demand, we define the important segments as "CGM events" (or events), which is a sub-sequence of a CGM series that is significantly distinguishable from stable and normal CGM sequences. An event normally includes glucose elevation, peak, and decline, but the shapes may be varying. Figure 2.2 shows three illustrative examples of CGM events in one-day CGM data. CGM events are easy to recognize manually when the entire glucose stream is short. However, it is necessary to develop automated detection algorithms to recognize and summarize CGM events from huge volumes of CGM streaming data from a sizeable group of patients. To extract information at event level, the objective of this chapter is to develop an automatic algorithm that can efficiently and accurately detect CGM events, and extract relevant information for each detected event.

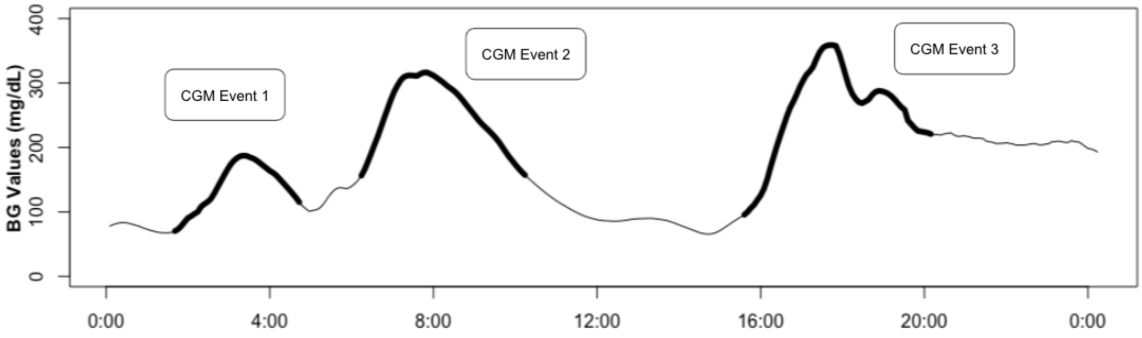


Figure 2.2: Illustrative examples of CGM events. Three CGM events are indicated by bold segments and text labels

2.3 Related topics

Regarding the method of detecting the defined CGM events, there are some similar topics in the literature. Peak detection, change point detection, and pattern matching are the most relevant topics in our scenario.

2.3.1 Peak detection

The peak detection problem is an important topic that is commonly discussed and studied in signal processing applications. Peaks often indicate significant events, for example, sudden price changes in the stock market and bursts in utilization of CPU [Palshikar et al., 2009]. For univariate time series with small size, peaks can be easily identified manually through visualizing the time series. However, when the data size and dimension increase, one needs algorithms to automatically detect peaks and to avoid subjectivity.

A data point in a time series is a local peak if it is 1) a local maximum within a time window; and 2) isolated, meaning that the number of data points with similar values in the same time window should be sufficiently small [Palshikar et al., 2009]. It is worth noting that a local peak is not necessarily a global peak, whose values are sufficiently large in the global context. Many methods have been developed to detect peaks in time series data, such as traditional window-threshold techniques [Pan and Tompkins, 1985, Jacobson, 2001], wavelet transform [Du et al., 2006], Hilbert transform [Benitez et al., 2001], artificial neural networks [Xue et al., 1992], filtering methods [Zhang and Lian, 2011], clustering methods [Mehta et al., 2010], etc.

2.3.2 Change point detection

Change points are sudden changes of observation magnitudes or variations in time series, which may potentially indicate transitions between hidden states. Change point detection is widely used for data mining and modeling in many areas, such as medical condition monitoring, climate change detection, speech and image analysis, etc [Aminikhanghahi and Cook, 2017].

Summarised in a survey study by Aminikhanghahi et al. [Aminikhanghahi and Cook, 2017], related studies includes segmentation, edge detection, and anomaly detection. Both supervised and unsupervised methods are employed in change point detection. Supervised methods are essentially classification models, such as decision tree, support vector machine, nearest neighbor, hidden Markov model, etc [Reddy et al., 2010]. Unsupervised methods includes likelihood ratio methods [Kawahara and Sugiyama, 2012], probabilistic methods [Adams and MacKay, 2007], kernel-based methods [Harchaoui et al., 2009], clustering methods [Keogh et al., 2001], etc.

2.3.3 Pattern match

A pattern in time series is a set of sequential observation describing a meaningful shape or tendency during a period of time. Patterns could imply important phenomena in the monitored objects and widely discussed in finance, medical, and other areas. In general, pattern matching problem can be divided into two scenarios: 1) full sequence matching, in which matching is based on the full length of the series; and 2) pattern detection on streaming time series, in which matched subsequences are searched from a long time series [Chen et al., 2007].

The latter scenario is more relevant in our context and many methods have been proposed for the second type of pattern matching. Some popular methods are based on neural network [Zapranis and Tsinaslanidis, 2010], geometrical similarity measure [Zhou et al., 2006], regression [Lo et al., 2000], clustering [Barnaghi et al., 2012], principal component analysis [Rao and Principe, 2002], etc. Searching for matched segments normally involves measuring the distance between segments with different length in most cases. Many distance measures have been proposed, such as norm-based distances, shape-based (dis)similarity measures, etc. Another challenge in pattern matching is the segmentation of a long time series. Using different segmentation methods on time series can have a profound effect on the pattern matching results [Wan et al., 2016]. In most cases, the length of potential matched sub-sequences are unknown and varying. Most existing studies applied fixed window sliding or disjoint windows to query sub-sequences [Chen et al., 2007].

2.4 Limitations in existing methods

Reviewing these relevant topics, we noticed the following challenges as an event detection problem:

1. Consecutive points detection

Unlike the peak or change point detection, which focus more on single point detection, our task on event detection is not limited to some important points, beginning or peak of an event, but cares more on the entire subsequence.

2. Challenges in regular pattern matching

The demand in our task, identifying en entire subsequence, makes our study more like a pattern matching problem. However, there are some main challenges in pattern matching problems. 1) The searching window size is difficult to decide in segmentation of a long time series. The boundaries or length of the potential matching subsequences are normlly unknown and varying. Therefore, the size of the detection window critically affects the detection performance. 2) The shifting and scaling issues may also influence the matching performance since it requires more careful measure of distances or (dis)similarities. 3) Pattern matching techniques are computationally expensive and inefficient. Most shape-based distances or (dis)similarities measures have high computational cost. However, in practice, data sizes are normally large in both length of single target streaming series, and number of subjects. In addition, most pattern matching methods use sliding searching windows, and this causes a large amount of redundant computations.

2.5 Methods

To overcome the mentioned difficulties in pattern matching problem, one possible solution is to characterize the time series by a simpler alternative which reserves the main characteristics of the original time series and ignores minor details. Back to 1980's, Kedem and Slud [Kedem and Slud, 1982,Kedem and Slud, 1981] proposed using the binary differential representation to characterize and discriminate time series. They also investigated the asymptotic properties of the representation series and developed novel statistics and methods for the Goodness-of-fit tests based on higher order crossing. In study [Kedem and Slud, 1982], the authors defined the binary differential representation series as follows. Let $Z = \{Z_t\}_{t=-\infty}^{\infty}$ be a stationary discrete-time process. Denote ∇ as the backward-difference operator, which yields $(\nabla Z) = Z_t - Z_{t-1}$ and $(\nabla^2 Z) = (\nabla(\nabla Z))_t = Z_t - 2Z_{t-1} + Z_{t-2}$. Define the clipping operator U as

$$(UZ)_t = \begin{cases} 1 & \text{if} \quad Z_t \geq 0 \\ 0 & \text{if} \quad Z_t < 0 \end{cases}$$

Then, for $j \geq 1$, the binary differential representation of Z is defined by

$$X_t^{(j)} \equiv (U\nabla^{j-1}Z)_t$$

In this method, series $\{X_t\}$ is an encoded binary representation of the original series $\{Z_t\}$. The authors also studied the asymptotic properties of the binary representatives under certain conditions and showed that binary representatives based on low degree differences are useful for discrimination.

Enlightened by the idea of higher order crossing method, we proposed an approach for event detection based on the ideas of encoding the raw series and applying the exact pattern matching on the encoded sign change series instead of the raw data series. It is worth noting that our method can be potentially extended to other event detection problems in univariate time series mining.

2.5.1 Definitions and Main steps

Definitions

CGM event (or event): a subsequence of glucose readings that is distinguishable from the stable glucose status. It often consists of an increasing segment, a stable segment (optianal), and a decreasing segment but the shapes are varying.

Slope series (d(GL)): series of first derivatives of the smoothed CGM series

Sign series (of slope series) : indicating the sign of each data point in slope series by three categories:

$$\text{sign} = \begin{cases} + & \text{if} \quad d(GL) > 0 \\ 0 & \text{if} \quad d(GL) = 0 \\ - & \text{if} \quad d(GL) < 0 \end{cases}$$

Sign change: indicating the sign change between a data point and its previous data point in slope series by two categories:

$$\text{sign change} = \begin{cases} +1 \quad \text{if} \quad d(GL) \text{ change from } - \text{ to } 0 \text{ or from } 0 \text{ to } + \\ -1 \quad \text{if} \quad d(GL) \text{ change from } 0 \text{ to } - \text{ or from } + \text{ to } 0 \end{cases}$$

Main Steps

To complete the task and overcome the difficulties discussed above, the main idea of our method is to match the patterns in the sign change series, instead of the raw glucose series. By doing so, we characterized the raw series (including the patterns and the long target series) by its sign change series, and then resolve the above mentioned challenges in typical pattern matching problem: 1) lengths of potential matching sequences are fixed and known; 2) shifting and scaling issues in the raw data do not affect the sign change pattern; 3) redundant computation and distance measuring are avoided during the exact matching.

The main steps are as follows:

1. Data preparation: we manually labeled an indicator variable for each CGM reading which 1 indicates that an observation belongs to a CGM event, and 0 otherwise. The labeled data are used in method developing, parameter optimizing, and performance testing.

2. Pattern Recognition: we identify and encode the typical event patterns in our sample as reference patterns.

3. Raw data encoding: we encode both the reference patterns and the long streaming series as their sign change series.

4. Pattern searching: we search for the reference patterns (in term of sign change series) from the sign change series of the long series by exact matching.

5. Event attributes extraction: extracting attributes, event starting and ending time, etc. from the detected events (subsequences).

Methods and techniques applied at each step are introduced and described in the following subsections.

2.5.2 Pattern Recognition

The goal of pattern recognition is to identify and summarize the reference patterns in the sample data for the later step, pattern searching. Briefly, reference patterns can be recognized by analyzing the pre-labeled segments (events): grouping the similar segments and extracting the typical pattern from each group. Grouping the similar segments involves methods to measure distance between segments with different lengths, and then, properly cluster all segments into multiple subgroups. Extracting typical patterns requires defining and computing a pattern that represents each subgroup. Methods applied during the pattern recognition are discussed as follows.

2.5.2.1 Distance measures: Dissimilarity of Time Series

Many methods were proposed and developed to measure distance or (dis)similarity between time series for a long time. The ℓ_1 and ℓ_2 norms, or Manhattan and Euclidean distances respectively, are the most commonly used distance measures for both cross-sectional or time series data [Aggarwal et al., 2001]. However, the disadvantages of ℓ_p distance measures are also obvious in the context of measuring the (dis)similarities of time series data. They can be only defined for equal length time series, and are sensitive to noise, scale and time shifts.

Alternatively, many other distance measures for time-series have been proposed to overcome these limitations and other challenges associated with time series data. The shape-based time series clustering methods provided many insights and could be properly applied in our study. In the shape-based time-series clustering, some methods of measuring the distance are widely used, such as, Dynamic time warping distance (DTW), Global alignment kernel distance, and Shape-based distance.

DTW [Berndt and Clifford, 1994] is a dynamic programming algorithm that compares two series and finds the optimal warping path between them under certain constraints, such as monotonicity. It is widely used by the data mining community to overcome the limitations associated with the Euclidean distance [Ratanamahatana and Keogh, 2004].

Cuturi [Cuturi, 2011] proposed an algorithm, Global alignment kernel distance, to assess similarity between time series by using kernels. Denote an alignment between two series x and y as π, and defined the set of all possible alignments as $\mathcal{A}(n,m)$, which is constrained by the lengths of x and y. $|\pi|$ is the length of π . Then the distance is defined as

$$d(x,y) = \sum_{\pi \in \mathcal{A}(n,m)} \prod_{i=1}^{|\pi|} \kappa(x_{\pi_1(i)}, y_{\pi_1(i)})$$

where κ is a local similarity function defined as

$$\kappa(x,y) = e^{-\phi_\sigma(x,y)},\ \phi_\sigma(x,y) = \frac{1}{2\sigma^2}\|x-y\|^2 + \log(2 - e^{-\frac{\|x-y\|^2}{2\sigma^2}})$$

and σ is the bandwidth. However, limitations of the method are diagonal dominance and a complexity in computation.

The shape-based distance was proposed as part of the k-Shape clustering algorithm [Paparrizos and Gravano, 2015]. It is based on the cross-correlation with normalization (NCC) sequence between two series. The NCC sequence is obtained by convolving the two series, so different alignments can be considered. In particular, the cross-correlation (CC) between $x = (x_1, ..., x_m)$ and $y = (y_1, ..., y_m)$ is defined as

$$CC_\omega(x,y) = R_{\omega-m}(x,y), \omega = 1, 2, ..., 2m-1$$

where

$$R_k(x,y) = \begin{cases} \sum_{l=1}^{m-k} x_{l+k} \cdot y_l & \text{if} \quad k \leq 0 \\ R_{-k}(x,y) & \text{if} \quad k < 0 \end{cases}$$

In calculation of CC_ω, ω indicates the shift of x. Different normalizations for CC_ω are discussed in [Paparrizos and Gravano, 2015]. One popular method is the coefficient normalization, NCC_c, which is defined as follows:

$$NCC_c(x,y) = \frac{CC_\omega(x,y)}{\|x\|_2\|y\|_2} = \frac{CC_\omega(x,y)}{\sqrt{R_0(x,x)\cdot R_0(y,y)}}$$

where $\|\cdot\|_2$ is the ℓ_2 norm of the series. Then, the shape-based distance can be calculated with

$$d(x,y) = 1 - \max_\omega(NCC_c(x,y))$$

Its range lies between 0 and 2, with 0 indicating perfect similarity. The calculation of CC_ω is sensitive to scale, which may need to adjust the scale issue if using other methods for normalization of CC_ω.

Among different measures, DTW distance is widely applied as dissimilarity measure for time series and performed well in many studies [Aghabozorgi et al., 2015]. It allows to measure dissimilarity between two series with different length. One main drawback of DTW algorithm is that the method is computationally expensive in time and memory. However, DTW is proper to apply in our analysis during the pattern recognition step (instead of directly using in pattern matching), given that length of each event series in our analysis is relatively small. Note that other distance measures can also be applied if they are able to measure the distance based on shape and for series with different length.

2.5.2.2 Dynamic time warping (DTW)

Suppose two time series are denoted by $X = (x_1, ...x_n)$ and $Y = (y_1, ...y_m)$, and the aim is to measure the dissimilarity between X and Y. A cross-distance matrix (CDM) with $n \times m$ dimensions is defined between X and Y for each pair of x_i and y_j by a non-negative, local distance function, $d(x_i, y_j)$. The most common choice for $d(x_i, y_j)$ is ℓ_p norm. In particular,

$$\mathrm{CDM}(i, j) = d(x_i, y_j) = (\sum_{\nu} |x_i - y_j|^p)^{1/p} \tag{2.1}$$

Note that the calculation of the CDM matrix requires X and Y have the same dimension, ν, in the multivariate case. For univariate time series, the CDM will be identical regardless of the value of p. The core of DTW is to define the *warping curve* $\phi = \{\phi(k), k = 1, ..., K\}$:

$$\begin{aligned} \phi(k) &= (\phi_x(k), \phi_y(k)) \quad \text{where} \\ \phi_x(k) &\in \{1, ..., n\} \\ \phi_y(k) &\in \{1, ..., m\} \end{aligned}$$

For a given warping curve ϕ, we compute the average accumulated distortion between X and Y as

$$d_\phi(X, Y) = \sum_{k=1}^{K} \frac{d(\phi_x(k), \phi_y(k)) m_\phi(k)}{M_\phi} \tag{2.2}$$

where $m_\phi(k)$ is a per-step weighting coefficient and M_ϕ is the corresponding normalization constant, which ensures that the accumulated distortions are comparable along different paths. To ensure reasonable warps, constraints are usually imposed on ϕ. For example, monotonicity is imposed to preserve their time ordering and avoid meaningless loops:

$$\begin{aligned}\phi_x(k+1) &\geq \phi_x(k)\\ \phi_y(k+1) &\geq \phi_y(k)\end{aligned}$$

The DTW algrithm optimizes the warping curve that minimizes d_ϕ, and then defines the DTW distance as

$$DTW_d(X,Y) = \min_{\phi} d_\phi(X,Y) \tag{2.3}$$

In our case, we need to compute the global alignment which

$$\phi_x(1) = \phi_y(1) = 1; \qquad \phi_x(K) = n; \quad \phi_y(K) = m.$$

In equation 2.2, the choice of $m_\phi(k)$ and its corresponding M_ϕ depends on the step-patterns, which is a local constraint that determines which directions are allowed when moving ahead in the CDM with the associated per-step weights. Many step-patterns have been discussed, and *symmetric*1 pattern in [Giorgino et al., 2009] is used in our analysis. The symmetric1 pattern allows vertical, horizontal, and diagonal step with equal weight, and is widely used in many studies [Rakthanmanon et al., 2012, Fu, 2011].

2.5.2.3 Time series clustering

Major types of time series clustering include shape-based, feature based, edit-based clustering, and structure-based clustering [Aghabozorgi et al., 2015, Hennig et al., 2015]. Shape-based clustering is based on distance that measures (dis)similarities of the overall shape of time series using the original or scaled data. Feature based clustering often is applied to obtain reduction in dimension and noise. It requires extracting features from the original time series first, and then measures the distance based on these features. Edit-based distance is based on the minimum number of operations that are required to transform one time series into the other time series. Structure-based distance compares higher level structures that obtained by modelling or compressing the time series.

In general, the widely used clustering methods can be divided into partitional clustering and hierarchical clustering [Hennig et al., 2015]. The partitional method divides data objects into multiple subsets or clusters, such that each data object exactly belongs to one cluster. K-means clustering method is one of the most known partitional clustering methods. One challenge problem in partitional clustering is that the number of clusters needs to be prespecified.

The hierarchical clustering builds a nested structure that organizes data objects as a hierarchical tree. Two major approaches for hierarchical clustering are agglomerative and divisive method. The agglomerative approach starts with all objects as individual clusters, and for each step, merge the closest pair of clusters until only one cluster left. In agglomerative hierarchical clustering analysis, linkage criterion determines the distance between two nodes (sets of observations). Some widely used linkage criteria have been discussed, such as single linkage, complete linkage, weighted/unweighted average linkage, and centroid linkage. In contrast, the Divisive approach starts with one cluster that includes all objects, and then, at each step, split a cluster until each cluster contains only one object. The hierarchical clustering avoids to assume the number of clusters ahead, and provides the structure or the dendrogram which may contain meaningful information.

In our analysis, we applied the agglomerative hierarchical clustering based on the (dis)similarity matrix addressed via DTW method. Complete linkage criterion, Equation 2.4, was used for the clustering algorithm to determine distance between two clusters A and B.

$$d(A,B) = \max_{a \in A, b \in B} DTW_d(a,b) \tag{2.4}$$

One advantage of hierarchical clustering is that the nested structure can be visualized and one has more flexibility to decide the number of clusters without repeating the computation procedure. Without knowledge of possible number of patterns of events, it is crucial to flexibly select the number of clusters and calculate the typical series for each cluster as the reference event patterns.

2.5.2.4 Pattern Extraction

After clusters of time series being decided, we need to extract the typical shape of time series for each cluster. In the context of time series clustering, a typical shape could be either the medoid series of the cluster, or the average series of the cluster, which both can refer to the center of a cluster. In general, center of a cluster of series can be defined as the series that minimizes the $within - cluster\ sum\ of\ square$. The center is called *medoid series* if the center is an series in the cluster; otherwise, it is called the *average series* if the definition of the center is not restricted to a real series in the cluster. In our study, the average series is referred as the typical pattern of a cluster.

Based on DTW, several methods have been proposed to average a set of time series. We applied a global averaging strategy, DTW Barycenter Averaging (DBA) [Petitjean et al., 2011], to calculate the average series of each cluster. Overall, DBA is an iterative algorithm that refines an initial (potentially arbitrary) average series to minimize the sum of the squared distance (SSD) to the other series.

Suppose $L^k = \{l_1, ..., l_k\} \in R^k$ is the average series with length k, then SSD is defined by coordinate-to-coordinate distances in Equation 1.4, which the contribution of one coordinate of the average series to the total SSD is the sum of distances between this coordinate and the coordinates of other series associated to it in the computation of DTW. The choice of $d(\cdot)$ is Euclidean distance in our study.

$$SSD(L^k) = \sum_{i=1}^{k} \sum_{x \in assoc(l_i)} d^2(l_i, x) \tag{2.5}$$

The minimization of Equation 1.4 is achieved through a coordinate-wise calculation, which minimizes the partial sum for each coordinate by calculating the barycenter. In particular, let $S = \{S_1, ..., S_N\}$ be the cluster on which one need to average, then each iteration of the algorithm contains two steps: 1) compute $DTW_d(L^k, S_i),\ \forall S_i \in S$, where L^k is the temporary average at current iteration, to find associations between coordinates of the average and coordinates of all S_i's; 2) update each coordinate of the average as the average of coordinates associated to it in step 1).

Let $l_i,\ i = 1, ..., k$ be the coordinates of the average series at current iteration, and $l'_i,\ i = 1, ..., k$ be the coordinates of the average series at next iteration. Denote the set of coordinates associated with l_i in the computation of $DTW_d(L^k, S_i),\ \forall S_i \in S$ as $assoc(l_i)$. Then, the DBA algorithm refines the average series by:

$$l'_i = \frac{\sum\limits_{x \in assoc(l_i)} x}{|assoc(l_i)|} \quad \forall\ i = 1, ..., k \tag{2.6}$$

As we discussed, the DBA method minimizes the SSD iteratively starting from an initial average series. The initiation of the algorithm consists of two parts: 1) the length of the average, k; 2) coordinates of the initial average. Regarding the length of the average, we set the median length of series in each cluster as the length of the average for the corresponding cluster to avoid great loss of information and huge computational burden. Regarding the coordinates of the average, in Petitjean *et al.*'s study [Petitjean et al., 2011], randomly choosing an series in the cluster leads to the best performance in most cases. Combining both parts, we set the initial average series by randomly selecting one series in the cluster with length equals the median length of all series in that cluster.

2.5.3 Series Encoding

The core of our approach is to encode the raw series to its sign change series for both the reference pattern series and the long target series. For the encoding process, it involves smoothing and differentiating the raw series, adjusting the slope series to properly define the sign series, and finally extracting the sign change series for later steps.

2.5.3.1 Smoothing and differentiating

Noise is very common in time series data, and affects the performance of models and methods. In addition, minor events, very small and insignificant waves, should not be counted as separate events by the detection algorithm. To minimize the influence of noise, minor events and controlling the detection resolution, the raw data series are smoothed before encoding.

Many smoothing techniques, such as moving average, cubic spline, etc. could be applied in our analysis. We applied the smoothing splines [Green and Silverman, 1993] to the raw series, and differentiating the smoothed series for the slope series.

The advantage of smoothing splines is to avoid the problem of selecting knots. It uses a maximum set of knots and controls the modeling calculation by adding the regularization term:

$$f(x) = \arg\min RSS(f,\lambda) = \arg\min \ \{\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int_a^b (f''(x))^2\}$$

where λ is a smoothing parameter. The second term in RSS penalizes the curvature of the smoothing function. In this analysis, B-splines are used as the basis functions.

The ordinary smoothing spline is known to exhibit biased estimation near boundaries [Oehlert, 1992, Chunfeng, 2001]. Therefore, we adjusted the one-day smoothing window by an overlapping moving time window to avoid dispersion near the boundaries of the original time window. In particular, the center of each time window is from everyday 00:00 to 23:59, then we extend the time window to both directions for extra n minutes. After smoothing and differentiating the series in the extended time window, only the results from 00:00 to 23:59 are recorded to reconstruct the full smoothed series and the corresponding slope series,d(GL). The *smoothing parameter* (λ) is introduced during smoothing to control the degree of smoothness.

2.5.3.2 Encoding

To increase the computational efficiency and avoid the issues due to the length of matching window and scales, we propose a method to encode the original data series into a sign change series, which is much shorter, simpler, and free of scale. The encoding process is used to simplify both the series of typical patterns we extracted from each cluster, and the target long series. A step-by-step description of encoding process is provided next.

1. After the slope series d(GL) is created, it is firstly adjusted by

$$d'(GL)_i = \begin{cases} d(GL)_i \quad \text{if} \quad |d(GL)| > \theta_e \\ 0 \quad \text{otherwise} \end{cases}$$

where θ_e is a pre-specified *adjusting threshold*, and i is the index of each observation in d(GL).

2. Based on the adjusted slope series, create the sign series for d'(GL) by the definition introduced in Section 2.5.1.

3. Extract the sign change series from the sign series according to the definition Section 2.5.1.

Note that the sign change series only contains elements corresponding to a change of signs in the slope series, which makes the sign change series much shorter and simpler than the raw series. As an example, let the series

$$\text{sign}(d'(GL)) = \{0, 0, 0, +, +, +, 0, 0, -, -, -, -, 0, +, +, -, -, -, 0, 0, 0\}$$

be the sign series of a given adjusted slope series. We can observe 7 changes of sign in the series, then the sign change series would only contain 7 elements by definition.

In particular, the sign change series for the example series is

$$\text{sign change}(d'(GL)) = \{+1, -1, -1, +1, +1, -1, +1\}$$

according to definition of sign change.

2.5.4 Exact Pattern Matching and Overall Algorithm

The last step is the exact pattern matching, which we search for the encoded reference patterns in the encoded long target series. Both the reference patterns and the long target series are characterized by the corresponding sign change series, which makes the computation in the matching step more efficient and less expensive. We highlight that for the matching step, the algorithm matches and searches for the exact same patterns instead of fuzzy matching.

The overall algorithm is summarized as Algorithm 1. It implements the methods we have introduced in the above subsections and outputs 1) event attributes table for all detected events, including event start and end time, etc., and 2) a binary series indicating whether a data point belongs to an event.

Note that at step 4 (a,b), evens are detected by matching reference patterns in the sign change series of target series, and the algorithm records the index of the raw series S_i to denote the start and end of each detected event. For example, m_j is a segment of idx for jth matching, then $\min(m_j)$ and $\max(m_j)$ are the start and end index of jth event in S_i.

```
Algorithm 1: Overall algorithm for event detection
Input:
Reference patterns $\mathbf{P}$=$\{P_1, ..., P_n\}$;
Target series $\mathbf{S}$=$\{S_1, ..., S_N\}$;
Parameter $\Theta = (\lambda, \theta_e)$ ;
Output:
Event information table $T_e$;
Event indicator series $I_e$;
Initialize $I_e = 0$;
for ( $i = 1, ..., N$ ) {
    1. Smooth $S_i$ as $S_i^s$ by a smoothing degree $\lambda$;
    2. Differentiate $S_i^s$ as $S_i'$;
    3a. Encode $S_i'$ as $C_i$ by an adjusting threshold $\theta_e$;
    3b. Record the corresponding index of $S_i$ for each element in $C_i$ as $idx$;
    4a. Search for all $P_j$'s from $C_i$ ;
    4b. Record values of $idx$ for each matching as start and end index of an
     event;
    5a. Extract event attributes based on a segment from $S_i$ between start
     and end index;
    5b. Update $I_e = 1$ between start and end index to indicate the detected
     events ;
}
```

2.5.4.1 Optimize the parameters

In algorithm 1, it involves two parameters, $\Theta = (\lambda, \theta_e)$: smoothing degree and slope adjusting threshold, through which we are able to control the resolution of the detection. In particular, the algorithm is designed to ignore minor peaks and combine grouped minor peaks. To what extend those minor peaks would be ignored or grouped is determined and adjusted via different values of $\Theta = (\lambda, \theta_e)$. The optimization of Θ could be achieved through a training process based on pre-labeled event data, such that the optimized Θ is well adopted to customized local data and various demands.

In the training, we optimize a parameter set $\Theta = (\lambda, \theta_e)$ by minimizing an modified 0-1 loss function defined in Equation 1.6. According to the characteristics of the proposed approach, we expect the 0's and 1's in the error series, $e = |y - \hat{y}|$, to be clustered, which $e_i = 1$'s are more likely grouped instead of scattered or randomly spaced. Hence, we propose an $modified\ \ 0-1\ loss\ function$ which penalizes more for the larger grouped e_i's by an upper bounded step weight coefficient.

Define the error series $e = \{e_i = |y_i - \hat{y_i}|,\ i = 1, ..., n\}$ and assume it contains m clustered segments with $e_i = 1$. The clustered error segment can be formally defined as $e^c_j = \{e_k = 1, e_{k+1} = 1, ..., e_{k+n_j} = 1 \mid e_{k-1} = 0, e_{k+n_j+1} = 0\}$ where $\ j = 1, ..., m$.

Then the adjusted 0-1 loss function is defined as

$$\begin{aligned}\ell(\lambda, \theta_e|\hat{y}, y) &= \sum_{j=1}^{m}\sum_{i=1}^{n_j} w_j|y_i - \hat{y}_i| \qquad \text{where} \\ w_i &= \frac{\min(\lfloor \sum_{e_i \in e_j^s} e_i/6 \rfloor, 5)}{c}\end{aligned} \tag{2.7}$$

where c is an arbitrary constant to scale the loss for interpretation.

The modified 0-1 loss function in Equation 1.6 assigns higher weight to missed events or long missed partial events, and lower weight to missed or over-detected boundaries due to the smoothing and adjusting in the encoding process. The proposed loss fits the task of event detection, because, theoretically, there is no exact start and end of a CGM event, and we commonly used the local minima as start and end during data preparing.

2.6 Real world data study

In the Welldoc CGM data, many users' data have long recording history. However, due to the capacity of manual labeling, our sample data contained 48121 CGM readings in the first 30-day use of 6 random patients among all Welldoc provided CGM users. To optimize parameters and examine the performance of our approach, all events in the sample data are manually labeled independently by multiple team members supervised by an expert.

We split the labeled data into two separate subsets: the first 25 days data as training set for optimizing the parameters, and the later 5 days data as the testing set to examine the performance. Overall, we labeled 394 and 93 events respectively in training and testing datasets. We trained for optimal parameters at both individual and global levels, and tested their performance separately.

2.6.1 Pattern recognition results

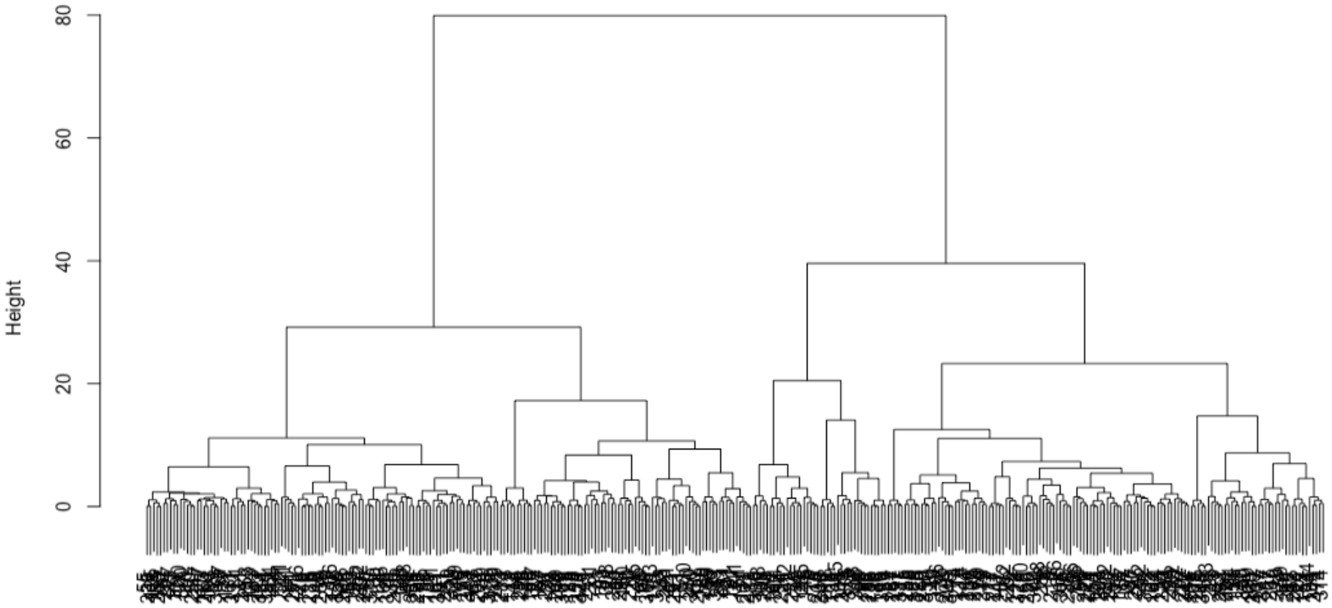


Figure 2.3: Dendrogram in hierarchical clustering. A complete structure for all pre-labeled events (sub-sequences).

At the pattern recognition step, we identified five major CGM event patterns as reference patterns. In the hierarchical clustering based on DTW similarity measure, we initially discovered 6 main clusters according to the complete dendrogram (Figure 2.3), and summarised the clusters by their average series using DBA method in Figure 2.4. Then, we smoothed and encoded the average series to extract the reference patterns. Six cluster averages collapsed to five encoded reference patterns in Table 2.1.

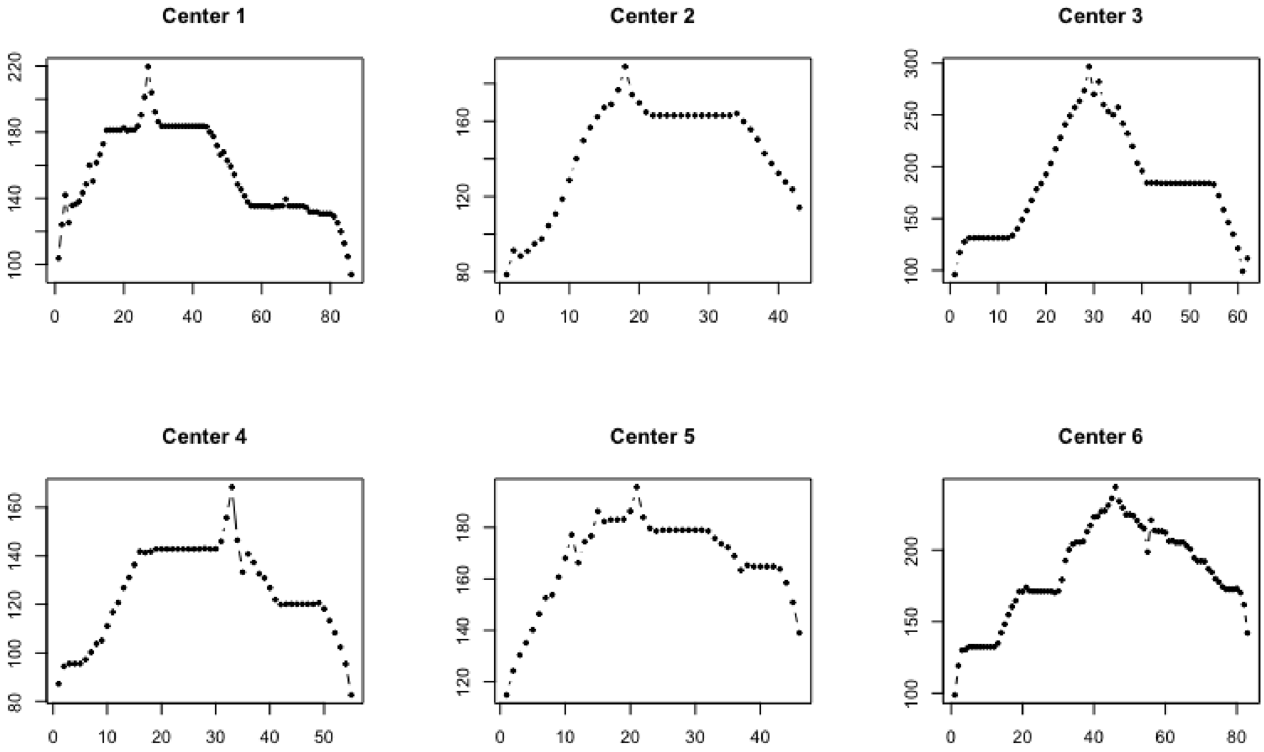


Figure 2.4: Typical events (average series) for 6 clusters.

While encoding the reference patterns, we have considered multiple smoothing degrees to identified all possible patterns. We noticed that some patterns in Table 2.1 are nested, for example, pattern 5 is contained in all other patterns, because it is the most fundamental peak shape. Therefore, it is crucial to carefully order the major patterns for next step, searching and matching reference patterns from the long CGM series. One should search for parent patterns before child patterns. For example, before searching for pattern 4 and 5, we need to exhaust events that matches with pattern 3, and remove the matched segments from the sign change series for later searching.

No.	Pattern	Cluster	Searching Order
1	”+1,-1,-1,+1,-1,+1,-1,+1”	1	1
2	”+1,-1,-1,+1,-1,+1”	1,2,5	2
3	”+1,-1,+1,-1,-1,+1,-1,+1”	3,4	1
4	”+1,-1,+1,-1,-1,+1”	4	2
5	”+1,-1,-1,+1”	5,6	3

Table 2.1: Six center sequences collapsed to five major event patterns after encoding.

2.6.2 Parameter optimization

To optimize the parameters in the approach, we tested two strategies: *individualized parameters*, which we train and minimize the loss function only based on the user's own data, and *global parameters*, which we mixed all users' data for training. The full ranges for two parameters, over which the algorithm is optimized, are $[0.003, 0.006]$ and $(0, 0.1)$ for adjusting threshold θ_e and smoothing degree λ. The training results are presented in Table 2.2. The individualized optimal parameters exhibited variation on both parameters among different users.

2.6.3 Testing results

We applied both individualized and global optimal parameters on the testing set, and measured their performance separately in terms of the proposed loss. Table 2.3 shows performance including testing errors and number of detected events for each user using different sets of parameters. We noted that the algorithm performed better when it incorporated with global optimal parameters.

As an one-day example, Figure 2.5 and Table 2.4 shows the detected events and the corresponding attributes. For the Start/End status, we define

Strategy	User ID	Number of events	Optimal parameters (λ^*, θ_e)	Detected events	Training error
Global	All	394	(0.7074, 0.0040)	352	454.5833
Individual	1	62	(0.5072, 0.0050)	54	80.8500
	2	64	(0.5072, 0.0042)	53	44.9500
	3	64	(0.1869, 0.0042)	68	52.2167
	4	78	(0.9866, 0.0038)	68	40.5833
	5	67	(0.5072, 0.0050)	58	32.3167
	6	59	(0.7074, 0.0048)	51	34.5000

Table 2.2: Optimal Parameters based on different training strategies. $\lambda^*=\lambda \times 10^4$.

		Individualized parameters		Global parameters	
User	Number of events	Detected Events	Testing Error	Detected Events	Testing Error
1	16	14	7.73	14	4.13
2	12	13	2.17	13	2.12
3	17	18	41.65	18	46.95
4	13	11	9.60	11	10.15
5	19	16	13.07	16	6.88
6	16	14	11.30	14	9.33
Total	93	86	85.52	86	79.57

Table 2.3: Performance evaluation using individualized and global optimal parameters

H: glucose>180

N: glucose∈[70,180]

L: glucose∈[54,70)

VL: glucose<54

and the severity score [1] is a defined value represents the severity of an event with range 0 to 9.

[1] Detailed definition of severity scores are in Appendix A Table A.1

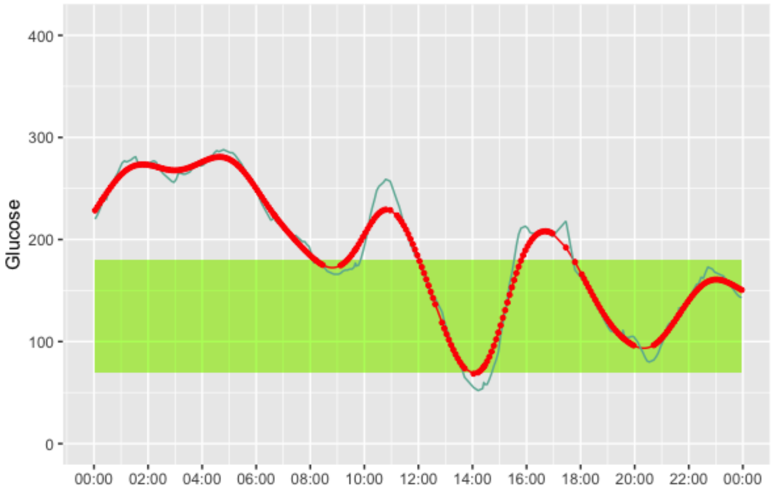


Figure 2.5: Detection for an example day

Event	Start Time	End Time	Start Status	End Status	Severity Score
1	(Previous day) 18:57	08:27	N	N	9
2	09:07	14:02	N	L	3
3	14:12	19:57	L	N	3
4	20:42	23:57	N	N	0

Table 2.4: Event attributes corresponding to example day in Figure 2.5

2.7 Discussion

In this chapter, we developed an automated algorithm for event detection and event information extraction in the context of univariate time series. The method is based on exact pattern matching after characterizing the raw data series by simpler and scale free sign change series via encoding techniques. In the typical pattern matching problems, one suffers from the following challenges: the uncertainty of target time window length; the issues of scaling and shifting; and redundant and expensive computation cost. Our approach mitigates these problems by encoding the long streaming series to a sign change series based on its slope series. By doing so, the shape of the potentially matched segments are characterized by the sign change series and free of considering length, scale and shifting of segments. The computation is more efficient while matching exact number/text series with finite elements. In addition, the redundant computation for matching is avoided if assuming events are mutually exclusive, for example, in our scenario.

The proposed approach performed accurately and efficiently in the real data study, and it has potentials to be applied to other streaming time series data analysis for accurate event detection and pattern matching problems. In the approach, two parameters are used to control the resolution of the detection, which provides flexibility to adjust the algorithm according to different tasks and applications in different areas.

In addition, characterizing the shape of reference patterns and long target series is not restricted to use first order derivative. For more complicated event patterns, the encoding process can be extended with higher order derivatives. Furthermore, by carefully defined and adjust the encoding process, the current approach can be potentially extended to multivariate time series data mining problems as well. More future works are needed to develop and validate the extension of the current method.

The limitations of the approach are as follows. Firstly, the reference patterns are identified based on limited labeled data. It may not cover all possible major patterns generically due to the data limitation. This encourages the further studies on larger and heterogeneous user samples. On the other hand, the parameter optimization process was based on the manually labeled data, which could potentially be sub-optimal in more general population. Data labeling quality may influence the training results and the measure of performance. The proposed modified 0-1 loss function mitigated the influence of labeling quality regarding to the minor disagreements of event boundaries. However, the general understanding of a CGM event and the desired resolution of the detection task among different members during labeling may still affect the training and testing.

Chapter 3: Analysis of Continuous Glucose Monitoring Data: Within-subject Data Analysis

3.1 Introduction

In CGM industry and related health care area, there are many metrics are widely used in CGM reports with well defined targets [Battelino et al., 2019]. In Table 3.1, we listed 10 metrics with corresponding descriptions and industry recognized targets. It is obvious that some metrics are associated. For example, glucose management indicator (GMI) is linearly deterministic by mean glucose. Therefore, not all metrics are necessary to describe patients' status, and which suggests a possible low dimensional framework to measure patients' status. A proper low-dimensional framework can be more applicable in practise, and it highlights the most important features without redundant information. With the development of CGM devices and its collaboration with personal smart devices, a low-dimensional framework with visualization can provide more intuitive descriptions of users' glucose status and present more readable reports for CGM users.

The objective of this chapter is to build a low dimensional framework with a few selected important metrics to measure users status, and visualize the change

Metrics Name	Description	Target
Number of days CGM worn	–	All days
% of time CGM is active	–	>70%
Mean glucose (Mgl)	Average glucose	–
Glucose management indicator (GMI)	3.31 + 0.024 × Mgl	<7%
Glycemic variability (CV)	Coefficient of variation	≤36%
Time above range (TAR_{vh})	% of readings >250 mg/dL	<5%
Time above range (TAR_h)	% of readings 181–250 mg/dL	<25%
Time in range (TIR)	% of readings 70–180 mg/dL	>70%
Time below range (TBR_l)	% of readings 54–69 mg/dL	<4%
Time below range (TBR_{vl})	% of readings <54 mg/dL	<1%

Table 3.1: Standardized CGM metrics for clinical care

of status over time. In this chapter, we will first explore the correlation between metrics, then, select important metrics via principal component analysis (PCA), and finally build the framework to measure and visualize user's status longitudinally.

For the later analysis, we note that metrics in Table 3.1 could be calculated within any time window depending on the customized demands, such as daily, weekly, biweekly, monthly, etc. In order to track the change of users' status, metrics should be repeatedly calculated within a fixed time window for a certain length of period. Therefore, the data structure in our study is a mixture of repeated measures from each single user on different metrics, which challenges the independence assumptions in the cross-sectional correlation study and PCA.

3.2 Challenges and related work

Analysis on repeated measure data, unlike the cross-sectional data, challenges the independence assumption in most cases due to the mixture of within- and between- subject observations. Violation of independence may produce biased results due to different patterns between- versus within- subjects. Applying methods and techniques that require the independence assumption to analyze dependent data is a common practise. However, many studies showed that it may produce erroneous results [Molenaar, 2004, Aarts et al., 2014].

One possible solution is aggregating observations from same subjects. In such way, one may solve the violation of independence. However, aggregating multiple observations by summary statistics, for example taking averages, could lead to biased results [Myung et al., 2000]. On the other hand, using average or other summary statistics may cause information loss, such as time variant effects. Then it prevents the researchers from discovering possible trend or change over time.

To avoid information loss and biased results, some studies in literature have suggested the idea of splitting between- and within- subjects variations to release the independence assumption. The extended ANOVA for repeated measures introduced the method of splitting between- and within- subjects variations to adjust the effect caused by multiple measures from the same individuals [Girden, 1992]. Bland and Altman [Bland and Altman, 1995b, Bland and Altman, 1995a] later applied the within-subject correlation in biostatistics to analyze the common within-subject association for paired repeated measures, which are two corresponding measures assessed for each participant on two or more occasions. Westerhuis et al. [Westerhuis et al., 2010] also suggested a mixed model framework to separate the between- and within- subjects variations. In addition, multilevel modeling is proposed to analyze between- and within- subjects variations simultaneously, instead of focusing on one type of variations.

To our demand of generating more complete reports for users and physicians, and reducing the influence caused by violation of independence, we will split between- and within- subjects variations and focus on the within-subject analysis.

3.3 Within-subject correlation

Correlation is a commonly used measure to quantify the association between two variables. However, widely used techniques for correlation, such as Pearson, Kendall, and Spearman correlation, all assume independence between observations.

In our scenario, repeated metrics are calculated based on CGM data within a particular length of time, for example daily, to measure the status of a user. As discussed, we care more about the within-subject association between different metrics instead of the between-subject association. First, based on the analysis of covariance (ANCOVA), we can split the total variation and statistically adjust for within-subject variation. Then, we will address the within-subject correlation in the manner of Pearson correlation. After eliminating the between-subject variation, it is worth to note that we still need to relax the independence assumption in order to calculate the Pearson correlation for the within-subject data. Moreover, we also assume common correlation for all subjects in current analysis.

Let x_{ij} be an observation for any metric, then it can be decomposed into:

$$x_{ij} = x.. + \underbrace{(x_{i.} - x..)}_{\textbf{between-subject}} + \underbrace{(x_{ij} - x_i.)}_{\textbf{within-subject}} \tag{3.1}$$

where

$$x.. = \frac{1}{N}\sum_i \sum_j x_{ij}$$

$$x_i. = \frac{1}{n_i}\sum_j x_{ij}$$

which, $x..$ is the grand mean of all x_{ij}'s, and $x_i.$ is the group mean of x_{ij}'s within subject i. Then, equation (3.1) can be rewritten into the matrix format as:

$$\boldsymbol{X} = \boldsymbol{X}.. + \underbrace{\boldsymbol{X_b}}_{\textbf{between-subject}} + \underbrace{\boldsymbol{X_w}}_{\textbf{within-subject}} \tag{3.2}$$

The within-subject correlation would be calculated via the covariance matrix of $\boldsymbol{X_w}$:

$$\boldsymbol{\Sigma}_{\boldsymbol{X_w}} = \{\sigma_{ij} = cov(\boldsymbol{x}^i, \boldsymbol{x}^j)\}_{p\times p} = c \cdot \boldsymbol{X_w}^T \boldsymbol{X_w} \tag{3.3}$$

In the spirit of Pearson correlation, the within-subject correlation between two metrics can be defined as

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} \tag{3.4}$$

In addition to correlations, the corresponding p-values are also desired to test if the association is significant. In the hypothesis testing of regular Pearson correlation, the sampling distribution of a certain function of Pearson's correlation coefficient follows Student's t-distribution with degrees of freedom $N-2$ under the null hypothesis, $\rho = 0$, by assuming the joint distribution of the paired variables are bivariate normal distribution and the observations are not associated Specifically, let $r = \hat{\rho}$, then

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \sim t\ (df = N-2)$$

under the null hypothesis $H_0 : \rho = 0$

However, in the case of within-subject correlation, the sampling distribution on which the calculation of p-values based may not be reliable. The main reason is that we relaxed the independence assumption for Pearson correlation in calculation of within-subject correlation. Therefore, an analytic p-value is difficult to decide.

One alternative method is to achieve the empirical sampling distribution via the bootstrap method and construct confidence intervals for the correlation. Bootstrapping methods are free of distribution assumptions and estimate parameter accuracy through random resampling approach [Efron and Tibshirani, 1994]. The bootstrap in the within-subject case is stratified by subjects, which implemented by randomly sampling observations with replacement within each user's data for all users, and then, yielding a bootstrapped sample.

Table 3.2 describes the steps of the bootstrapping in our scenario and the method we applied to decide an empirical confidence interval at a certain level of confidence. The method we applied to generate the the empirical confidence interval is the bias-corrected and accelerated (BC$_a$) interval. In particular, the BC$_a$ interval at $1 - \alpha$ coverage is defined as

$$\text{BC}_a : (r_{low}, r_{up}) = (r^*_{\alpha_1}, r^*_{\alpha_2})$$

where

$$\begin{aligned}\alpha_1 &= \Phi(\hat{z}_0 + \frac{\hat{z}_0 + z_{\alpha/2}}{1 - \hat{a}(\hat{z}_0 + z_{\alpha/2})}) \\ \alpha_2 &= \Phi(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha/2}}{1 - \hat{a}(\hat{z}_0 + z_{1-\alpha/2})})\end{aligned}$$

Here $\Phi(\cdot)$ is the standard normal cumulative distribution function and $z_{\alpha/2}$ is the $100 \times \alpha/2$th percentile of a standard normal distribution. Values of $\hat{z}_0$ and $\hat{a}$ can be calculated by

$$\begin{aligned}\hat{z}_0 &= \Phi^{-1}(\frac{\#\{r^* < r\}}{B}) \\ \hat{a} &= \frac{\sum_{i=1}^{n}(r_{(\cdot)} - r_{(i)})^3}{6\{\sum_{i=1}^{n}(r_{(\cdot)} - r_{(i)})^2\}^{3/2}}\end{aligned}$$

where B is the number of replicates, r is the original parameter estimate, n is the number of observations in the original sample. The bias-correction factor $\hat{z}_0$ is estimated by the proportion of the bootstrap estimates less than the original parameter estimate. The acceleration factor $\hat{a}$ is estimated through jackknife resampling, which involves generating n replicates of the original sample. The ith jackknife replicate is obtained by leaving out the ith observation of the original sample until n samples are obtained. For each jackknife sample, the calculated estimate is denoted as $r_{(i)}$. The average of these estimates is

$$r_{(\cdot)} = \sum_{i=1}^{n} \frac{r_{(i)}}{n}.$$

Step	Description
1	A certain number of pairs (x_k^i, x_k^j) are sampled with replacement from each user
2	Combine the selected data from multiple users as one bootstrapped sample
3	The sample correlation coefficient r is calculated based on the sampled data
4	Repeat Step 1 − 3 a large number of times to reach the empirical distribution of r
4	A $100(1-\alpha)\%$ confidence interval for ρ can be defined by the BC_a method as (r_{low}, r_{up})

Table 3.2: Steps for bootstrapping and empirical confidence interval

3.4 Important variable selection

The second step to build up the framework is to reduce the dimension by selecting important metrics among all available ones, which can explain the data variation as much as possible.

To achieve dimension reduction, principal component analysis (PCA) is widely applied in many studies and areas. Traditional PCA decomposes multivariate data into orthogonal components and approximates the total variation by the first few principal components (PC). However, one important disadvantage of regular PCA is that the interpretation of PCs are difficult. Weak interpretability of PCs affects the direct application of PCA in our analysis, because physicians and users need precisely defined metrics instead of a linear combination of multiple metrics. Moreover, there are no well defined and industry recognized targets for PCs, which will be applied as thresholds to classify users' status.

Against the disadvantage of regular PCA, we decided to apply an unsupervised variable selection procedure to recognize important metrics based on PCA, instead of using PCs directly. To apply PCA, we still need to split the within-subject variation from between-subject variation and relax the independence assumption. In section 3.3, the within-subject matrix, $\boldsymbol{X_w}$, is calculated. The mail idea of the selection procedure is to apply the regular PCA on $\boldsymbol{X_w}$ and determine the important metrics using the first k PCs. A step-by-step description of our approach is provided next.

Step 1: conducting within-subject PCA

Apply PCA on centered and scaled $\boldsymbol{X_w^{N \times p}}$ to avoid the influence of metrics with very different scales. With PCA, the output contains p PC eigenvalues, $\boldsymbol{\lambda^p} = [\lambda_1 \;\; ... \;\; \lambda_p]$ and a $p \times p$ PC loading matrix $\boldsymbol{r^{p \times p}} = [\boldsymbol{r_1} \;\; ... \;\; \boldsymbol{r_p}]$.

Step 2: determining number of important PCs

Several criteria have been proposed to determine the number of PCs should be selected. Some widely used criteria are as following: 1) check the scree plot, which shows the decrease of the marginal explained variance for additional PCs. 2) include first k PCs until the proportion of cumulative explained variance reaches a predetermined threshold, for example, 90%. 3) include PCs with explained variance, λ_i, is greater than 1.

In our analysis, all above mentioned criteria are applicable. We decided to determine value of k based on the scree plot. A Scree Plot is a line segment plot that shows the eigenvalues for each individual PC with eigenvalues on the y-axis and the number of PCs on the x-axis. The scree plot criterion selects all PCs before the curve flattens out. That is, we will select the first k PCs where the $(k+1)$th and later PCs offer little additional explained variance.

Step 3: selecting the important metrics based on first k PCs

The principal of the third step is to select the metrics which preserve variation based on PCA. A common way to locate important original variables is based on loadings in PCA or the correlation between original variables and PC scores [Guo et al., 2002]. Three criteria could be used in our analysis: 1) for first k PCs, the variable with the largest loadings are selected; 2) retain variables with highest association with each of the first k PC scores [Jolliffe, 1972]; 3) select k variables with the highest association with all first k PC scores.

For the selection using criteria 1 and 2, the variable less correlated with previous selected metric(s) should be selected if multiple metrics are candidates based on the criterion. For criterion 3, one should select the highest associated metric as first variable, and all later selections should not be strongly correlated with previous selection based on the within-subject correlation. The calculation of correlation between variable and all first k PC scores can be addressed based on linear regression after removing the between-subject variation.

For instance, the correlation between variable X_i and fist k PC scores can be access via the following linear model:

$$\boldsymbol{X_w^i} = \beta_0 + \boldsymbol{X_w^{N\times p}} \boldsymbol{r^{p\times k}} \boldsymbol{\beta^{k\times 1}} + \boldsymbol{e}$$

where $\boldsymbol{X_w^{N\times p}}$ is the within-subject matrix, $\boldsymbol{r^{p\times k}}$ is the loading matrix for the first k PCs, and $\boldsymbol{e}$ is the error term. Then the correlation between variable $\boldsymbol{X_w^i}$ and $\boldsymbol{X_w^{N\times p}} \boldsymbol{r^{p\times k}}$ can be calculated through the coefficient of determination

$$r = \sqrt{r^2} = \sqrt{1 - \frac{\sum(x_w^i - \hat{x}_w^i)^2}{\sum(x_w^i - \bar{x}_w^i)^2}}$$

Finally, k metrics are selected to build a low dimensional framework. Patients glucose status can be classified into 2^k levels based on the k selected metrics and the corresponding targets. Visualization of status and the trajectory of status over a period can be build accordingly depends on the selection results.

3.5 Real world data analysis

Among the data provided by Welldoc, the CGM data included more than 200 users CGM records with different length of activation period. The CGM data contains the glucose readings, the corresponding time stamps, and other information. The length and pattern of CGM usage vary among included users. To avoid users with longer history dominating the results, our analysis was conducted within a fixed time window, three months, for all users. The three month period fits the guidance that diabetic patients are recommended to meet their physicians at least once every three months. Therefore, the designed study period would be meaningful and applicable in clinical practice.

Metrics No.	Metrics	Mean (SD)
2	% of time CGM is active (Act%)	93.52 (17.04)
3	Mean glucose (Mgl)	159.65 (47.03)
5	CV	24.64 (9.25)
–	SD-gl	40.04 (20.29)
6	TAR$_{vh}$	9.83 (19.68)
7	TAR$_{h}$	18.15 (16.45)
8	TIR	70.83 (28.24)
9	TBR$_{l}$	0.90 (2.49)
10	TBR$_{vl}$	0.28 (1.64)
6+7	TAR=TAR$_{h} + TAR_{vh}$	27.98 (28.30)
9+10	TBR=TBR$_{l} + TBR_{vl}$	1.19 (3.49)

Table 3.3: Simple summary of selected metrics

Among 10 widely used metrics in regular CGM report in Table 3.1, we excluded metrics 1 and 4, number of days CGM worn and GMI, because the values of metric 1 in the Welldoc data are almost constant, and metric 4 is linearly determined by the mean glucose. We added three additional metrics, *time above range* (TAR) and *time below range* (TBR), to introduce less granular metrics comparing to the original metrics; and the standard deviation of glucose (SD-gl) to measure the variability of glucose. Values of each metric was calculated at daily level for later analysis. Table 3.3 shows the included metrics and summary statistics by mean and standard deviation among all users.

Then, we assessed the within-subject correlation coefficients and the corresponding empirical 95% confidence intervals for all pairs of selected metrics. For bootstrapping, we sampled 50% days from each user in each replicate and repeated the procedure 5000 times to address the CIs. The calculation results are presented in Table 3.4. Among 11 included metrics, we observe that 1) TIR, TAR, and Mgl are strongly correlated with each other; 2) SD and CV are highly correlated; The strong or moderate correlations between metrics confirmed that not all metrics are necessary to measure users glucose status while building a low-dimensional framework and visualizing users glucose status change.

Metrics	Act%	Mgl	CV	SD	TAR$_{vh}$	TAR$_h$
Act%	1	-0.033	0.122	0.116	-0.042	-0.011
Mgl	(-0.076,0.006)	1	-0.147	0.341	0.832	0.408
CV	(0.079,0.165)	(-0.183,-0.114)	1	0.834	-0.034	-0.0.033
SD	(0.075,0.158)	(0.299,0.379)	(0.823,0.844)	1	0.343	0.170
TAR$_{vh}$	(-0.094,0.005)	(0.819,0.844)	(-0.075,0.008)	(0.294,0.386)	1	-0.025
TAR$_h$	(-0.048,0.022)	(0.368,0.447)	(-0.064,-0.003)	(0.138,0.203)	(-0.065,0.011)	1
TIR	(-0.008,0.079)	(-0.838,-0.818)	(-0.042,0.024)	(-0.409,-0.346)	(-0.644,-0.606)	(-0.769,-0.723)
TBR$_l$	(-0.065,0.014)	(-0.276,-0.231)	(0.277,0.362)	(0.058,0.135)	(-0.113,-0.072)	(-0.140,-0.099)
TBR$_{vl}$	(-0.105,-0.020)	(-0.146,-0.102)	(0.219,0.282)	(0.070,0.133)	(-0.064,-0.026)	(-0.055,-0.018)
TAR	(-0.073,0.003)	(0.851,0.868)	(-0.081,-0.017)	(0.324,0.386)	(0.614,0.653)	(0.733,0.778)
TBR	(-0.087,0.004)	(-0.274,-0.228)	(0.318,0.399)	(0.080,0.157)	(-0.113,-0.072)	(-0.129,-0.087)

Metrics	TIR	TBR$_l$	TBR$_{vl}$	TAR	TBR	
Act%	0.042	-0.017	-0.053	-0.036	-0.038	
Mgl	-0.829	-0.253	-0.124	0.860	-0.251	
CV	-0.011	0.327	0.251	-0.048	0.366	
SD	-0.379	0.098	0.100	0.356	0.121	
TAR$_{vh}$	-0.625	-0.092	-0.046	0.633	-0.091	
TAR$_h$	-0.747	-0.119	-0.037	0.757	-0.108	
TIR	1	-0.010	-0.051	-0.986	-0.016	
TBR$_l$	(-0.014,0.033)	1	0.289	-0.153	0.898	
TBR$_{vl}$	(-0.075,-0.028)	(0.228,0.345)	1	-0.058	0.680	
TAR	(-0.988,-0.985)	(-0.176,-0.131)	(-0.079,-0.037)	1	-0.144	
TBR	(-0.041,0.008)	(0.864,0.919)	(0.640,0.719)	(-0.166,-0.122)	1	

Table 3.4: Within-subject correlation and 95% CIs. Splitted by the diagonal, the upper triangle shows the estimated correlation, and the lower triangle shows the empirical 95% CIs.

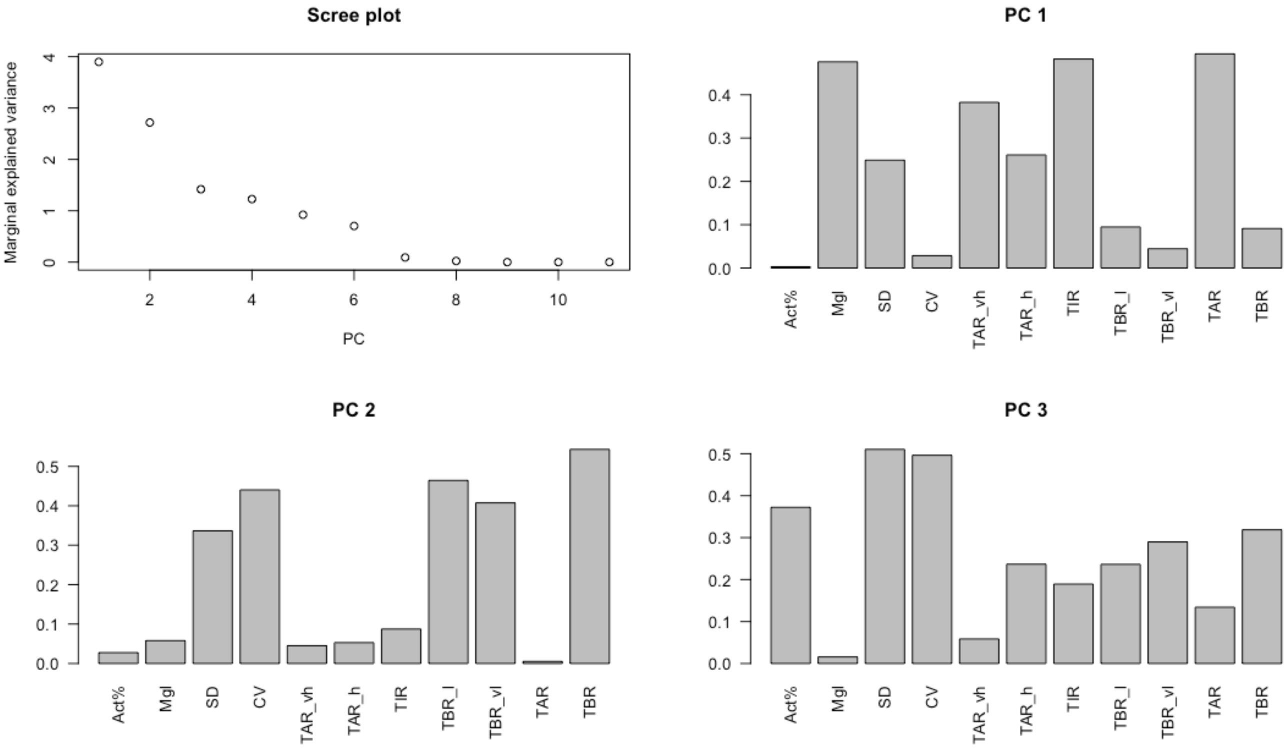


Figure 3.1: Within-subject PCA. Scree plot and loadings for PC 1–PC 3 in absolute values

Next, we applied the within-subject principal component analysis based on the within-subject matrix which we calculated in the previous step. Figure 3.1 exhibits the scree plot that shows the explained variance by each PC and the absolute loading coefficients for the first three components. According to the scree plot, three or six components are two proper values for k. In our study, lower dimension will be more applicable to interpret and visualize in clinical reports. The cumulative variation explained by the first three components exceeds 75% of total variation.

To select important metrics based on within-subject PCA, three previously discussed criteria are presented in Figure 3.1 and Table 3.5. Based on different

Metric	PC 1	PC 2	PC 3	First 3 PCs
Act%	0.0118	-0.1338	-0.5497	0.4626
Mgl	-0.952	-0.0636	0.2336	0.9771
SD	-0.4737	-0.8604	-0.5845	0.9525
CV	-0.0375	-0.9292	-0.7079	0.9248
$\text{TAR}_v h$	-0.7762	-0.2005	0.0296	0.7849
TAR_h	-0.4905	0.0394	0.4703	0.5185
TIR	0.9605	0.2265	-0.3696	0.9865
TBR_l	0.1313	-0.5161	-0.0874	0.6325
$\text{TBR}_v l$	0.0444	-0.4197	-0.0213	0.4242
TAR	-0.9736	-0.1255	0.3795	0.9857
TBR	0.1173	-0.5867	-0.0738	0.7628

Table 3.5: Within-subject correlations between original metrics and first 3 PCs

	Criteria Description	Selection Results
1	Max loading	TIR, TBR, CV
2	Max correlation: single PC	TIR, CV, Act%
3	Mas correlation: multiple PCs	TIR, CV, TBR

Table 3.6: Important metrics selected by 3 criteria

criteria, we selected important metrics and summarized the results in Table 3.6. Using criteria 1 and 2, we observed that, in PC1, TAR, TIR, and mean glucose were dominating, and among them TIR is selected because it is more commonly used in practice. The ensemble of selections from three criteria gave the final selection, TIR, TBR, and CV.

After the selection process, we selected TIR, TBR, and CV to build the low-dimensional framework to measure users' status and visualize a longitudinal trace of their status. In general, TBR is not frequently being observed, and with little variation. Therefore, TIR and CV are used as major dimensions (y- and x-axis) and TBR is an additional indicator as the third dimension in a 2-D visualization.

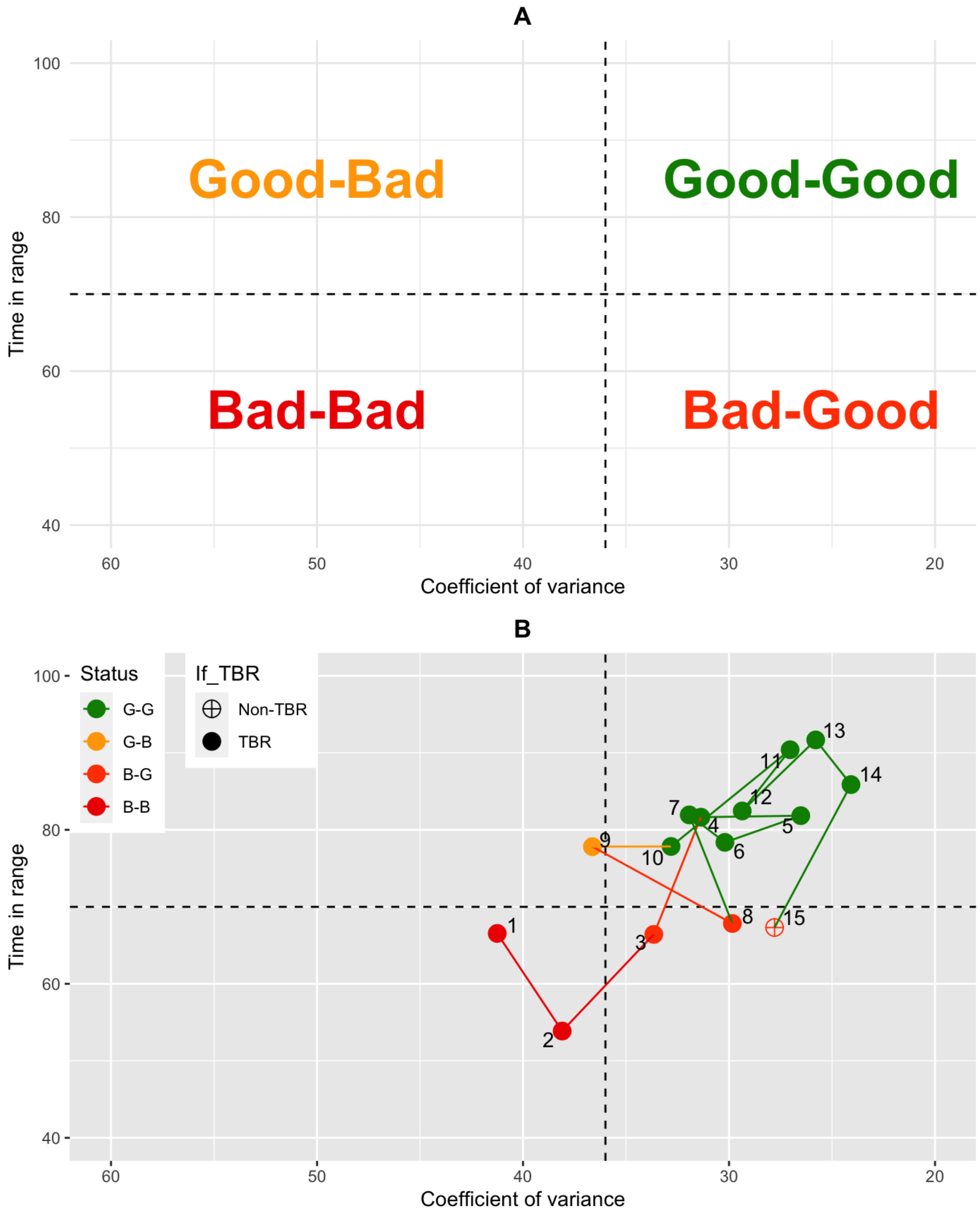


Figure 3.2: (A) Four quadrants classification. Values on x-axis are reversed to position the "G-G" status on the top-right quadrant and the "B-B" status on the bottom-left quadrant. (B) Example of status trajectory for a user over three-month period

Using the targets for TIR (70%) and CV (36%) in Table 3.1, we classified a user's status into 4 quadrants (categories) as illustrated in Figure 3.2-A. In the 2 × 2 labeling, the first digit indicates the status of TIR, and the second digit is for CV. "Good" (G) denotes user reaches the target for a particular metric; otherwise, "Bad" (B) is labeled. The third dimension, using different point shapes, indicates whether TBR in the corresponding period is greater than 5%. As an example, Figure 3.2-B shows a status trajectory of one user over three-month period, and each point summarizes the user's status in one week based on daily metrics by averages.

3.6 Discussion

In this chapter, we focused on analysis methods for data with multiple measures from the same group of subjects. We addressed the within-subject correlation, and further applied an approach based on within-subject PCA for unsupervised variable selection. Finally, we built a low-dimensional framework to measure and visualize users' glucose status across a period of time.

The idea of selecting important variables based on PCA can be extended to general data mining problems or some dimension reduction problems when using the PCs directly is not applicable. The important variables in our context refers to the variables that preserves the most variation of the original data, which fits the task of ordinary PCA well. For the case of different data mining tasks, PCA may not be proper without further validation.

The limitations in our study are as follow. The first limitation is that the unsupervised variable selection approach based on within-subject PCA mainly focuses on maintaining more variation of the original data after variable selection procedure. In fact, data variation may not be the optimal way to measure the information contained in the data. Variable selection methods oriented to variation may cause information loss and further affect the power of the low-dimensional framework in some cases. In the context of CGM and diabetes, the total number of metrics is moderate, and medical knowledge and studies could help confirming and examining the current selection results [Gabbay et al., 2020, Rama Chandran et al., 2018, Urakami et al., 2020]. Another limitation is that the current study is retrospective. We excluded users whose records were shorter than three months. It may cause the current results to be unrepresentative of the population with diabetes.

For the future work, we plan to 1) explore more measures besides variation to guide the variable selection; 2) replicate the analysis on data involves more users, and recognize variations in results.

Chapter 4: Analysis of Continuous Glucose Monitoring Data: Association Analysis Between Event and Time Series

4.1 Introduction

In this chapter, we aim to investigate the immediate effect of real life activities on glucose, such as foods, medications, and physical exercises. That is, the correlation between real life activities and the glucose trajectories in a certain period after the activities. The domain knowledge tells that some foods will increase the glucose, and some medications, for example the rapid insulin, will decrease the glucose. We intended to further statistically verify the immediate effects for activities of combination of foods and medications, and exercises on glucose trajectories.

In Luo *et al.*'s study, the authors defined the similar problem as "event vs. time series" correlation analysis, and transformed the problem to a multivariate two-sample hypothesis testing problem [Luo et al., 2014]. It tests whether two multivariate samples are following the same underlying distribution, which could be assumed known or unknown. In our scenario, we restate the problem to whether the glucose trajectories sampled after a type of activity is significantly different from a random series sample. The dimension of the multivariate sample depends on the length of glucose trajectories we aim to investigate and monitor after a particular activity. In addition, we only focused on the immediate influence of an activity in the current analysis instead of long term effects.

In our context, if a type of activities A and a glucose series S are significantly correlated, then the occurrence of activity A would cause significant change on the glucose series S. This is a qualitative study which we investigate whether the correlation is significant, and the level of correlation is not discussed in this study. Figure 4.1 shows the glucose series and some activities, includes foods and exercise, in one-day time window. It obviously presents rapid glucose increase after food intake.

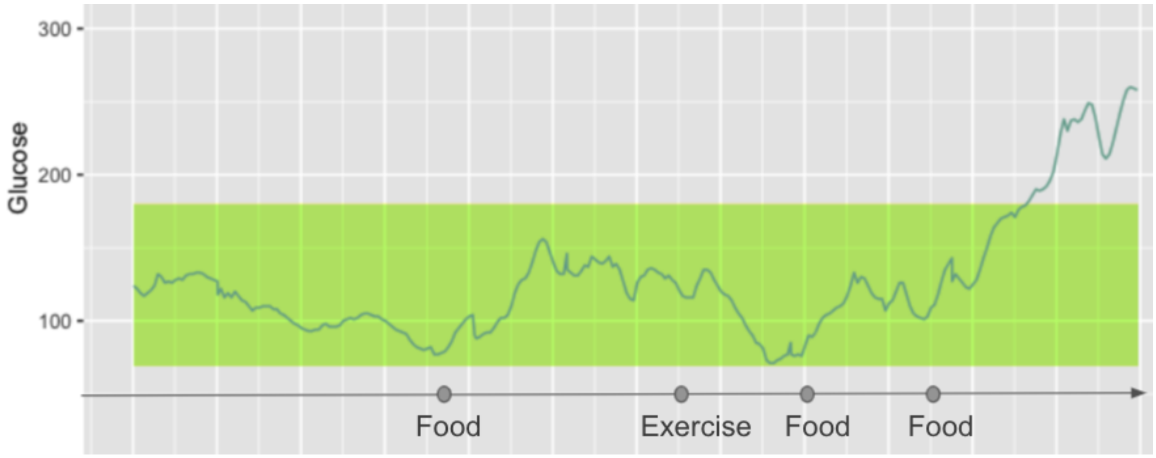


Figure 4.1: Illustrative example of correlation: real life activities and glucose change

4.2 Related study

Box and Tiao discussed the effect of interventions on a response time series through a dynamic linear model based on ARMA model and the binary intervention series [Box and Tiao, 1975]. They estimated the coefficients through maximum likelihood method with properly assumed error distribution and stationary time series. Without strong assumptions, Luo *et al.* transformed the problem to a multivariate two-sample hypothesis testing problem [Luo et al., 2014] and introduced the non-parametric multivariate two-sample tests to study the association between event and time series for telemetry data.

In a multivariate two-sample problem, one tests for the equality of two underlying distributions based on two sets of independent observations. If assuming two underlying distributions are same except their locations, it transmits to a two-sample location problem. For example, if two underlying distributions are assumed to be multivariate normal with equal covariance matrices, Hotelling's T^2 test can be applied to test the equality of their means. In addition, many nonparametric tests have been developed for the multivariate two-sample location problem for different settings [Hettmansperger et al., 1998, Choi and Marden, 1997, Mondal et al., 2015].

Considering the underlying distributions are unknown, many non-parametric tests for more general two-sample comparison have been proposed. Friedman and Rafsky extended the Wald–Wolfowitz run test and the Kolmogorov–Smirnov maximum deviation test to multivariate case [Friedman and Rafsky, 1979]. Biswas and Ghosh proposed a criterion for two-sample tests based on the inter-point distance [Biswas and Ghosh, 2014]. Schilling and Henze proposed methods based on nearest neighbor coincidences [Schilling, 1986, Henze, 1988]. Some non-parametric tests for the general two sample problem are based on pairwise distances between the observations [Liu and Modarres, 2011, Hall and Tajvidi, 2002]

In our study, we would apply a general two-sample test using non-parametric methods, which Schilling proposed based on nearest neighbor coincidences and performed well in many studies [Schilling, 1986, Luo et al., 2014, Mondal et al., 2015]. We noted the differences between a general two-sample multivariate test and the two-sample test in our study. In a general two-sample multivariate test, samples are independently sampled from the underlying distribution. However, in the streaming data scenarios, we need to relax the independence assumption to applied the two-sample test. To ensure the effectiveness of the test, we will firstly verify the performance of the test after relaxing the independence assumption through simulation studies. Then, we apply the method on the real data to investigate the correlation between activities and glucose series.

4.3 Definitions and Theorem

Let $\mathbf{G} = \{G_1, ..., G_{n_1}\}$ and $\mathbf{X} = \{X_1, ..., X_{n_2}\}$ be two independent samples in R^d from distributions $F_1(\cdot)$ and $F_2(\cdot)$. Here, $\mathbf{G}$ denotes the random glucose sample after one type of activity, and $\mathbf{X}$ denotes the random glucose sample with no condition. The hypotheses in the test are:

$$H_0 : F_1 = F_2 \qquad H_a : F_1 \neq F_2$$

The rejection of H_0 means the two distributions are statistically different; that is, the glucose after a certain type of activity is significantly different from random series.

Denote the pooled sample as $\mathbf{Z} = \mathbf{G} \cup \mathbf{X} = \{Z_1, ..., Z_n\}$ with $n = n_1 + n_2$, $\Omega_1 = \{1, 2, ..., n_1\}$, $\Omega_2 = \{n_1 + 1, ..., n\}$, and

$$Z_i = \begin{cases} G_i, & i \in \Omega_1 \\ X_{i-n_1}, & i \in \Omega_2 \end{cases}$$

Definition Z_j is the k**th nearest neighbor** (kNN) of Z_i if exactly $k-1$ Z'_j satisfy $\parallel Z_i - Z'_j \parallel < \parallel Z_i - Z_j \parallel$, where $j' \in [1, n], j' \neq i, j$, and $\parallel \cdot \parallel$ is a norm; denote $Z_j = NN_i(k)$ as kth nearest neighbor of Z_i.

Define an indicator function:

$$I_i(k) = \begin{cases} 1, & \text{if } NN_i(k) \text{ in the same sample with } Z_i \\ 0, & \text{otherwise} \end{cases}$$

Then, the test statistic for the null hypothesis is defined as

$$T_{K,n} = \frac{1}{nK} \sum_{i=1}^{n} \sum_{k=1}^{K} I_i(k) \tag{4.1}$$

It shows the proportion of cases that a data point and its kth nearest neighbor are in the same sample. Intuitively, a small value is expected under the null, because two samples are from the same distribution. Thus, a large test statistic indicates significant difference between two distributions. The asymptotic distribution of $T_{K,n}$ under the null hypothesis is stated in Theorem 1 [Schilling, 1986].

Theorem 1 *If $n_1, n_2 \to \infty$, $\frac{n_i}{n} = \lambda_i$, $i = 1, 2$, d and K are surficiently large, then*

$$(nK)^{1/2}(T_{K,n} - \mu_K)/\sigma_K \sim N(0,1)$$

where the mean and variance

$$\begin{aligned} \mu_K &= (\lambda_1)^2 + (\lambda_2)^2 \\ \sigma_K^2 &= \lambda_1\lambda_2 + 4\lambda_1^2\lambda_2^2 \end{aligned}$$

A large $T_{K,n}$ indicates small p-value which is the evidence against the null. In our analysis, we use p-value $= 0.05$ as the significance level.

4.4 Simulation study

As discussed, the performance of the test method need to be examined after assumption releasing. In addition, the KNN method involves three parameters which need to be decided, K, d, n before applying to real data study. In Schilling's study, the author recommended K should not be smaller than 3; for d, the method worked well when sample size $n \geq 200$ and $d \leq 10$. We will examine the performance of the method using more values of K and d for different sample sizes, and decide optimal K, d, and n by the results.

In the simulation study, we set up two cases, correlated case and uncorrelated case. In the correlated case, we use food as activity which the domain knowledge shows the significant influence, then the two samples we will test are glucose series after food records and a random set of series. For the uncorrelated case, we will test for two random samples.

Besides the performance, we also test for the influence of different values of K, d and the sample size, n. The simulation studies were conducted using samples with size 200, 250, and 300. For each sample size, we generated 200 random trials by random sampling without replacement, and calculated the KNN test statistics in each trial based on different combinations of d and K. The distance measure used in the simulation is the L_2-norm or Euclidean distance.

Table 4.1 and 4.2 exhibit the simulation results in terms of mean and standard deviation of Z statistics in each setting, and the percentage of correct conclusions at significance level $\alpha = 0.05$.

Size=200	K				
d	10	15	20	25	30
10	1.60,1.33,65.5%	1.46,1.32,70%	1.24,1.30,78%	1.06,1.33,81%	0.91,1.27,84%
15	1.32,1.22,74.5%	1.07,1.22,75%	0.84,1.19,82%	0.64,1.16,87%	0.48,1.09,91%
20	1.13,1.21,78%	0.78,1.20,85.5%	0.49,1.17,89.5%	0.26,1.11,93%	0.11,1.11,94.5%
25	1.00,1.10,79%	0.67,1.03,85%	0.52,1.08,87%	0.27,1.10,93.5%	0.11,1.06,95%
30	0.89,1.05,81%	0.65,1.10,86.5%	0.46,1.10,90%	0.25,1.04,94.5%	0.09,1.08,97%
Size=250	K				
d	10	15	20	25	30
10	2.20,1.41,46.5%	2.04,1.39,48.5%	1.88,1.40,54.5%	1.75,1.36,58%	1.65,1.32,60.5%
15	1.70,1.44,60%	1.50,1.44,65%	1.28,1.38,73%	1.07,1.38,76.5%	0.82,1.31,78.5%
20	1.42,1.31,68%	1.17,1.32,77%	0.92,1.32,80.5%	0.67,1.35,84.5%	0.45,1.31,88%
25	1.12,1.11,77%	0.80,1.08,87.5%	0.63,1.05,90.5%	0.45,1.02,91.5%	0.31,1.05,93.5%
30	0.94,1.14,81.5%	0.64,1.13,88.5%	0.46,1.06,91.5%	0.35,1.03,91%	0.18,1.06,95%
Size=300	K				
d	10	15	20	25	30
10	3.05,1.53,24%	2.84,1.54,30.5%	2.70,1.57,31.5%	2.55,1.52,38%	2.29,1.53,42.5%
15	2.19,1.30,45%	1.93,1.34,53.5%	1.65,1.30,62.5%	1.51,1.31,66%	1.33,1.31,71.5%
20	2.15,1.27,48.5%	1.86,1.28,56%	1.53,1.24,65%	1.24,1.25,72%	.99,1.25,81.5%
25	1.75,1.30,58%	1.36,1.23,67%	1.11,1.22,74%	0.94,1.22,78%	0.79,1.21,88.5
30	1.04,1.12,76.5%	0.82,1.13,81.5%	0.67,1.10,87%	0.54,1.13,90%	0.42,1.09,92.5%

Table 4.1: Simulation results for uncorrelated case. Values in the table represent $\overline{z}$, s_z, % of $|z| < 1.96$ out of 200 random trials

For the uncorrelated case, H_0 is true, choice for K and d significantly influenced the underlying distribution for all levels of sample size. For different sample sizes, 200, 250, and 300, the reliable tests can be generated with $K \geq 25$ and $d \geq 25$, $K \geq 30$ and $d \geq 30$, and $K > 30$ and $d > 30$, respectively. For the correlated case, H_0 is false, the method performed well for all combinations of n, K and d in all random trials. Combining the results in both cases, based on different sample sizes, K and d should be at lease 25 to ensure the reliable tests and confident conclusions.

Size=200	K				
d	10	15	20	25	30
10	9.35,1.45,100%	11.22,1.84,100%	12.14,2.18,100%	12.41,2.15,100%	12.29,2.38,100%
15	10.53,1.49,100%	13.56,1.96,100%	15.27,2.34,100%	16.15,2.60,100%	16.65,2.82,100%
20	10.49,1.41,100%	12.98,1.92,100%	14.57,2.28,100%	15.77,2.57,100%	16.69,2.75,100%
25	13.76,1.93,100%	15.31,2.23,100%	17.24,2.79,100%	17.53,2.99,100%	17.97,3.08,100%
30	14.57,2.28,100%	15.77,2.57,100%	17.49,2.62,100%	17.47,2.84,100%	17.89,2.99,100%

Size=250	K				
d	10	15	20	25	30
10	11.10,1.29,100%	13.39,1.92,100%	14.83,2.24,100%	15.57,2.39,100%	15.77,2.41,100%
15	11.94,1.28,100%	15.13,1.73,100%	17.23,2.12,100%	18.52,2.35,100%	19.28,2.55,100%
20	11.85,1.23,100%	14.76,1.75,100%	16.72,2.14,100%	18.15,2.47,100%	19.26,2.69,100%
25	14.67,2.03,100%	17.27,2.14,100%	19.04,2.45,100%	20.13,2.54,100%	21.22,2.87,100%
30	17.42,2.12,100%	18.82,2.33,100%	21.05,2.65,100%	22.96,2.61,100%	23.21,2.85,100%

Size=300	K				
d	10	15	20	25	30
10	12.83,1.36,100%	15.45,1.94,100%	18.51,2.52,100%	19.13,2.69,100%	19.37,2.75,100%
15	13.84,1.32,100%	17.49,1.95,100%	20.18,2.31,100%	21.96,2.63,100%	23.12,2.81,100%
20	13.49,1.84,100%	17.65,2.11,100%	19.87,2.34,100%	22.40,2.66,100%	24.05,3.01,100%
25	18.03,2.12,100%	20.21,2.57,100%	22.35,3.05,100%	23.89,2.99,100%	25.21,3.35,100%
30	21.94,2.56,100%	23.34,2.65,100%	24.05,2.70,100%	26.40,3.56,100%	27.84,3.76,100%

Table 4.2: Simulation results for correlated case. Values in the table represent $\overline{z}$, s_z, % of $|z| > 1.96$ out of 200 random trials

4.5 Real world data study

Using the Welldoc data, we tested the correlation for two types of activities, 1) activities with both foods and medications (rapid insulin) intake at the same time, and 2) physical exercises. In the data processing step, we found the available sample sizes for tests were 286 for activities with both foods and medications (rapid insulin) intake, and 254 for physical exercises. Based on the data availability, two tests were conducted with $K = d = 30$ to ensure the reliability. Euclidean distance was applied to identify nearest neighbors.

Type of activity	Sample size	Test statistics	p-value	Conclusion
Foods+Medications	286	24.2857	<0.0001	Significant
Physical exercises	254	0.7681	0.4424	Insignificant

Table 4.3: Correlation test between activities and CGM

Table 4.3 presents the testing results including test statistics and corresponding p-values. The test showed that combination of foods and medications intake still significantly changed the glucose level after the activity. However, the physical exercise did not immediately influence the glucose trace significantly. This might be because the test was conducted without considering the lag time between the activity and change of glucose.

4.6 Discussion

In this chapter, we investigated relationship between real life activities and their corresponding influence of CGM series as the correlation between event and time series. We extended the KNN method, which is a non-parametric test for general two-sample multivariate testing, to medical streaming data mining problems for multiple subjects and validated the performance and reliability by simulation study.

To applied the KNN test, we relaxed the independence assumption in the general settings of two-sample multivariate test. However, the method successfully distinguished the correlated and uncorrelated cases in multiple simulation studies with different choices of sample sizes, number of nearest neighbors, and dimensions of series. It encourages the potential opportunities in application of more methods under the regular settings to medical time series data analysis with proper relaxing of some assumptions.

In the current study, we applied the kNN test to analyze the correlation. Besides the kNN test, many other non-parametric methods are proposed in literature and could be applicable in our study or similar data mining problems. While applying alternative methods, further simulation studies are needed to examine the effectiveness and power in a given scenario.

For future work, we can apply different two sample multivariate testing methods in the current scenario to verify and compare the performance of different tests. In addition, our current test only considered the immediate influence of limited types of activities on glucose change. In study [Box and Tiao, 1975], the author used autocorrelations to identify the possible values for lags. We can start with similar strategy to explore the lagged influence. Moreover, more types of activities can be incorporated in future study to better serve the design of therapy on lifestyle change and medication regimen for diabetic patients.

Chapter 5: Predicting Progression of Diabetes among Prediabetic Patients: Binary Outcome Prediction

5.1 Overview and Background

In the US, the number of people with type 2 diabetes mellitus (T2D) reached 30 million according to the National Diabetes Statistical Report 2020 [CDC, 2020a]. T2D can cause serious health problems including heart disease, blindness, and kidney disease [CDC, 2020c] as well as high health care costs [ADA, 2020]. In efforts to prevent diabetes, more than 1,500 CDC-recognized organizations offer different programs in T2D prevention, and costs of such programs are now covered by Medicare as well as approximately 40 commercial health plans [CDC, 2020a].

The management of prediabetes is a crucial step in curbing the growth of diabetes [ADA, 2021]. Prediabetes, in which glycemic levels are above normal but below the diabetes threshold, is a key stage where actions can be taken to delay or prevent T2D [CDC, 2020b, Tabák et al., 2012]. Studies show that there exists substantial variation in the risk of progression to diabetes from prediabetes [Tabák et al., 2012, DeJesus et al., 2017, Richter et al., 2018]. The considerable variation suggests that it is critical to assess risk of progression for appropriate intervention, thus, resources in diabetes prevention can be used more efficiently. More resources can be focused on high-risk patients to delay or prevent the progression [CDC, 2020b, Tabák et al., 2012, Richter et al., 2018], while avoiding unnecessary intervention, extra psychological and financial burdens for low-risk groups [Yudkin and Montori, 2014, Van den Bruel, 2015, Yudkin, 2016, Rooney et al., 2021, Echouffo-Tcheugui and Selvin, 2020].

There is limited research on personalized risk assessment to support and differentiate interventions for the prediabetes population. Most existing studies focus on predicting the risk of diabetes in the general population [Wilson et al., 2007, Schmidt et al., 2005, Nguyen et al., 2019, López et al., 2018], which may exhibit very different baseline characteristics compared to prediabetic population. An extensive literature search identified two studies on diabetes risk prediction for prediabetic population. Yokota et al. [Yokota et al., 2017] applied a multivariate logistic regression analysis to predict the development of diabetes among patients with prediabetes in Japan. In this study, the diagnostic method of prediabetes is the oral glucose tolerance test (OGTT), which is not pervasively available in the US except for pregnant women [Meijnikman et al., 2017, Bonora and Tuomilehto, 2011]. Glauber et al. [Glauber et al., 2018] proposed a simple calculation of two-year risk of diabetes based on population segmentation using hemoglobin A1c (HbA1c) and body mass index (BMI) for individuals with prediabetes in the US. The incidence rate in each subgroup is used as a risk measure and they did not consider other individual characteristics for prediction.

This study builds on these pioneering works to improve the risk predictions of diabetes for a US population with prediabetes. Motivated by the earlier papers, we set out to incorporate a broader range of risk factors in individualized risk predictions. Given the increasing availability of electronic health record (EHR) systems, our study extracted a wide range of factors, including demographic characteristics, lab measures, and behavior and social factors, from patients' EHRs. Leveraging the different types of factors, we aim to build an accurate and individualized risk prediction model and identify important factors associated with diabetes prognosis.

5.2 Material

The American Diabetes Association recommends that the diagnostic criteria of prediabetes be based on any of the following: HbA1c (5.7%–6.4%), fasting plasma glucose (100-125 mg/dl), or oral glucose tolerance test (plasma glucose 2 hours after a 75-g oral glucose load, 140-199 mg/dl) [ADA, 2021,Association et al., 2020]. Since HbA1c is most pervasively available in our population's EHR, the study population are individuals with HbA1c values between 5.7% and 6.4%. We initially identified 15,895 unique individuals based on EHR information consisting of all outpatient visits between 2012-2016, and extracted their EHR data.

Exclusion of previously diagnosed diabetic patients, including those with diagnosed Type-1 diabetes, was performed based on ICD-9 and ICD-10 diagnosis codes, and diabetes-related medications in patients' EHRs (A.2). Patients under age 18 were also excluded. Figure 5.1 summarizes the inclusion/exclusion process and the corresponding sample sizes for each step.

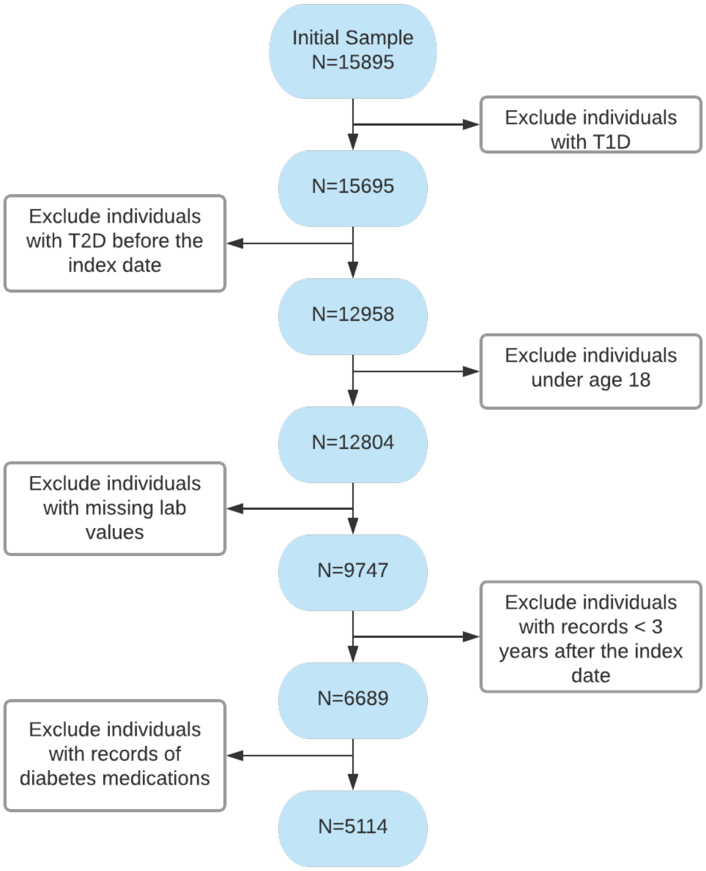

Initial Sample
N=15895
Exclude individuals with T1D
N=15695
Exclude individuals with T2D before the index date
N=12958
Exclude individuals under age 18
N=12804
Exclude individuals with missing lab values
N=9747
Exclude individuals with records < 3 years after the index date
N=6689
Exclude individuals with records of diabetes medications
N=5114

Figure 5.1: Patient inclusion/exclusion diagram

5.2.0.1 Definition of variables

The seventeen features used in the models were selected based on the literature on T2D risk prediction and glycemic control [Wilson et al., 2007,Schmidt et al., 2005, Nguyen et al., 2019, López et al., 2018, Yokota et al., 2017, Yeh et al., 2010, Schulze et al., 2007, Ley et al., 2014, Hodge et al., 2006, Christine et al., 2015, Strom and Egede, 2012], and their availability in the EHR data. We define the earliest date that a patient's HbA1c reading falls within 5.7%-6.4% as the index date. Demographic features are gender, age (on the index date), and race. Clinical factors are built based on information from three years prior to the index date, including body mass index (BMI), blood pressure (BP), high-density lipoproteins (HDL), low-density lipoproteins (LDL), Serum glutamic pyruvic transaminase (SGPT), and glycated hemoglobin (HbA1c, on the index date). Similarly, we extracted key social and behavioral determinants of health, specifically smoking status, exercise commitment, diet comment, alcohol use, caffeine use, living/housing status; and family history of T2D. If more than one value is available, the average value is used for clinical lab values, and the most frequent value for any behavioral and social factors.

Behavioral factors

We applied rule-based Nature Language Processing (NLP) methods to extract the behavioral features recorded in free-text format including smoking, exercise, diet, alcohol use, and caffeine use. For each behavior, we first identified relevant data fields with free-text descriptions of each behavior and developed rules to classify those text-based descriptions into predefined levels. Given our efforts to use readily available clinical information, these classification rules were based on an evaluation of the most common text descriptions for each field.

Social Determinants

The Moonstone system is an open-source, Java-based NLP system that uses linguistic and dictionary-based rules to infer social determinant information from free text [Conway et al., 2019]. The Moonstone system was used to extract information about social determinants from clinical notes: housing situation (homeless/marginal housing, lives at home/not homeless, lives in a facility), living alone (lives alone or does not live alone), and social support (social support or no social support).

Outcome variable

We define the outcome variable as the development of T2D at any point in the three years following the index date. The main diagnosis is according to T2D ICD-9 or ICD-10 diagnosis codes. Considering that only 72% – 74% patients meeting criteria of diabetes were formally diagnosed [Kazemian et al., 2019], we further included patients with at least two HbA1c measures exceeding 6.4% within one month as diabetic following ADA guidelines [Association et al., 2020].

5.3 Preliminary Study and Motivation

The objective of our study, predicting the short-term risk of progression to diabetes among adults with prediabetes, can be achieved as a classic binary classification problem. Many methods and models have been proposed and applied for classification in different areas. Some widely used and well performing models are Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), K-nearest Neighbor (kNN), and Artificial neural network (ANN).

For a preliminary study, we first applied the above mentioned models using risk factors defined in previous section, and compared their performance with two baseline models. The first baseline model, the Framingham model [DeJesus et al., 2017], was initially developed to predict risk of diabetes among the general population using a multivariate logistic regression on categorized variables. Considering the different target populations in two studies, we adopted their approach by training the model on predictors provided in their study using our training data. The second baseline model was developed to predict two-year diabetes risk for a prediabetic population [Glauber et al., 2018]. The approach uses a simple risk index based on HbA1c and BMI. We adapt the approach to our population by bucketing the HbA1c and BMI values in the training data using the ranges suggested in their study, and computed the risk for each subgroup by the proposed formula and the subgroup incidence rate in the training set.

All models were trained and tested separately. The sample is split into training (75%) and testing (25%) set for model development and evaluation During the model development using training data, 10-fold cross validation is used to optimize the necessary tuning parameters if needed. The performance of models is measured by Area under the ROC curve (AUC), which represents the overall ability of discrimination between individuals who develop diabetes and those who do not [Bradley, 1997]. An empirical 95% confidence interval of each AUC is calculated using stratified bootstrap with 2000 replicates, which ensures equal incidence rates in all replicates.

The performance of all models are reported in Table 5.1 by AUC and empirical 95% confidence intervals. As it is showing, all predictive models are outperformed the two baseline models. The AUC's vary between 0.718 and 0.77, which are fairly accurate. Observing these models, the best performing model, random forest, has the highest AUC, but it is difficult to interpret and address the risk factors. Logistic model has high interpretability and it performed similarly as the RF model. Next,

Model	AUC	(95% CI)
LR	0.768	(0.739, 0.799)
KNN	0.718	(0.691, 0.743)
ANN	0.733	(0.705, 0.761)
SVM	0.764	(0.740, 0.791)
RF	0.770	(0.745, 0.800)
Framingham	0.574	(0.539, 0.609)
Glauber	0.554	(0.517, 0.591)

Table 5.1: AUC of models applied in preliminary study

we aim to develop alternative models to improve the performance and maintain good interpretability.

5.4 Method

One way to cope with difficulties in balancing performance and interpretability is to develop a nonlinear model composing of a number of sub-models which are interpretable and responsible for respective sub-domains. The idea ofa multi-model approach [Binder et al., 1981] was proposed as a multivariable control methodology using several models representing a physical system in different operating points. This algorithm is characterized by a parallel structure, which allows flexibility for different sub-domains. Later, based on the the concept of fuzzy sets theory [Zadeh et al., 1996], the idea of fuzzy modeling [Takagi and Sugeno, 1985] was developed and provided new techniques to build multi-models.

5.4.0.1 Fuzzy modeling

In contrast with traditional set theory, fuzzy set theory provides a new way to view the relationship between objects and sets with different degrees of membership [Zadeh et al., 1996]. Let $X = \{x\}$ be a space of objects with a generic element x. A fuzzy set A in X is characterized by a membership function $f_A(x)$ which associates with each point in X a real number in the interval [0, 1]. The value of $f_A(x)$ at x represents the degree of membership of x in A. Thus, the closer the value of $f_A(x)$ to 1, the higher the degree of membership of x in A. The original set theory can be viewed as a special case of fuzzy set thery where the possible values of $f_A(x)$ are either 0 or 1.

Cooperating with fuzzy set theory, fuzzy modeling attempts to take the uncertainty into account while building the models. In this study, the main framework is based on Takagi–Sugeno (TS) fuzzy modeling [Takagi and Sugeno, 1985]. The TS fuzzy model is described by fuzzy IF-THEN rules which represents local input-output relations of a nonlinear system. The main feature of a TS fuzzy model is to express the local dynamics of each fuzzy set by a linear model. The overall fuzzy model of the system is achieved by fuzzy mixture of the linear models. Many studies have applied TS fuzzy modeling to medical topics for classification and prediction [Viegas et al., 2017, Curto et al., 2016, Fernandes et al., 2014]. In particular, the final output in TS fuzzy model is a weighted average of individual local outputs:

$$y = \frac{\sum_{i=1}^{c} \beta_i y_i}{\sum_{i=1}^{c} \beta_i} = \frac{\sum_{i=1}^{c} \beta_i (a_i' x + b_i)}{\sum_{i=1}^{c} \beta_i}$$

where the degree of activation β_i for the ith rule is calculated by

$$\beta_i = \mu_{A_i}(x) \quad \text{where} \ \ \mu_{A_i}(x) : \mathcal{X} \to [0,1]$$

The consequent parameters a_i for each rule could be obtained as an weighted ordinary least square estimate in general settings. To determine the antecedent fuzzy sets, we applied a fuzzy clustering algorithms, fuzzy c–means (FCM) [Bezdek et al., 1984].

FCM is used to decompose a given set of objects into clusters based on similarity via a cost function, c-means function. Let C be the number of clusters and Q be the number of input variables, then each cluster is characterized by the center or prototype ν_i denoted as

$$\nu_i = (\nu_1, ..., \nu_Q) \quad \text{where} \ \ i = 1, 2, ..., C$$

The prototypes ν_i can be computed by

$$\nu_i = \frac{\sum_{n=1}^{N} \mu_{in}^m x_n}{\sum_{n=1}^{N} \mu_{in}^m}$$

where $m \in [1, \infty)$ is the weighting exponent that determines the degree of fuzziness in the clustering procedure. The degree of membership for object n in cluster i $\mu_{in} \in [0,1]$ is defined as

$$\mu_{in} = \left[\sum_{k=1}^{C} (\frac{d_{in}}{d_{kn}})^{\frac{2}{m-1}} \right]^{-1}$$

for each sample x_n, $n = 1, 2, \ldots, N$. $\mu_{in} = 0$ implies sample n does not belong to cluster i, and $\mu_{in} = 1$ implies sample n belongs to cluster i with absolute certainty. Assigning observations in sample data into C clusters is achieved by minimizing the following objective function

$$L(X, U, V) = \sum_{i=1}^{C} \sum_{n=1}^{N} \mu_{in}^{m} d_{in}^{2}$$

where d_{in} is a distance measure, such as Eucledian distance, Mahalanobis distance, etc.

In the ordinary TS fuzzy modeling, the input variables used to determine the antecedent fuzzy sets and the local linear models are normally the same. However, introducing more flexibility to the choice of variables for the local linear models can potentially lead to more insights on different risk factors in subgroups. We will apply more flexible choice of variables as input to fit the local linear models.

5.4.0.2 Variable selection

Variable or feature selection is a key step which can help identify and extract the most relevant factors for a given classification task [Dash and Liu, 1997]. In this framework, the term relevant refers to the influence of a given variable on the possible error in the classification problem.

Guyon et al. did a comprhansive review, and summarized variable selection methods into univariate and multivariate methods [Guyon et al., 2008]. The univariate method considers one variable at a time, and the multivariate method considers a subset of variables. The disadvantage of the univariate method is that the conditional dependencies among different subsets of variables significantly affects the performance of the method. It has been pointed out that combining individually selected best variables does not imply a best subset of variables [Cover, 1999]. Therefore, a multivariate method is considered as better approach.

The most popular multivariate variable selection methods apply forward, backward or floating sequential (stepwise) schemes. Forward or backward selection method usually add or remove one variable at each step conditioning on the selected variables. The forward and backward methods form a nested ranking of variables, and the selection is fixed once a variable is added (or removed), which often results in a sub-optimal selection [Cover, 1974, Jain et al., 2000]. Floating search attempts to overcome the nested problem in forward and backward methods by allowing adding back the removed variables or excluding the added variables. However, the computational cost easily increases to evaluate all the possible subsets of features, and the selection set is not guaranteed to be optimal [Jain et al., 2000].

In addition to these methods, exhaustive search approach guarantees to find the optimal subset. However, it is computationally expensive due to the great number of possible subsets. Considering the total number of variables is not over large in our study, we decided to access the optimal subset via an exhaustive searching strategy based on a non-parametric multivariate two-sample test, KNN test [Schilling, 1986, Henze, 1988].

Suppose $X = \{x_1, x_2, x_3, ..., x_n\}$ is the full set of predictor variables, Y is the response variable, and X_s^d is an arbitrary subset of X with size d. The main idea of our selection approach is that the selected subset X_s^d is significantly associated with Y if

$$F(X_s^d|Y=1) \neq F(X_s^d|Y=0)$$

As discussed in Chapter 4, multivariate two-sample tests are used to check whether two samples come from the same underlying distribution, which is assumed to be unknown. In our context, the two samples are $X_s^d|Y=1$ and $X_s^d|Y=0$, and the hypothesis in the test are

$$H_0 : F(X_s^d|Y=1) = F(X_s^d|Y=0) \qquad H_a : F(X_s^d|Y=1) \neq F(X_s^d|Y=0)$$

If H_0 is rejected, it means two distributions are statistically different, and X_s^d and Y are significantly associated. Otherwise, X_s^d and Y are not associated.

Similar to Chapter 3, the pooled sample is $\mathbf{X_s^d} = \mathbf{X_s^d}|\mathbf{Y}=\mathbf{0} \cup \mathbf{X_s^d}|\mathbf{Y}=\mathbf{1}$ with $N = N_0 + N_1$, where N_i is the size of $Y = i$. Denote the index of the kth nearest neighbor of ith observation as $NN_i(k)$ based on a certain distance measure. Then, the indicator

$$I_i(k) = \begin{cases} 1, & \text{if } Y_{NN_i(k)} = Y_i \\ 0, & \text{otherwise} \end{cases}$$

Then, the test statistic for the hypothesis is defined in the same way

$$T_{K,N} = \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K} I_i(k) \tag{5.1}$$

Given the null, the asymptotic distribution of $Z_K = (NK)^{1/2}(T_{K,N} - \mu_K)/\sigma_K$ follows standard normal with mean and variance as

$$\begin{aligned} \mu_K &= (\lambda_0)^2 + (\lambda_1)^2 \\ \sigma_K^2 &= \lambda_0\lambda_1 + 4\lambda_0^2\lambda_1^2 \end{aligned}$$

where $N_0/N = \lambda_0$, $N_1/N = \lambda_1$, and K, d are sufficiently large. Study [Schilling, 1986] suggested values of σ_K^2 for small K and d. For a given significance level α, large value of Z_K shows high confidence of hypotheses H_a.

Based on the KNN two sample test, our approach selects the subset X_s^d which maximizes Z_K for given K. Algorithm 2 summarizes the selection procedure. Different values of k may lead to different selection sets. Then, the final selection could be determined by plurality vote using multiple k values. We will validate and test the performance of the proposed method through simulation studies before applying it to the real data analysis.

5.4.1 Model assessment

As in the preliminary study, our sample is split into training (75%) and testing (25%) set for model development and evaluation. Model performance is measured by AUC and an emprical 95% confidence interval. We also report additional metrics, accuracy, sensitivity, specificity, and precision using a classifying threshold which matches the actual incidence rate for additional insights. To explore more possibilities, the compared models include 1) ordinary fuzzy modeling, which we applied the fuzzy models using selected variables to decide the fuzzy sets and fit local linear

Algorithm 2: An exhaustive searching algorithm based on multivariate two-sample kNN test

```
Input:
Predictors $X = \{x_1, x_2, x_3, ..., x_n\}$;
Response $Y$;
Number of nearest neighbors $k$;
Output:
Selection set $X_s \subset X$;
for ( $i = 1, ..., n$ ) {
    A. Decide all possible subsets with size $i$ by $\binom{n}{i}$;
    for ( Each subset $X_s'$ ) {
        1. Apply the kNN test for k=k;
        2. Record the value of $Z_k$;
    }
}
Return the subset with $\max Z_k$ as $X_s$ ;
Algorithm End.
```

models; 2) flexible fuzzy modeling, which we used selected variables to access the fuzzy sets, and the local linear models were fitted based on all variables with a local selection.

5.5 Simulation study

To test the performance of the variable selection methods, we designed two simulation studies under different settings.

In the first case, we created 6 significant variables and 30 noise (random) variables. Among 6 important variables, 4 variables are linearly associated with the response, and the other 2 are nonlinearly associated. Three of the four variables with linear association are set to be mutually correlated, and the two variables with nonlinear effects are also set as correlated. Noise variables are sampled randomly from multiple continuous distributions, such as Uniform, normal, and exponential, with different parameters. In particular, we draw the sample data from

$$Y \sim Bernoulli(p)$$

$$\log \frac{p}{1-p} = \beta_0 + \sum_{i=1}^{4} \beta_i x_i + \beta_5 (x_5 + a)^2 + \beta_6 \sin(x_6^2) + e$$

We set X_1, X_2 and X_3 following multivariate normal with arbitrary means and variances and the correlations greater than 0.8; X_5 and X_6 following multivariate normal with arbitrary means and variances and the correlation greater than 0.8; X_4 following an arbitrary continuous distribution. Noise variables were added later.

For the second case, we generated datasets which the response variable is significantly related with two variables simultaneously but not related to the single variables separately. We also added two extra variables which were moderately correlated with the response. Noise variables are created by the same methods in case one. We first sampled Y, X_1 and X_2 by setting X_1 and X_2 following independent and identical normal, and assigning

$$y = \begin{cases} 1 & \text{if} \quad y > x \quad \text{and} \quad y > -x \\ 1 & \text{if} \quad y < x \quad \text{and} \quad y < -x \\ 0 & \quad\text{otherwise} \end{cases}$$

Next, Y', X_3 and X_4 can be sampled similarly as case 1 through

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

and set σ_e^2 large such that the linear relationship is moderate. Then, we merged two by value of Y and Y'.

Table 5.2 summaries the simulation designs and the expected target variables in each design.

	Design description	Target

Design 1	X1–X3: linear, mutually correlated
	X5,X6: nonlinear, correlated
	X4: linear; X7–X36: noise
		X1 or X2 or X3;
		X5 or X6;
		X4

Design 2	X1∪ X2: strong, non-linear X3,X4: linear, moderate X5-X34: noise	X1,X2

Table 5.2: Simulation design for variable selection. Data size in both designs n=1000

To apply the Exhaustive kNN method, we decided to use $k \in \{5, ..., 14\}$ and the final selected variables are the most frequent subset across different k values. Table 5.3 presents the selected subset for different k values in both designs in a random trial.

k	Design 1	Design 2
5	x1,x4,x5	x1,x2,noise13
6	x1,x4,x5	x1,x2
7	x1,x4,x5,noise8	x1,x2
8	x1,x4,x5,noise8	x1,x2
9	x1,x4,x5,noise8	x1,x2
10	x1,x4,x5	x1,x2
11	x1,x4,x5	x1,x2
12	x1,x4,x5	x1,x2
13	x1,x4,x5,noise3	x1,x2
14	x1,x4,x5,noise13	x1,x2,noise8
Final selection	x1,x4,x5	x1,x2

Table 5.3: Exhaustive kNN selection results for different k in simulation studies

To compare the performance, we applied another five widely used variable selection methods, Random Forest (RF), XGboost, LASSO, and stepwise selection methods based on AIC and BIC via logistic regression. For RF and XGboost methods, we selected the variables based on their importance rank, and the number of variables equals the number of target variables. Table 5.4 shows the selection results using different methods in 200 random trials. We reported two criteria to measure and compare the performance of different selection methods, percentage of selection results that include all expected variables (% C1), and percentage of selection results that include unexpected variables (% C2).

As shown in Table 5.4, in Design 1, Exhaustive kNN, RF, and XGboost methods selected the correct subset of variables in most trials, which presented the ability of handling both linear and nonlinear effects. However, LASSO, AIC and BIC can only identify variables with linear associaitons. AIC and BIC methods were very likely to include unnecessary variables. In addition, Exhaustive kNN, Random Forest, XGboost, and LASSO methods all tended to select one variable when there exits multiple correlated variables, while AIC and BIC are more likely to keep all. In design 2, only the Exhaustive kNN method selected the expected variables in most trials. That is, when the decision boundary to differentiate the response labels depends on two or more variables simultaneously, the exhaustive method outperformed all other baseline methods.

Method	Design 1		Design 2	
	% C1	% C2	% C1	% C2
Exhaustive kNN	94.5%	6.0%	95.5%	8.5%
RF	95.5%	4.5%	0%	100%
XGboost	94.0%	6.0%	0%	100%
LASSO	1.0%	4.5%	0%	100%
AIC	62.5%	100%	1.5%	100%
BIC	3.5%	100%	0%	100%

Table 5.4: Variable selection results in simulation studies

5.6 Real world data analysis

5.6.1 Population Characteristics

Our study population of 5,114 patients are majority female (61.9%) and are racially diverse with 44.4% of the patients being White and 46.0% Black. The median age at the index visit is 58 years old, with an interquartile range (IQR) of 16 years. Among them, 1,341 patients (26.2%) developed T2D in the three-year follow-up period. Detailed baseline characteristics of all patients are reported in Table A.3. Categorical variables are summarized by counts (proportion in percentage) of each level, and numerical variables are presented by median (IQR) due to the skewness of their distribution in our sample.

5.6.2 Variable selection

We applied the Exhaustive kNN methods on the training sample for $k \in \{5, 6, ..., 14\}$, and reported the selection result for different k values in Table 5.5. To identify the nearest neighbors, the distance measure used for 1) numeric variables, was normalized Euclidean distance, and 2) categorical variables, 0 if the attribute is of the same class, and 1 otherwise. The final distance between two observations (x and y) is calculated as $d(x, y) = \sqrt{\sum_{j=1}^{m} d^2(x_j, y_j)}$ where j is the index of predictors.

k	Selected variables
5	HbA1C, LDL
6	HbA1C, HDL
7	HbA1C, HDL
8	HbA1C, HDL
9	HbA1C, HDL
10	HbA1C, HDL
11	HbA1C, HDL
12	HbA1C, Hypertension
13	HbA1C, Hypertension
14	HbA1C, HDL
15	HbA1C, HDL
Final selection	HbA1C, HDL

Table 5.5: Exhaustive kNN variable selection

Voting by outputs from different k values, the Exhaustive kNN method selected variables HbA1C and HDL as optimal subset of predictors.

5.6.3 Fuzzy models

Using HbA1C and HDL as inputs, we applied the Fuzzy C-means for clustering step. We decided the number of clusters based on *sum of squared within-cluster distances*, which is the objective function introduced in Section 4.4.0.1, through an *elbow plot* Figure 5.2.

After examining the elbow plot, we chose to apply the ordinary and flexible fuzzy modeling with seven fuzzy clusters. After deciding the fuzzy clusters, local models were fitted by weighted linear regression, where the weights were the mem-

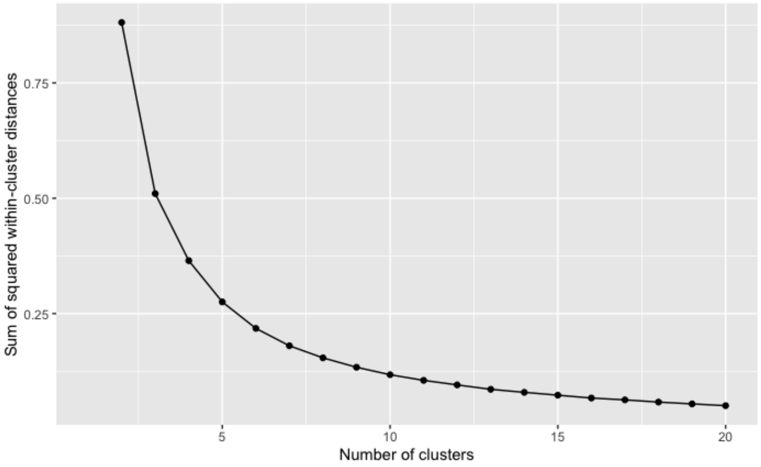


Figure 5.2: Elbow plot for fuzzy C-means clustering

bership degree in fuzzy clustering step. Table 5.6 presents the center for each fuzzy cluster, and the corresponding significant predictors for two approaches. In ordinary fuzzy model, HbA1c and HDL were significant in all fuzzy sets. However, in flexible fuzzy model, significant predictors varied in different fuzzy clusters. AUC and 95% CI are also reported in Table 5.6. We noticed that the flexible fuzzy model performed well with AUC 0.789, which was higher than all models in the preliminary study, and the ordinary fuzzy modeling. To test whether the AUC of flexible fuzzy model is significantly different compared to the AUC of the ordinary model, we applied the DeLong's test for two correlated ROC curves [DeLong et al., 1988, Sun and Xu, 2014], which reported a p-value 0.0103 (test statistics $z = 2.565$).

Fuzzy clusters		Ordinary Fuzzy modeling	Flexible Fuzzy modeling
Cluster No.	Center (HbA1c, HDL)	Significant factor (sign of coefficient)	Significant factor (sign of coefficient)
1	5.7, 24.7	HbA1c(+),HDL(-)	HbA1c(+),HDL(-),BMI(+), Hypertension(+), Exercise:Exercise(-)
2	6.2, 30.8	HbA1c(+),HDL(-)	HbA1c(+),SGPT(+),Age(+), BMI(+), Exercise:Exercise(-),Housing:Unknown(-)
3	5.7, 4.1	HbA1c(+),HDL(-)	HbA1c(+),HDL(-),LDL(-), BMI(+),Age(+),Hypertension(+)
4	6.0, 17.2	HbA1c(+),HDL(-)	HbA1c(+),HDL(-),SGPT(+), BMI(+),Hypertension(+)
5	6.3, 7.4	HbA1c(+),HDL(-)	HbA1c(+),HDL(-),SGPT(+), BMI(+),Exercise:Exercise(-),Gender:M(-)
6	6.0, 2.6	HbA1c(+),HDL(-)	HbA1c(+),HDL(-),SGPT(+), BMI(+), Hypertension(+),Gender:M(-), Housing:Unknown(-)
7	5.8, 47.2	HbA1c(+),HDL(-)	HbA1c(+),HDL(-),SGPT(+), Age(+),BMI(+),Exercise:Exercise(-)
Performance AUC (95% CI)		0.765 (0.736,0.794)	0.789 (0.762,0.816)

Table 5.6: Summary of Ordinary and Flexible Fuzzy modeling and their performance

5.7 Discussion

In this chapter, we developed an accurate diabetes risk prediction model for patients with prediabetes based on commonly available information in EHR systems. Our model's performance is compared with two baseline models and other widely used classification models. One of our models, the flexible fuzzy model, outperformed all other methods in our study. We also noted that two baseline models worked poorly using our sample, which highlights the importance of personalized modeling specific to the prediabetic population.

In the proposed variable selection method, we incorporated the exhaustive search strategy with a selection criterion based on a nonparametric multivariate two-sample test. We applied the kNN method in our current study, but it worth to note that other nonparametric tests may be applied with careful validation. One disadvantage for exhaustive search is the high computational cost. It is proper in our study because the number of variables in our sample is relatively small. However, in the high dimension case, we could extend the searching strategy to stepwise or floating approach. Further studies are need to examine the performance and generalizability of our method and possible extensions.

Of note, our variable selection method finally selected a subset containing only two variables. The method performed well in simulation study. In addition, the ordinary fuzzy model worked comparably better than most models in the preliminary study by only using the selected variables. This highlighted that the applied variable selection method selected the most influential variables.

In addition to the ordinary fuzzy modeling, we relaxed the inputs variables for local models in the flexible fuzzy model. Different local models presented variations on the significant predictor. It may imply that, within prediabetic population, patients in different subgroups are sensitive to very different factors, and thus, need specific types of interventions to improve the efficiency and control the financial and psychological burdens. The flexible fuzzy model has higher AUC compared to the ordinary fuzzy model. One potential problem in the flexible fuzzy model is that all risk factors are included in multiple local linear models, and it may cause overfitting issues. For future work, we will work on examining and solving the overfitting problem.

Chapter 6: Predicting Progression of Diabetes among Prediabetic Patients: HbA1c Prediction

6.1 Overview

In the care for diabetes, glycaemic control is one of the most important goals of clinical management [Assessment, 2021]. Many studies showed that glycaemic control is closely associated with risk of diabetes-related complications and fatal outcomes [Patel et al., 2008]. Many studies focused on HbA1c prediction or risk factor investigations on cross-section data with a specific follow-up period [Rauh et al., 2017, Chiu and Wray, 2010, Yazidi et al., 2016]. Rauh *et al.* [Rauh et al., 2017] developed a model based on linear regression to predict the HbA1c measure using a six-year follow-up data. However, these studies may omit the opportunities of capturing the information available in a longitudinal view of HbA1c measurements. Existing studies suggested that HbA1c trajectories are heterogeneous among patients with diabetes [Luo et al., 2018, Gebregziabher et al., 2010]. In particular, different patterns or shapes of HbA1c trajectories over time were identified for patients with the same baseline HbA1c.

Among the reviewed studies focused on longitudinal change of HbA1c trajectories, many studies concentrated on identifying the HbA1c trajectory types based on grouping methods. In a recent review study [Luo et al., 2018], the authors reviewed and summarized over twenty studies that analyzed HbA1c trajectories based on grouping methods. These methods included latent class growth analysis, latent class growth mixture modeling, hierarchical and multi-stage clustering analysis, etc. Through group-based studies, most studies identified 4 to 5 types of HbA1c trajectories using different predictors for later classification or prediction.

Beyond the group-based methods, some studies predicted HbA1c trajectories through mixed effect models. Ngufor *et al.* in their study [Ngufor et al., 2019] developed a derived mixed-effect machine learning model to predict the longitudinal change in HbA1c trajectories, which is an extension of generlized mixed-effect model (GLMM) by integrating the random-effects structure of GLMM into non-linear machine learning models. Their response was then transformed to binary outcomes as final prediction.

Based on the reviewed studies, our study aims to model the change of HbA1c trajectories by non-group-based methods and predict the individualized HbA1c trajectories. We developed a model based on the well known state-space modeling in time series analysis to estimate and predict HbA1c by introducing the predictor to the change of HbA1c.

6.2 A state-space model: dynamic linear model

In general, state-space models treat the time series Y_t as an incomplete and noisy function of some unobserved hidden process θ_t (the state) [Petris et al., 2009]. A general *dynamic linear model*, also known as Gaussian state-space model, can be characterized by a set of equations:

$$y_t = F_t\theta_t + e_t$$

$$\theta_t = G_t\theta_{t-1} + w_t$$

where $e_t \sim N(0, \Sigma_t)$, $w_t \sim N(0, W_t)$ and e_t, w_t are independent. The first equation is called the *observation equation*, and the second is called *state equation.*

In our particular case, we model the system as

$$\begin{aligned} Y_t &\sim N(\mu_t, \sigma_0^2) \\ \mu_t &= \mu_{t-1} + \phi_{t-1} + e_{1t} \qquad e_{1t} \sim N(0, \sigma_1^2) \\ \phi_t &= \lambda\phi_{t-1} + X_t\beta^T + e_{2t} \qquad e_{2t} \sim N(0, \sigma_2^2) \end{aligned} \tag{6.1}$$

where Y_t is the response variable and X_t is the predictor vector. Here, we assume discrete time. Then the state and parameter set in our system are

$$\theta_t = (\mu_t, \phi_t)$$

$$\Psi = (\lambda, \beta, \sigma_1^2, \sigma_2^2)$$

There are two assumptions in a state-space model:

A1. the process of $\theta_t, t = 0, 1, 2, ...$ is a Markov chain

A2. $Y_t|\theta_t$ are independent and Y_t only depends on θ_t.

By A1, the process of $\theta_t, t = 0, 1, 2, ...$ is Markovian so that one can characterize the hidden process by an initial distribution $p(\theta_0)$ and a transition probability $p(\theta_t|\theta_{t-1})$. By A2, the joint conditional distribution for $Y_t|\theta_t$'s can be written as

$$f(Y_1, ..., Y_n|\theta_1, ..., \theta_n) = \prod_{t=1}^{n} f(y_t|\theta_t)$$

Then, with the estimated states, the log-likelihood can be written as

$$-2 \log L(Y_1, ..., Y_n|\Psi) = C + \sum_t (y_t - \hat{y}_t)^T \Sigma_{y,t}^{-1} (y_t - \hat{y}_t) + \log(|\Sigma_{y,t}|)$$

The parameter Ψ is involved in modeling the $\hat{y}_t = F_t\hat{\theta}_t$ and $\Sigma_{y,t}$ for all $t > 0$, and thus, it is challenging to estimate parameters directly by maximizing the likelihood function.

One alternative way to estimate of the parameters is via the Expectation-Maximization (EM) algorithm. Briefly, the EM algorithm includes two steps, E and M steps. In the E-step, we estimate the process of the latent state using *Forward Filtering* (FF). In this method, we estimate the latent state θ_t using the values of the observed input X_t and the state value of the previous time step $\hat{\theta}_{t-1}$. One variant of this method is to apply *Forward Filtering Backward Smoothing* (FFBS) which uses the estimated state values of future time steps to correct the estimated latent state backwards after applying the FF method. In the M-step, we estimate the values of the parameters Ψ that maximize the likelihood of having the latent patient processs previously estimated in the E-step. Then, we repeat the E and the M steps until convergence.

6.2.1 E-Step: Forward Filtering

The Forward Filtering is solved by the well-known Kalman filter. Let D_t denote the information until time t, and let $D_t = D_{t-1} \bigcup \{y_t\}$. Then, in general, we have:

$$\begin{aligned}
p(\theta_t|D_{t-1}) &= \int p(\theta_t, \theta_{t-1}|D_{t-1})\, d\theta_{t-1} \\
&= \int p(\theta_t|\theta_{t-1})p(\theta_{t-1}|D_{t-1})\, d\theta_{t-1} \\
p(y_t|D_{t-1}) &= \int p(y_t, \theta_t|D_{t-1})\, d\theta_t \\
&= \int p(y_t|\theta_t)p(\theta_t|D_{t-1})\, d\theta_t
\end{aligned}$$

and finally by Bayes rule

$$p(\theta_t|D_t) \propto p(y_t|\theta_t)p(\theta_t|D_{t-1}).$$

The computation of the involved conditional probabilities is very challenging in general cases. However, by assuming all error terms follow normal distributions, the marginal and conditional distributions of (θ_t, Y_t) are normal for all $t > 0$. Then, the process of the latent state can be estimated through the posterior distribution. To simplify the notation, we rewrite our model equation in the format of the general observation equation and state equation where

$$\theta_t = \begin{pmatrix} \mu_t \\ \phi_t \end{pmatrix}, \qquad F_t = \begin{pmatrix} 1 & 0 \end{pmatrix}, \quad e_t \sim N(0, V = \sigma_0^2)$$

$$G_t = \begin{pmatrix} 1 & 1 \\ 0 & \lambda \end{pmatrix}, \quad w_t = \begin{pmatrix} e_t^1 \\ X_t\beta^T + e_t^2 \end{pmatrix}, \quad W_t = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

Assume $\theta_t|D_t \sim N(\pi_t, \Lambda_t)$, and $D_t = y_1, ..., y_t$. By $\theta_t = G_t\theta_{t-1} + w_t$, we have

$$\begin{aligned} E(\theta_t|D_{t-1}) &= G_tE(\theta_{t-1}|D_{t-1}) + E(w_t) = G_t\pi_{t-1} + w_t \\ Var(\theta_t|D_{t-1}) &= G_tVar(\theta_{t-1}|D_{t-1})G_t^T + W = G_t\Lambda_{t-1}G_t^T + W \end{aligned}$$

By $y_t = F_t\theta_t + e_t$, we have

$$\begin{aligned} E(y_t|D_{t-1}) &= F_tE(\theta_t|D_{t-1}) = F_t(G_t\pi_{t-1} + w_t) \\ Var(y_t|D_{t-1}) &= F_tVar(\theta_t|D_{t-1})F_t^T + V = F_t(G_t\Lambda_{t-1}G_t^T + W)F_t^T + V \end{aligned}$$

Thus,

$$\begin{aligned} \theta_t|D_{t-1} &\sim N(G_t\pi_{t-1} + w_t, G_t\Lambda_{t-1}G_t^T + W) \\ y_t|D_{t-1} &\sim N(F_t(G_t\pi_{t-1} + w_t), F_t(G_t\Lambda_{t-1}G_t^T + W)F_t^T + V) \end{aligned}$$

and they are jointly normal with mean and variance

$$\begin{pmatrix} G_t\pi_{t-1} + w_t \\ F_t(G_t\pi_{t-1} + w_t) \end{pmatrix}, \begin{pmatrix} G_t\Lambda_{t-1}G_t^T + W & (G_t\Lambda_{t-1}G_t^T + W)F_t^T \\ F_t(G_t\Lambda_{t-1}G_t^T + W) & F_t(G_t\Lambda_{t-1}G_t^T + W)F_t^T + V \end{pmatrix}$$

Then the conditional distribution $P((\theta_t|D_{t-1}) \mid (y_t|D_{t-1})) = P(\theta_t|Y_t)$ can be directly derived based on the joint normal properties. That is, a normal distribution with mean and variance

$$\begin{aligned} E(\theta_t|Y_t) &= G_t\pi_{t-1} + w_t \\ &+ \Sigma_t F_t^T(F_t\Sigma_t F_t^T + V)^{-1}(y_t - F_t(G_t\pi_{t-1} + w_t)) \\ Var(\theta_t|Y_t) &= \Sigma_t + W \\ &- \Sigma_t F_t^T(F_t\Sigma_t F_t^T + V)^{-1}F_t\Sigma_t \\ \text{where } \Sigma_t &= G_t\Lambda_{t-1}G_t^T + W \end{aligned}$$

Therefore the state at t is updated based on the distribution at $t-1$.

6.2.2 M-Step: Parameter Estimate

In the E-step, we have the process of the latent state estimated. Now, we estimate the parameters that maximize the likelihood of the estimated process.

Recall the state equation in our scenario in Equation 6.1,

$$\begin{aligned}\mu_t &= \mu_{t-1} + \phi_{t-1} + e_{1t} \qquad e_{1t} \sim N(0, \sigma_1^2) \\ \phi_t &= \lambda\phi_{t-1} + X_t\beta^T + e_{2t} \qquad e_{2t} \sim N(0, \sigma_2^2)\end{aligned}$$

Then, the overall log-likelihood function for all observations can be written as following:

$$\begin{aligned}\ell(\mu_{i,t}, \sigma_1^2) &= C - \frac{1}{2}\sum_i\sum_t \log\sigma_1^2 - \frac{1}{2}(\sigma_1^2)^{-1}\sum_i\sum_t(\mu_{i,t} - \mu_{i,t-1} - \phi_{i,t-1})^2 \\ \ell(\phi_{i,t}, \lambda, \beta, \sigma_2^2) &= C' - \frac{1}{2}\sum_i\sum_t \log\sigma_2^2 - \frac{1}{2}(\sigma_2^2)^{-1}\sum_i\sum_t(\phi_{i,t} - \lambda\phi_{i,t-1} - X_{i,t}\beta^T))^2\end{aligned}$$

and set up equations to solve for MLE of multiple parameters:

$$\begin{aligned}\frac{\partial}{\partial\sigma_1^2}\ell(\mu_{i,t}, \sigma_1^2) &= -\frac{1}{2}\sum_i\sum_t(\sigma_1^2)^{-1} - \frac{1}{2}(\sigma_1^2)^{-2}\sum_i\sum_t(\mu_{i,t} - \mu_{i,t-1} - \phi_{i,t-1})^2 = 0 \\ \frac{\partial}{\partial\sigma_1^2}\ell(\phi_{i,t}, \lambda, \beta, \sigma_2^2) &= -\frac{1}{2}\sum_i\sum_t(\sigma_2^2)^{-1} - \frac{1}{2}(\sigma_2^2)^{-2}\sum_i\sum_t(\phi_{i,t} - \lambda\phi_{i,t-1} - X_{i,t}\beta^T)^2 = 0\end{aligned}$$

Let $B = (\lambda \;\; \beta)$, $Z = (\phi_{i,t-1} \;\; X_{i,t})_{i,t}$, and $\phi = (\phi_{i,t})_{i,t}$

$$\frac{\partial}{\partial B}\ell(\phi_{i,t}, \lambda, \beta, \sigma_2^2) = -\frac{1}{2}(\sigma_2^2)^{-1}(-2Z^T\phi + 2Z^TZB^T) = 0$$

Then, the MLE's are

$$
\begin{aligned}
\hat{\sigma}_1^2 &= \sum_i \sum_t (\mu_{i,t} - \mu_{i,t-1} - \phi_{i,t-1})^2 \\
\hat{\sigma}_2^2 &= \sum_i \sum_t (\phi_{i,t} - \lambda\phi_{i,t-1} - X_{i,t}\beta^T)^2 \\
\hat{B} &= (Z^T Z)^{-1} Z^T \phi
\end{aligned}
$$

6.2.3 Predicting the response

After the parameters are estimated, the response variable y_t can be predicted based on the forward filter through

$$
\begin{aligned}
\hat{y}_t|D_{t-1} &= F_t^T E(\theta_t|D_{t-1}) + e_t \\
E(\hat{y}_t|D_{t-1}) &= E(\mu_t|D_{t-1}) \qquad (6.2)
\end{aligned}
$$

Besides a typical one-step-ahead prediction, by assuming the baseline condition of predictors X will not change, we also can make a k-step-ahead prediction for small k by

$$\mu_{t+k} = \mu_t + X_t\beta^T + \frac{1-\lambda^k}{1-\lambda}\theta_t + \frac{1-\lambda^{k-1}}{1-\lambda}X_t\beta^T$$

6.3 Real world data study

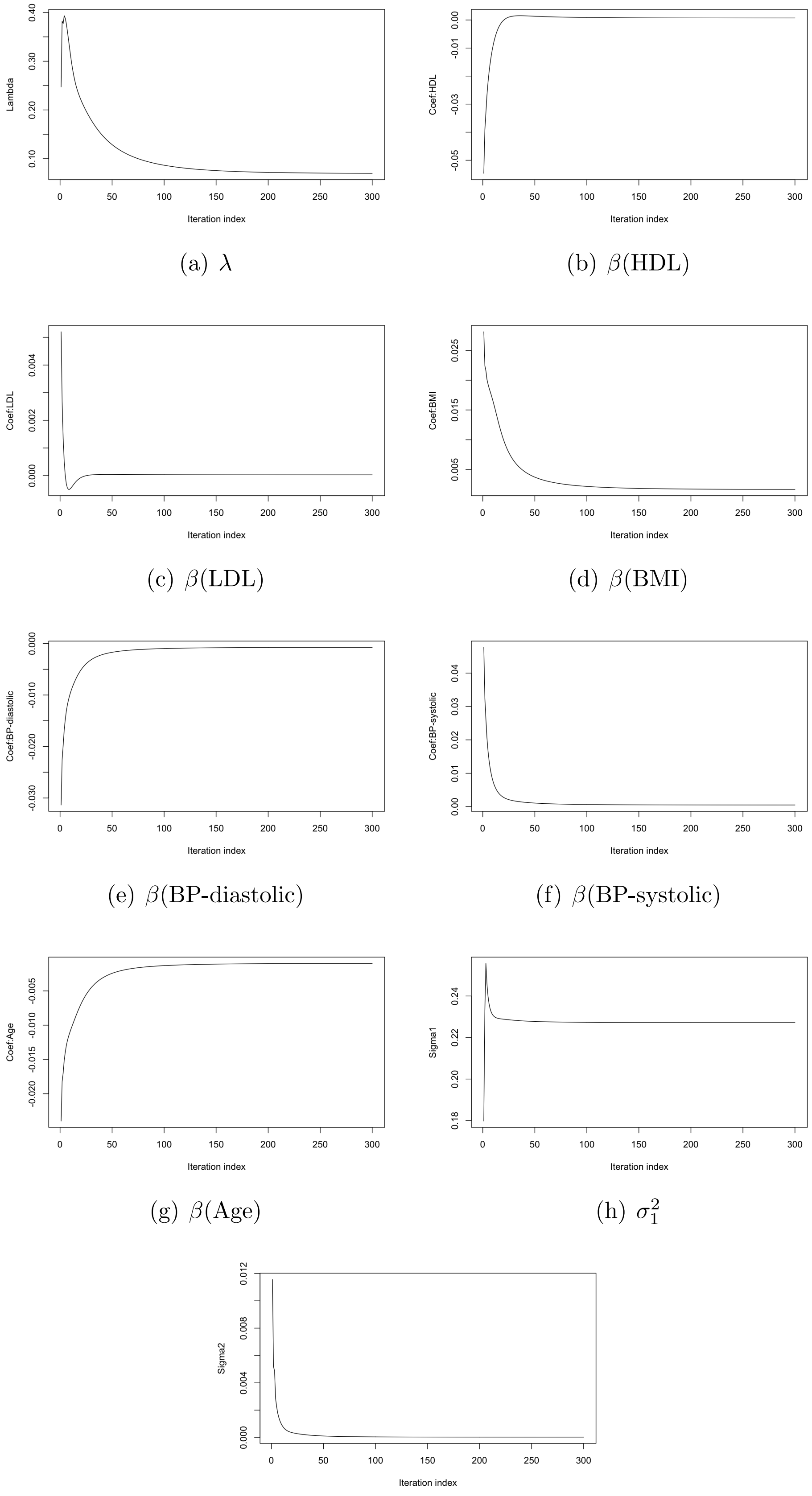
The real data analysis was based on the same set of data discussed in Chapter 5. This study focused on the HbA1c trajectory analysis which only included patients who had HbA1c measures at least twice a year for at least three years. According to the model setting, the discrete time was used with 6-month as increment unit, and the initial HbA1c recording time was set as $t = 0$. In most cases, patients would not have records exactly at beginning or end of a 6-month period in the following-up years. Thus, we used the record closest to each time point within 3 months. Finally, we included 881 patients data with more than five years records for each patient. Then we split the included patients into training (75%) and testing (25%) set randomly. For the predictor variables, we included HDL, LDL, BMI, BP (diastolic and systolic), and age. All predictor variables were centered and scaled before fitting the model.

6.3.1 Parameter estimation: EM algorithm

As described in Section 2, the parameters $\Psi = (\lambda, \beta, \sigma_1^1, \sigma_2^2)$ were estimated by iterating the E- and M step. Figure 6.1 exhibits the convergence for all parameters. Most parameters except λ were convergent quickly after 50 iterations, and λ converged after 200 iterations. Table 6.1 reports the initial and estimated values at iteration 200 of all parameters. The initial values of β were addressed by an ordinary linear regression, and the initial values of $\lambda, \sigma_1^2, \sigma_2^2$ were assigned randomly between 0 to 1.

6.3.2 Prediction

For HbA1c trajectory prediction, we applied the developed model in two scenario: one-step-ahead prediction and k-step-ahead prediction. For one-step-ahead



(a) λ

(b) β(HDL)

(c) β(LDL)

(d) β(BMI)

(e) β(BP-diastolic)

(f) β(BP-systolic)

(g) β(Age)

(h) σ_1^2

(i) σ_2^2

Figure 6.1: Convergence of parameters in EM algorithm

Parameter	Initial value	EM estimates (iteration=200)	p-value
λ	0.24	0.0864	< 0.001
β(HDL)	-0.0546	0.0009	< 0.001
β(LDL)	0.0052	0.0003	0.753
β(BMI)	-0.0239	0.0022	< 0.001
β(BP-diastolic)	0.0281	-0.0010	< 0.001
β(BP-systolic)	-0.0313	0.0007	< 0.001
β(Age)	-0.0313	-0.0013	< 0.001
σ_1^2	0.1	0.2273	-
σ_2^2	0.1	0.0001	-

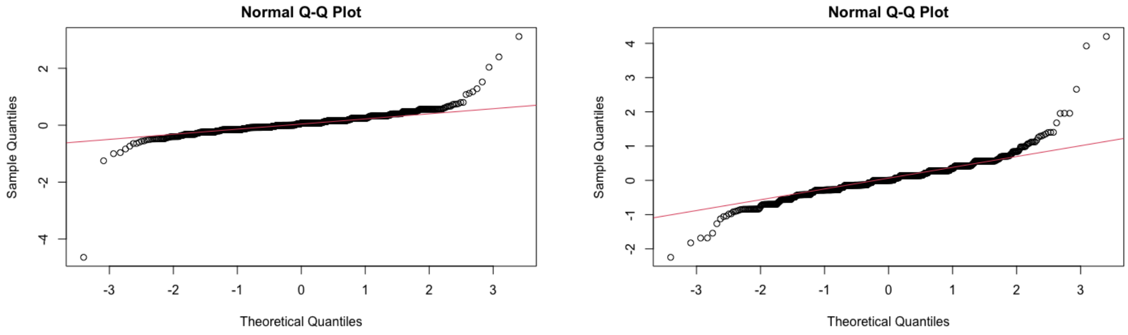
Table 6.1: Initial values and final parameter estimation in EM-algorithm

prediction, we used the information at t to predict the HbA1c at $t+1$; and for k-step-ahead prediction, only the information at the baseline $t = 0$ were used to predict the HbA1c values for $t = 1, ..., k$. The value of k was varying among patients based on their length of records. The performance of prediction in both scenarios were measured by *mean absolute error* (MAE) in Table 6.2. We noted that the one-step-ahead prediction had better prediction performance in terms of MAE. In addition, the k-step-ahead prediction performed better when k was small compared to large k's.

	One-step-ahead	k-step-ahead
MAE	0.1587	0.3372

Table 6.2: Prediction performance measured by MAE in two Scenarios

For a diagnostic checking, we investigated the distribution of the $\hat{e}_0 = y_t - \hat{y}_t$ by a Q-Q plot for normality. Figure 6.2 presents the Q-Q plots for both prediction settings. For both cases, majority of observations were following the theoretical pattern with some deviations for the tails.



(a) One-step-ahead prediction

(b) k-step-ahead prediction

Figure 6.2: Diagnose checking for normal errors

6.4 Discussion

In this chapter, we built an dynamic linear model based on the state-space model settings to fit the HbA1c trajectories for patients with prediabetes. To estimate the involved parameters, an EM algorithm was applied by using the well-known Kalman forward filtering to estimate the latent states and finally calculated the MLE for parameters. The prediction was generated on a separate test set in two scenarios: one-step-ahead and k-step-ahead prediction to fit the practical demands of short-term or long-term prediction. The one-step-ahead prediction performed better in the testing data compared to the k-step-ahead prediction, which highlighted the challenges in the long-term prediction.

The limitations of the approach are as follows. Firstly, in the sample inclusion and exclusion process, we selected patients with more available HbA1c records to build the current model. As a retrospective study, the current selected sample may not be a comprehensive representative of the entire population with prediabetes.

Moreover, in the data inclusion/exclusion step, we selected patients with regular medical records to build the model, and the selected subset may behave differently from the subset without regular records. Secondly, due to the data availability, we included the current list of predictors to model the dynamic change of HbA1c. There are potentials to add more predictors to improve the prediction performance. Finally, discrete time is a genetic setting in our current system which can be further extended as continuous time settings. Although diabetic or prediabetic patients are commonly required to test their HbA1c regularly, a model can provide predictions for more flexible time period is more applicable in practice.

Of the note, our current model settings are based on discrete time, however, we could potentially extend the dynamic setting to continuous time for future work. One possible way is that we can remodel the state equation of μ_t by an exponential function instead of plain linear. For example,

$$\mu_t = \mu_{t-h} e^{f(\phi_{t-h})h} + e_t$$

The estimation of the latent states and parameters may require log transformation to simplify the calculation in forward filtering process. The input predictors X_t can be incorporated in $f(\cdot)$ to model the dynamic changes as in the current model. In addition, the distribution of error terms should also be well defined. Alternatively, we could introduce non-parametric or semi-parametric functions to model the change respect to the continuous time to vary the pattern of the trajectories among different patients.

Another future direction is enlightened by the distribution of residuals. In Q-Q plots 6.2, there exists deviation of normal distribution on both tails. This may suggest violation of normality or a mixed normal distribution. We can further include categorical variables, such as demographics, to adjust the mixed normal issue. On the other hand, we can consider nonlinear relationship and nonparametric models while modeling the states.

Chapter 7: Conclusion

Numerous statistical tools and methods have been developed and applied to health care of diabetes related topics. With more types of tasks and demands appearing in diabetic health services, adjustment and extension of existing methods and development of novel statistical methods can aid and support decision making and health care management for both experts and patients in more applications.

In this dissertation, we have applied and developed various statistical methods to address different demands in several projects. In addition, challenges and limitations in each study are also discussed. More future studies are encouraged by those challenges and limitations.

In chapter 2, we developed an automated algorithm for CGM event detection and information extraction for univariate time series data. To develop the detection algorithm, we applied DTW to measure (dis)similarities between time series with different lengths, hierarchical clustering and cluster averaging method for pattern recognition, and smoothing method and binary encoding method to characterize and simplify the original time series. The proposed method resolved some main challenging points in a typical pattern matching problem, including the uncertainty of target time window length, the issues of scaling and shifting, and redundant and expensive computation cost. In the real data study, the proposed approach performed accurately and efficiently. Moreover, it has potentials to be applied to other streaming time series data analysis for accurate event detection and pattern matching problems. More future works are needed to develop and validate the extension of the current method. First possible extension is that the encoding process can be extended with higher order derivatives for more complicated event patterns. Secondly, the current approach can be potentially extended to multivariate time series data mining problems.

In chapter 3, we focused on analysis methods for data with multiple measures from the same group of subjects. We addressed the within-subject correlation, and further applied approaches based on within-subject PCA for unsupervised variable selection by splitting the within-subject and between-subject variation. Then, we built a low-dimensional framework to measure and visualize users' glucose status across a period of time. The idea of selecting important variables based on PCA can be extended to general data mining problems or some dimension reduction problems when using the PCs directly is not applicable. Current analysis is driven by maintaining more variation of the original data in variable selection procedure, and we will explore alternative criteria to measure the information contained in the data to better guide the selection.

In chapter 4, we investigated whether the influence of some specific types of real life activities on CGM series is significant as the correlation between event and time series. We extended the KNN method, which is a non-parametric test for general two-sample multivariate testing, to medical streaming data mining problem and validated the performance by simulation study. For the future work, we plan to apply alternative two sample multivariate testing methods in the current scenario to verify and compare the performance of different tests. In addition, we will further study the lagged influence associated with more types of activities to better serve the design of therapy on lifestyle change and medication regimen for diabetic patients.

In chapter 5, we developed an accurate diabetes risk prediction model for patients with prediabetes based on commonly available information in EHR systems. In the proposed variable selection method, we incorporated the exhaustive searching strategy with a selection criterion based on nonparametric multivariate two-sample test. More future studies are needed to validate the performance of the proposed variable selection approach. Another future work direction is encouraged by the high computational cost in current approach caused by exhaustive searching. We plan to extend the approach to stepwise or floating scheme with the KNN method to high dimension problems.

In chapter 6, we built an dynamic linear model based on the state-space model settings to fit and predict the HbA1c trajectories for patients with prediabetes. To estimate the involved parameters, EM algorithm was used by applying the well-known Kalman forward filtering to estimate the latent states, and then calculating the MLE for parameters. Against the challenges in current method, we plan to 1) add more predictors to improve the prediction performance; 2) extend the discrete time setting to continuous time; 3) introduce non-parametric or semi-parametric methods to model the changes in addition to linear regression model.

Appendix A: Supplementary Tables

A.1 Chapter 2 Supplementary Table

Severity score	TAR (minute)	TAR_{vh}/TAR (%)
0	0	0
1	(0, 30]	<40%
2	(0, 30]	>40%
2	(30, 60]	<40%
3	(30, 60]	>40%
3	(60, 90]	<40%
4	(60, 90]	>40%
4	(90, 120]	<40%
5	(90, 120]	>40%
5	(120, 150]	<40%
6	(120, 150]	>40%
6	(150, 180]	<40%
7	(150, 180]	>40%
7	(180, 210]	<40%
8	(180, 210]	>40%
8	>210	<40%
9	>210	>40%

Table A.1: Definition of CGM event severity score

A.2 Chapter 5 Supplementary Table

Type of disease	ICD-9	ICD-10
Type 1 Diabetes	250.x1, 250.x3	E10
Type 2 Diabetes	250.x0, 250.x2	E11

Table A.2: ICD-9/-10 codes for Type 1 and 2 Diabetes

Categorical Variable				Numeric Variable		
Variable	Count (%)	Variable	Count (%)	Variable	Range	Median (IQR)
Gender		Race		Age at index	18 – 97	58.4 (16)
Female	3,164 (61.9)	Black	2,352 (46)	HbA1c at index	5.7 – 6.4%	6 (0.4)
Male	1,950 (38.1)	White	2,271 (44.4)	HDL	25 – 110	52.33 (19.7)
		Other/Unknown	491 (9.6)	LDL	20 – 261	109 (42)
Smoking status		Alcohol use		SGPT	6 – 197	21.6 (12.8)
Current	488 (9.5)	Never	313 (6.1)	BMI	14 – 79	30.9 (8.8)
Former	1,118 (21.9)	Infrequent	2,407 (47.1)	BP-diastolic	40 – 200	84 (10)
Never	2,544 (49.7)	Frequent	476 (9.3)	BP-systolic	88 – 250	140 (21)
Unknown	964 (18.9)	Unknown	1,918 (37.5)			
Exercise		Caffeine use				
Sedentary	1,126 (22)	Infrequent/Never	853 (16.7)			
Exercise	1,280 (25)	Frequent	980 (19.2)			
Unknown	2,708 (53)	Unknown	3,281 (64.2)			
Diet		Housing status				
Healthy	300 (5.9)	Mentioned	445 (8.7)			
Need to Improve	392 (7.7)	Not mentioned	4,669 (91.3)			
Unknown	4,422 (86.5)					
Hypertension		Family T2D history				
Diagnosed	3,623 (70.8)	Yes	679 (13.3)			
Not Diagnosed	1,491 (29.2)	No/Unknown	4,435 (86.7)			

Table A.3: Summary of baseline characteristics. Total Sample size = 5,114

Bibliography

[Aarts et al., 2014] Aarts, E., Verhage, M., Veenvliet, J. V., Dolan, C. V., and Van Der Sluis, S. (2014). A solution to dependency: using multilevel analysis to accommodate nested data. *Nature neuroscience*, 17(4):491–496.

[ADA, 2020] ADA (2020). The cost of diabetes. `https://www.diabetes.org/resources/statistics/cost-diabetes`. Accessed: 2021-05-15.

[ADA, 2021] ADA (2021). Diagnosis. `https://www.diabetes.org/a1c/diagnosis`. Accessed: 2021-05-15.

[Adams and MacKay, 2007] Adams, R. P. and MacKay, D. J. (2007). Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*.

[Adane et al., 2020] Adane, T., Getaneh, Z., and Asrie, F. (2020). Red blood cell parameters and their correlation with renal function tests among diabetes mellitus patients: A comparative cross-sectional study. *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, 13:3937.

[Aggarwal et al., 2001] Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, pages 420–434. Springer.

[Aghabozorgi et al., 2015] Aghabozorgi, S., Shirkhorshidi, A. S., and Wah, T. Y. (2015). Time-series clustering–a decade review. *Information Systems*, 53:16–38.

[Aminikhanghahi and Cook, 2017] Aminikhanghahi, S. and Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2):339–367.

[Assessment, 2021] Assessment, G. (2021). 6. glycemic targets: Standards of medical care in diabetesd2021. *Diabetes Care*, 44:S73.

[Association et al., 2020] Association, A. D. et al. (2020). 2. classification and diagnosis of diabetes: Standards of medical care in diabetes—2020. *Diabetes care*, 43(Supplement 1):S14–S31.

[Barnaghi et al., 2012] Barnaghi, P. M., Bakar, A. A., and Othman, Z. A. (2012). Enhanced symbolic aggregate approximation method for financial time series data representation. In *2012 6th International Conference on New Trends in Information Science, Service Science and Data Mining (ISSDM2012)*, pages 790–795. IEEE.

[Battelino et al., 2019] Battelino, T., Danne, T., Bergenstal, R. M., Amiel, S. A., Beck, R., Biester, T., Bosi, E., Buckingham, B. A., Cefalu, W. T., Close, K. L., et al. (2019). Clinical targets for continuous glucose monitoring data interpretation: recommendations from the international consensus on time in range. *Diabetes care*, 42(8):1593–1603.

[Benitez et al., 2001] Benitez, D., Gaydecki, P., Zaidi, A., and Fitzpatrick, A. (2001). The use of the hilbert transform in ecg signal analysis. *Computers in biology and medicine*, 31(5):399–406.

[Berndt and Clifford, 1994] Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, USA:.

[Bezdek et al., 1984] Bezdek, J. C., Ehrlich, R., and Full, W. (1984). Fcm: The fuzzy c-means clustering algorithm. *Computers & geosciences*, 10(2-3):191–203.

[Binder et al., 1981] Binder, Z., Fontaine, H., Magalhaes, M. F., and Baudois, D. (1981). About a multimodel control methodology, algorithms, multiprocessors implementation and application. *IFAC Proceedings Volumes*, 14(2):981–986.

[Biswas and Ghosh, 2014] Biswas, M. and Ghosh, A. K. (2014). A nonparametric two-sample test applicable to high dimensional data. *Journal of Multivariate Analysis*, 123:160–171.

[Bland and Altman, 1995a] Bland, J. M. and Altman, D. G. (1995a). Calculating correlation coefficients with repeated observations: Part 2—correlation between subjects. *Bmj*, 310(6980):633.

[Bland and Altman, 1995b] Bland, J. M. and Altman, D. G. (1995b). Statistics notes: Calculating correlation coefficients with repeated observations: Part 1—correlation within subjects. *Bmj*, 310(6977):446.

[Bonora and Tuomilehto, 2011] Bonora, E. and Tuomilehto, J. (2011). The pros and cons of diagnosing diabetes with a1c. *Diabetes care*, 34(Supplement 2):S184–S190.

[Box and Tiao, 1975] Box, G. E. and Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical association*, 70(349):70–79.

[Bradley, 1997] Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159.

[CDC, 2020a] CDC (2020a). National diabetes statistics report, 2020. https://www.cdc.gov/diabetes/data/statistics-report/index.html. Accessed: 2021-05-15.

[CDC, 2020b] CDC (2020b). Prediabetes - your chance to prevent type 2 diabetes. https://www.cdc.gov/diabetes/basics/prediabetes.html. Accessed: 2021-05-15.

[CDC, 2020c] CDC (2020c). Prevent complications. https://www.cdc.gov/diabetes/managing/problems.html. Accessed: 2021-05-15.

[Chen et al., 2007] Chen, Y., Nascimento, M. A., Ooi, B. C., and Tung, A. K. (2007). Spade: On shape-based pattern detection in streaming time series. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 786–795. IEEE.

[Chiu and Wray, 2010] Chiu, C.-J. and Wray, L. A. (2010). Peer reviewed: factors predicting glycemic control in middle-aged and older adults with type 2 diabetes. *Preventing chronic disease*, 7(1).

[Cho et al., 2018] Cho, N., Shaw, J., Karuranga, S., Huang, Y. d., da Rocha Fernandes, J., Ohlrogge, A., and Malanda, B. (2018). Idf diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes research and clinical practice*, 138:271–281.

[Choi and Marden, 1997] Choi, K. and Marden, J. (1997). An approach to multivariate rank tests in multivariate analysis of variance. *Journal of the American Statistical Association*, 92(440):1581–1590.

[Christine et al., 2015] Christine, P. J., Auchincloss, A. H., Bertoni, A. G., Carnethon, M. R., Sánchez, B. N., Moore, K., Adar, S. D., Horwich, T. B., Watson, K. E., and Roux, A. V. D. (2015). Longitudinal associations between neighborhood physical and social environments and incident type 2 diabetes mellitus: the multi-ethnic study of atherosclerosis (mesa). *JAMA internal medicine*, 175(8):1311–1320.

[Chunfeng, 2001] Chunfeng, H. (2001). Boundary corrected cubic smoothing splines. *Journal of Statistical Computation and Simulation*, 70(2):107–121.

[Chutani and Pande, 2017] Chutani, A. and Pande, S. (2017). Correlation of serum creatinine and urea with glycemic index and duration of diabetes in type 1 and type 2 diabetes mellitus: A comparative study. *National Journal of Physiology, Pharmacy and Pharmacology*, 7(9):914–919.

[Conway et al., 2019] Conway, M., Keyhani, S., Christensen, L., South, B. R., Vali, M., Walter, L. C., Mowery, D. L., Abdelrahman, S., and Chapman, W. W. (2019). Moonstone: a novel natural language processing system for inferring social risk from clinical narratives. *Journal of biomedical semantics*, 10(1):1–10.

[Cover, 1974] Cover, T. M. (1974). The best two independent measurements are not the two best. *IEEE Transactions on Systems, Man, and Cybernetics*, (1):116–117.

[Cover, 1999] Cover, T. M. (1999). *Elements of information theory.* John Wiley & Sons.

[Cryer et al., 2003] Cryer, P. E., Davis, S. N., and Shamoon, H. (2003). Hypoglycemia in diabetes. *Diabetes care*, 26(6):1902–1912.

[Curto et al., 2016] Curto, S., Carvalho, J. P., Salgado, C., Vieira, S. M., and Sousa, J. M. (2016). Predicting icu readmissions based on bedside medical text notes. In *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 2144–a. IEEE.

[Cuturi, 2011] Cuturi, M. (2011). Fast global alignment kernels. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 929–936.

[Dash and Liu, 1997] Dash, M. and Liu, H. (1997). Feature selection for classification. *Intelligent data analysis*, 1(1-4):131–156.

[DeJesus et al., 2017] DeJesus, R. S., Breitkopf, C. R., Rutten, L. J., Jacobson, D. J., Wilson, P. M., and Sauver, J. S. (2017). Incidence rate of prediabetes progression to diabetes: modeling an optimum target group for intervention. *Population health management*, 20(3):216–223.

[DeLong et al., 1988] DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845.

[Du et al., 2006] Du, P., Kibbe, W. A., and Lin, S. M. (2006). Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, 22(17):2059–2065.

[Echouffo-Tcheugui and Selvin, 2020] Echouffo-Tcheugui, J. B. and Selvin, E. (2020). Pre-diabetes and what it means: The epidemiological evidence. *Annual Review of Public Health*, 42.

[Efron and Tibshirani, 1994] Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap.* CRC press.

[Fernandes et al., 2014] Fernandes, M. P., Silva, C. F., Vieira, S. M., and Sousa, J. M. (2014). Multimodeling for the prediction of patient readmissions in intensive care units. In *2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1837–1842. IEEE.

[Friedman and Rafsky, 1979] Friedman, J. H. and Rafsky, L. C. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, pages 697–717.

[Fu, 2011] Fu, T. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164–181.

[Gabbay et al., 2020] Gabbay, M. A. L., Rodacki, M., Calliari, L. E., Vianna, A. G. D., Krakauer, M., Pinto, M. S., Reis, J. S., Puñales, M., Miranda, L. G., Ramalho, A. C., et al. (2020). Time in range: a new parameter to evaluate blood glucose control in patients with diabetes. *Diabetology & metabolic syndrome*, 12(1):1–8.

[Gebregziabher et al., 2010] Gebregziabher, M., Egede, L. E., Lynch, C. P., Echols, C., and Zhao, Y. (2010). Effect of trajectories of glycemic control on mortality in type 2 diabetes: a semiparametric joint modeling approach. *American journal of epidemiology*, 171(10):1090–1098.

[Giorgino et al., 2009] Giorgino, T. et al. (2009). Computing and visualizing dynamic time warping alignments in r: the dtw package. *Journal of statistical Software*, 31(7):1–24.

[Girden, 1992] Girden, E. R. (1992). *ANOVA: Repeated measures*. Number 84. Sage.

[Glauber et al., 2018] Glauber, H., Vollmer, W. M., and Nichols, G. A. (2018). A simple model for predicting two-year risk of diabetes development in individuals with prediabetes. *The Permanente Journal*, 22.

[Green and Silverman, 1993] Green, P. J. and Silverman, B. W. (1993). *Nonparametric regression and generalized linear models: a roughness penalty approach*. Crc Press.

[Guo et al., 2002] Guo, Q., Wu, W., Massart, D., Boucon, C., and De Jong, S. (2002). Feature selection in principal component analysis of analytical data. *Chemometrics and Intelligent Laboratory Systems*, 61(1-2):123–132.

[Guyon et al., 2008] Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L. A. (2008). *Feature extraction: foundations and applications*, volume 207. Springer.

[Hall and Tajvidi, 2002] Hall, P. and Tajvidi, N. (2002). Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, 89(2):359–374.

[Harchaoui et al., 2009] Harchaoui, Z., Vallet, F., Lung-Yut-Fong, A., and Cappé, O. (2009). A regularized kernel-based approach to unsupervised audio segmentation. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1665–1668. IEEE.

[Hennig et al., 2015] Hennig, C., Meila, M., Murtagh, F., and Rocci, R. (2015). *Handbook of cluster analysis*. CRC Press.

[Henze, 1988] Henze, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *The Annals of Statistics*, pages 772–783.

[Hettmansperger et al., 1998] Hettmansperger, T. P., Möttönen, J., and Oja, H. (1998). Affine invariant multivariate rank tests for several samples. *Statistica Sinica*, pages 785–800.

[Hodge et al., 2006] Hodge, A., English, D., O'dea, K., and Giles, G. (2006). Alcohol intake, consumption pattern and beverage type, and the risk of type 2 diabetes. *Diabetic Medicine*, 23(6):690–697.

[Jacobson, 2001] Jacobson, A. (2001). Auto-threshold peak detection in physiological signals. In *2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 3, pages 2194–2195. IEEE.

[Jain et al., 2000] Jain, A. K., Duin, R. P. W., and Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on pattern analysis and machine intelligence*, 22(1):4–37.

[Jolliffe, 1972] Jolliffe, I. T. (1972). Discarding variables in a principal component analysis. i: Artificial data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 21(2):160–173.

[Karpati et al., 2018] Karpati, T., Leventer-Roberts, M., Feldman, B., Cohen-Stavi, C., Raz, I., and Balicer, R. (2018). Patient clusters based on hba1c trajectories: a step toward individualized medicine in type 2 diabetes. *PloS one*, 13(11):e0207096.

[Kawahara and Sugiyama, 2012] Kawahara, Y. and Sugiyama, M. (2012). Sequential change-point detection based on direct density-ratio estimation. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(2):114–127.

[Kazemian et al., 2019] Kazemian, P., Shebl, F. M., McCann, N., Walensky, R. P., and Wexler, D. J. (2019). Evaluation of the cascade of diabetes care in the united states, 2005-2016. *JAMA internal medicine*, 179(10):1376–1385.

[Kedem and Slud, 1981] Kedem, B. and Slud, E. (1981). On goodness of fit of time series models: An application of higher order crossings. *Biometrika*, 68(2):551–556.

[Kedem and Slud, 1982] Kedem, B. and Slud, E. (1982). Time series discrimination by higher order crossings. *The Annals of Statistics*, pages 786–794.

[Keogh et al., 2001] Keogh, E., Chu, S., Hart, D., and Pazzani, M. (2001). An online algorithm for segmenting time series. In *Proceedings 2001 IEEE international conference on data mining*, pages 289–296. IEEE.

[Ley et al., 2014] Ley, S. H., Hamdy, O., Mohan, V., and Hu, F. B. (2014). Prevention and management of type 2 diabetes: dietary components and nutritional strategies. *The Lancet*, 383(9933):1999–2007.

[Liu and Modarres, 2011] Liu, Z. and Modarres, R. (2011). A triangle test for equality of distribution functions in high dimensions. *Journal of Nonparametric Statistics*, 23(3):605–615.

[Lo et al., 2000] Lo, A. W., Mamaysky, H., and Wang, J. (2000). Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. *The journal of finance*, 55(4):1705–1765.

[López et al., 2018] López, B., Torrent-Fontbona, F., Viñas, R., and Fernández-Real, J. M. (2018). Single nucleotide polymorphism relevance learning with random forests for type 2 diabetes risk prediction. *Artificial intelligence in medicine*, 85:43–49.

[Luo et al., 2014] Luo, C., Lou, J.-G., Lin, Q., Fu, Q., Ding, R., Zhang, D., and Wang, Z. (2014). Correlating events with time series for incident diagnosis. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1583–1592.

[Luo et al., 2018] Luo, M., Tan, K. H. X., Tan, C. S., Lim, W. Y., Tai, E.-S., and Venkataraman, K. (2018). Longitudinal trends in hba1c patterns and association with outcomes: A systematic review. *Diabetes/metabolism research and reviews*, 34(6):e3015.

[Mehta et al., 2010] Mehta, S., Shete, D., Lingayat, N., and Chouhan, V. (2010). K-means algorithm for the detection and delineation of qrs-complexes in electrocardiogram. *Irbm*, 31(1):48–54.

[Meijnikman et al., 2017] Meijnikman, A. S., De Block, C., Dirinck, E., Verrijken, A., Mertens, I., Corthouts, B., and Van Gaal, L. (2017). Not performing an ogtt results in significant underdiagnosis of (pre) diabetes in a high risk adult caucasian population. *International journal of obesity*, 41(11):1615–1620.

[Molenaar, 2004] Molenaar, P. C. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, 2(4):201–218.

[Mondal et al., 2015] Mondal, P. K., Biswas, M., and Ghosh, A. K. (2015). On high dimensional two-sample tests based on nearest neighbors. *Journal of Multivariate Analysis*, 141:168–178.

[Myung et al., 2000] Myung, I. J., Kim, C., and Pitt, M. A. (2000). Toward an explanation of the power law artifact: Insights from response surface analysis. *Memory & cognition*, 28(5):832–840.

[Naz and Ahuja, 2020] Naz, H. and Ahuja, S. (2020). Deep learning approach for diabetes prediction using pima indian dataset. *Journal of Diabetes & Metabolic Disorders*, 19(1):391–403.

[Ngufor et al., 2019] Ngufor, C., Van Houten, H., Caffo, B. S., Shah, N. D., and McCoy, R. G. (2019). Mixed effect machine learning: a framework for predicting longitudinal change in hemoglobin a1c. *Journal of biomedical informatics*, 89:56–67.

[Nguyen et al., 2019] Nguyen, B. P., Pham, H. N., Tran, H., Nghiem, N., Nguyen, Q. H., Do, T. T., Tran, C. T., and Simpson, C. R. (2019). Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. *Computer methods and programs in biomedicine*, 182:105055.

[Oehlert, 1992] Oehlert, G. W. (1992). Relaxed boundary smoothing splines. *The Annals of Statistics*, pages 146–160.

[Palshikar et al., 2009] Palshikar, G. et al. (2009). Simple algorithms for peak detection in time-series. In *Proc. 1st Int. Conf. Advanced Data Analysis, Business Analytics and Intelligence*, volume 122.

[Pan and Tompkins, 1985] Pan, J. and Tompkins, W. J. (1985). A real-time qrs detection algorithm. *IEEE transactions on biomedical engineering*, (3):230–236.

[Paparrizos and Gravano, 2015] Paparrizos, J. and Gravano, L. (2015). k-shape: Efficient and accurate clustering of time series. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1855–1870.

[Patel et al., 2008] Patel, A., MacMahon, S., Chalmers, J., et al. (2008). Action in diabetes and vascular disease: Preterax and diamicron modified release and controlled evaluation (advance) collaborative group. intensive blood glucose control and vascular outcomes in patients with type 2 diabetes. *N Engl J Med*, 358(24):2560–2572.

[Petitjean et al., 2011] Petitjean, F., Ketterlin, A., and Gançarski, P. (2011). A global averaging method for dynamic time warping, with applications to clustering. *Pattern recognition*, 44(3):678–693.

[Petris et al., 2009] Petris, G., Petrone, S., and Campagnoli, P. (2009). Dynamic linear models. In *Dynamic Linear Models with R*, pages 31–84. Springer.

[Rakthanmanon et al., 2012] Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q., Zakaria, J., and Keogh, E. (2012). Searching and mining trillions of time series subsequences under dynamic time warping. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 262–270.

[Rama Chandran et al., 2018] Rama Chandran, S., Tay, W. L., Lye, W. K., Lim, L. L., Ratnasingam, J., Tan, A. T. B., and SL Gardner, D. (2018). Beyond hba1c: comparing glycemic variability and glycemic indices in predicting hypoglycemia in type 1 and type 2 diabetes. *Diabetes technology & therapeutics*, 20(5):353–362.

[Rao and Principe, 2002] Rao, Y. N. and Principe, J. C. (2002). Time series segmentation using a novel adaptive eigendecomposition algorithm. *Journal of VLSI signal processing systems for signal, image and video technology*, 32(1):7–17.

[Ratanamahatana and Keogh, 2004] Ratanamahatana, C. A. and Keogh, E. (2004). Making time-series classification more accurate using learned constraints. In *Proceedings of the 2004 SIAM international conference on data mining*, pages 11–22. SIAM.

[Rauh et al., 2017] Rauh, S. P., Heymans, M. W., Koopman, A. D., Nijpels, G., Stehouwer, C. D., Thorand, B., Rathmann, W., Meisinger, C., Peters, A., De Las Heras Gala, T., et al. (2017). Predicting glycated hemoglobin levels in the non-diabetic general population: Development and validation of the direct-detect prediction model-a direct study. *PLoS One*, 12(2):e0171816.

[Reddy et al., 2010] Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., and Srivastava, M. (2010). Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks (TOSN)*, 6(2):1–27.

[Richter et al., 2018] Richter, B., Hemmingsen, B., Metzendorf, M.-I., and Takwoingi, Y. (2018). Development of type 2 diabetes mellitus in people with intermediate hyperglycaemia. *Cochrane Database of Systematic Reviews*, (10).

[Rooney et al., 2021] Rooney, M. R., Rawlings, A. M., Pankow, J. S., Tcheugui, J. B. E., Coresh, J., Sharrett, A. R., and Selvin, E. (2021). Risk of progression to diabetes among older adults with prediabetes. *JAMA internal medicine*, 181(4):511–519.

[Schilling, 1986] Schilling, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, 81(395):799–806.

[Schmidt et al., 2005] Schmidt, M. I., Duncan, B. B., Bang, H., Pankow, J. S., Ballantyne, C. M., Golden, S. H., Folsom, A. R., and Chambless, L. E. (2005). Identifying individuals at high risk for diabetes: The atherosclerosis risk in communities study. *Diabetes care*, 28(8):2013–2018.

[Schulze et al., 2007] Schulze, M. B., Hoffmann, K., Boeing, H., Linseisen, J., Rohrmann, S., Möhlig, M., Pfeiffer, A. F., Spranger, J., Thamer, C., Häring, H.-U., et al. (2007). An accurate risk score based on anthropometric, dietary, and lifestyle factors to predict the development of type 2 diabetes. *Diabetes care*, 30(3):510–515.

[Strom and Egede, 2012] Strom, J. L. and Egede, L. E. (2012). The impact of social support on outcomes in adult patients with type 2 diabetes: a systematic review. *Current diabetes reports*, 12(6):769–781.

[Sun and Xu, 2014] Sun, X. and Xu, W. (2014). Fast implementation of delong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters*, 21(11):1389–1393.

[Swapna et al., 2018] Swapna, G., Vinayakumar, R., and Soman, K. (2018). Diabetes detection using deep learning algorithms. *ICT express*, 4(4):243–246.

[Tabák et al., 2012] Tabák, A. G., Herder, C., Rathmann, W., Brunner, E. J., and Kivimäki, M. (2012). Prediabetes: a high-risk state for diabetes development. *The Lancet*, 379(9833):2279–2290.

[Takagi and Sugeno, 1985] Takagi, T. and Sugeno, M. (1985). Fuzzy identification of systems and its applications to modeling and control. *IEEE transactions on systems, man, and cybernetics*, (1):116–132.

[Urakami et al., 2020] Urakami, T., Yoshida, K., Kuwabara, R., Mine, Y., Aoki, M., Suzuki, J., and Morioka, I. (2020). Significance of "time below range" as a glycemic marker derived from continuous glucose monitoring in japanese children and adolescents with type 1 diabetes. *Hormone Research in Paediatrics*, 93(4):251–257.

[Van den Bruel, 2015] Van den Bruel, A. (2015). The triumph of medicine: how overdiagnosis is turning healthy people into patients.

[Viegas et al., 2017] Viegas, R., Salgado, C. M., Curto, S., Carvalho, J. P., Vieira, S. M., and Finkelstein, S. N. (2017). Daily prediction of icu readmissions using feature engineering and ensemble fuzzy modeling. *Expert Systems with Applications*, 79:244–253.

[Wan et al., 2016] Wan, Y., Gong, X., and Si, Y.-W. (2016). Effect of segmentation on financial time series pattern matching. *Applied Soft Computing*, 38:346–359.

[Westerhuis et al., 2010] Westerhuis, J. A., van Velzen, E. J., Hoefsloot, H. C., and Smilde, A. K. (2010). Multivariate paired data analysis: multilevel plsda versus oplsda. *Metabolomics*, 6(1):119–128.

[Wilson et al., 2007] Wilson, P. W., Meigs, J. B., Sullivan, L., Fox, C. S., Nathan, D. M., and D'Agostino, R. B. (2007). Prediction of incident diabetes mellitus in middle-aged adults: the framingham offspring study. *Archives of internal medicine*, 167(10):1068–1074.

[Xue et al., 1992] Xue, Q., Hu, Y. H., and Tompkins, W. J. (1992). Neural-network-based adaptive matched filtering for qrs detection. *IEEE Transactions on biomedical Engineering*, 39(4):317–329.

[Yazidi et al., 2016] Yazidi, M., Chihaoui, M., Chaker, F., Rjeb, O., and Slimane, H. (2016). Factors predicting glycemic control in type 1 diabetic patient. *Open Medicine Journal*, 3(1).

[Ye et al., 2020] Ye, J., Yao, L., Shen, J., Janarthanam, R., and Luo, Y. (2020). Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes. *BMC Medical Informatics and Decision Making*, 20(11):1–7.

[Yeh et al., 2010] Yeh, H.-C., Duncan, B. B., Schmidt, M. I., Wang, N.-Y., and Brancati, F. L. (2010). Smoking, smoking cessation, and risk for type 2 diabetes mellitus: a cohort study. *Annals of internal medicine*, 152(1):10–17.

[Yokota et al., 2017] Yokota, N., Miyakoshi, T., Sato, Y., Nakasone, Y., Yamashita, K., Imai, T., Hirabayashi, K., Koike, H., Yamauchi, K., and Aizawa, T. (2017). Predictive models for conversion of prediabetes to diabetes. *Journal of Diabetes and its Complications*, 31(8):1266–1271.

[Yudkin, 2016] Yudkin, J. S. (2016). "prediabetes": are there problems with this label? yes, the label creates further problems! *Diabetes Care*, 39(8):1468–1471.

[Yudkin and Montori, 2014] Yudkin, J. S. and Montori, V. M. (2014). The epidemic of pre-diabetes: the medicine and the politics. *Bmj*, 349.

[Zadeh et al., 1996] Zadeh, L. A., Klir, G. J., and Yuan, B. (1996). *Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers*, volume 6. World Scientific.

[Zapranis and Tsinaslanidis, 2010] Zapranis, A. and Tsinaslanidis, P. (2010). Identification of the head-and-shoulders technical analysis pattern with neural networks. In *International Conference on Artificial Neural Networks*, pages 130–136. Springer.

[Zhang and Lian, 2011] Zhang, F. and Lian, Y. (2011). Qrs detection based on morphological filter and energy envelope for applications in body sensor networks. *Journal of Signal Processing Systems*, 64(2):187–194.

[Zhou et al., 2006] Zhou, M., Wong, M.-H., and Chu, K.-W. (2006). A geometrical solution to time series searching invariant to shifting and scaling. *Knowledge and information systems*, 9(2):202–229.