Product Platform Approach to Personalized Type 2 Diabetes Mellitus Management

Ilke Boyaci, Jessica Chin, Abe Zeid, Sagar V. Kamarthi
Northeastern University
Boston, MA 02115, USA

Stephen Agboola, Kamal Jethwani
Partners Healthcare
Boston, MA 02114, USA

Abstract
Type 2 diabetes mellitus (T2DM) is one of the most common chronic disease and the seventh leading cause of death in the United States. Casting T2DM in a product platform framework, this research aims to establish the relationships between contributing factors and complications. Once the tree structure is created by analyzing the historical patient data, patients are clustered based on their complication characteristics and the contributing factors, such as age, race, blood pressure, glycemic levels, and hemoglobin levels. Along with patient clustering, treatment plans are also clustered simultaneously. This creates a mapping between patient groups and treatment groups, with one optimal treatment plan for each cluster of patients. When an individual patient’s membership is determined, the association between the patient and the optimal treatment plan is automatically identified. The healthcare providers can tailor the treatment plan based on the individual’s unique needs. The proposed approach is a more efficient way to treat T2DM patients than the current traditional approach. It also offers flexibility to customize the treatment plans for each individual patient without undue effort.

Keywords
Type 2 diabetes mellitus, product platform, personalized healthcare, clustering

1. Introduction
As of 2011, according to CDC’s national diabetes fact sheet, 25.8 million people, i.e., about 8.3% of the U.S. population are suffering from diabetes mellitus [1]. Type 2 diabetes accounts for 90-95% of all diabetes cases, and if the current trend continues, 1 in 3 people will be diagnosed with diabetes by 2050 [2]. The total direct and indirect cost of diabetes is determined as $147 billion in 2007 [2]. Management and controlling of T2DM via medical treatment plans, exercise and diet, reduce the risks of developing complications such as eye, kidney and nerve diseases. Thus, if the relationship between the contributing factors of T2DM and its complications are established, it will help both patients and healthcare providers cost effectively manage the disease and its complications. The association between the contributing factors and complications is represented in a tree structure in this research. This tree structure is created by using the product platform framework. The product platform concept (PPC) has been widely applied to increase the product variety to meet customers’ varying needs. PPC aims to ascertain and fulfill customers’ wants and needs without lumping them into one homogenous group [3]. This work introduces an engineering methodology to manage chronic diseases like T2DM. The challenge is to classify and cluster the T2DM patients to determine the association between the contributing factors and treatment plans by applying PPC. This is achieved by creating tree structure capturing contributing factors and complications from historical data. The historical patient data includes patient demographics, factors that contribute to T2DM, complications that each patient experience, and treatment methods for both T2DM and complications. Once the tree structure is created, patients are clustered based the characteristics of the complications and the contributing factors, such as age, race, blood pressure glycemic levels, and hemoglobin levels. Since the data set includes the treatment plans used for each patient, the treatment plans are clustered simultaneously, as well. The structuring process, equivalent to the design phase of the platform, is discussed in Section 4 in details.

1 Corresponding Author, Email: sagar@coe.neu.edu, Phone: 617-373-3070
After the platform structure and classifications of the patients are completed, a new diabetes patient is placed into the best-fitting patient cluster based on his/her attributes. The optimal treatment plan is determined through the association between patient clusters and treatment plan clusters. Once the optimal treatment plan is determined, it can be tailored by the physician based on patient’s unique needs. This last step introduces the personalization into the treatment process.

The proposed approach is an efficient way to treat T2DM patients compared to the current traditional approach. It offers physicians the flexibility to customize the treatment plans for each individual patient without undue effort. The paper suggests a novel methodology. There have been some association rule algorithms applied for classification of T2DM patients. Patil et al. [4] proposed an apriori association rule algorithm to generate rules to understand the relationships between measured fields whether the patient goes on to develop diabetes or not. An improved association rule has been presented for classification of T2DM patients to understand whether or not patient goes on to cultivate diabetes by implementing the multilevel based association rules [5]. Decision and classification trees have been applied to diabetes data from the DiabCare program in France; the authors stated that data mining techniques, especially decision trees, can be used successfully to extract knowledge from medical databases [6]. Table 1 summarizes the literature on data mining approaches applied to T2DM.

### Table 1: Review of classification methods for T2DM patients [7]

<table>
<thead>
<tr>
<th>Reference</th>
<th>Objective</th>
<th>Methods used</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>[8]</td>
<td>To improve the diagnostic accuracy of diabetes disease combining PCA and ANFIS</td>
<td>Principal component analysis and ANFIS</td>
<td>PCA is not capable of dealing with missing data and outliers, and large datasets [9]</td>
</tr>
<tr>
<td>[10]</td>
<td>Diagnosis and classification of diabetes</td>
<td>Generalized discriminant analysis and least square support vector machine</td>
<td>Discriminant analysis assumes that the each group follows normal distribution and very sensitive to outliers</td>
</tr>
<tr>
<td>[11-16]</td>
<td>Clustering/classification</td>
<td>CART/ classification tree</td>
<td>Increase or decrease of tree complexity, changes in splitting variables and values, if data have more complex structure then CART may not catch the correct structure of the data [17]</td>
</tr>
<tr>
<td>[18]</td>
<td>Clustering the leader genes and determining interactions among them with k-means algorithm</td>
<td>k-means</td>
<td>Problems with outliers and empty clusters. It does not work well with clusters (in the original data) of different size and different density</td>
</tr>
<tr>
<td>[19]</td>
<td>Feature selection for diabetes type 2</td>
<td>Naïve bayes</td>
<td>Assumes independence of features; it is highly susceptible to curse of dimensionality[20]</td>
</tr>
</tbody>
</table>

The remainder of this paper is organized as follows: Section 2, presents background about Type 2 Diabetes Mellitus. Basics of product platform concept and related literature are described in Section 3. Section 4 describes the proposed methodology in details. Section 5 provides an introduction to the clustering methods and an example on the application of the EM algorithm. Conclusions and future work are presented in the final section, i.e., Section 6.

### 2. What is Type 2 Diabetes Mellitus?

Because of population growth, aging, urbanization and increasing prevalence of obesity and physical inactivity, the number of people with diabetes is increasing [21]. Diabetes is the seventh leading cause of death in the U.S. [1]. Type 2 diabetes, also called non-insulin dependent diabetes mellitus (NIDDM) accounts for 90-95% of all diabetes cases, and if the current trend continues, 1 in 3 people will be diagnosed with diabetes by 2050 [2]. Usually exogenous insulin is not the first line therapy in management of T2DM but it may be required in complicated cases or failure to respond to diet or oral agents [22]. Studies have suggested that the number of people with diabetes increases approximately 1 million every year [23, 24]. According to the National Diabetes Fact Sheet [1], the rates of the complications among diabetes patients can be summarized as below:
• **Heart disease and stroke:** Risk of death from heart disease or stroke is 2-4 times higher in diabetics than in non-diabetics.

• **Hypertension:** 67% of diabetes patients experience hypertension that is greater than or equal to 140/90 millimeters of mercury (mmHg) or used prescription medications for hypertension.

• **Blindness and eye problems:** 28.5% are diagnosed with diabetic retinopathy and 4.4% have advance diabetic retinopathy which causes severe vision loss in 2005-2008.

• **Kidney disease:** 44% of incident cases of chronic renal failure is attributable to diabetes, and 220,290 diabetic people are on chronic dialysis or with a kidney transplant.

• **Neurological disease:** 60% - 70% of people with diabetes have mild to severe forms neuropathy, such as impaired sensation or pain in the feet or hands, slowed digestion of food in the stomach, and erectile dysfunction.

• **Amputations:** More than 60% of non-traumatic lower limb amputations occur in people with diabetes.

• **Other complications:** Many people with diabetes also experience dental problems, pregnancy problems and other metabolic problems.

The incidence of these complications can be reduced with improved glycemic (blood glucose) control, blood pressure control, blood lipid control and preventive care practices for eyes, kidneys and feet. Studies have shown that improved glycemic control significantly reduces the risk of experiencing complications [25].

3. **Product Platform Concept**

Product platform has been widely applied to product design to satisfy customers’ diverse needs. Instead of lumping customers into one homogenous group, their needs are considered on an individual basis. Satisfying individual customer’s needs leads to variation and differentiation in products. Mass customization (MC) is a manufacturing method that aims to satisfy individual customer’s needs within the economies of scale and economies of scope [26]. The focus of MC is to enhance product variety to satisfy customer needs. MC often exploits modular product designs to achieve product variety. Many companies and researchers draw attention to product platform as it is one of the most efficacious strategies for MC [27]. Yang et al. [28] state that product platform design is the key to constitute lower costs in a competitive market. Simpson [29] underscores that a successful product platform helps companies to increase product variety, shorten lead times and reduce costs. Product platform also gives companies ability to easily upgrade products based on new market segments. All the benefits notwithstanding, the biggest challenge is to determine the appropriate individual (differentiating) variables and platform (non-differentiating) variables that best identify the commonality among products. Thus, initial design phase of product platform is very crucial.

At the initial design phase, designers tries to understand very well the relationships among each end-product and manufacturing processes. Identification of the platform variables requires designers to perform longitudinal studies and analysis of the system. Designers also reckon the required approach, platform type and design type based on system characteristics by using numerous types of engineering methods before platform design.

4. **T2DM Product Platform Methodology**

Translating T2DM complications into a product platform framework helps one to understand the disease, complications and the contributing factors better. Deeper understanding eventually leads to better outcomes, such as effective treatment and prevention methods, and personalized treatments. Better understanding of the disease will also help both healthcare providers and patients to improve the management of the disease.

Mapping T2DM and its complications to product platform concept establishes the relationships between contributing factors and complications by creating a tree structure using statistical analysis of historical T2DM patient data and the inputs from medical experts. The data, which will be analyzed in the ongoing study, is gathered from patients at Massachusetts General Hospital (MGH) database which is affiliated with Harvard Medical School. The design phase of the platform concept for T2DM patients, as shown in Figure 1, consists of four phases: definition of the perspective of platform design for T2DM, statistical computation, clustering, and relating patients and treatments.

In Phase I, T2DM, complications and contributing factors are determined by literature review and inputs from medical experts. Understanding the associations between the contributing factors and the complications has a crucial role. Specifications of data for diabetes patients are summarized in Table 2 [22]. At this stage, the main purpose is to explore the disease, its complications, and the contributing factors in medical terms as much as possible. The recommendations for data collection include patient demographics, vitals, family history of the disease, duration of diabetes, obesity type and degree, complications, treatment methods, and other various metabolic conditions.
associated with diabetes and glucose intolerance such as pancreatic diseases, hormonal factors, drugs and chemical agents, insulin receptor abnormalities, and genetic syndromes [22]. Figure 2 represents some of the basic contributing factors and some of the main complications at high level.

Figure 1: Design phase of the platform concept for T2DM

Table 2: The recommended and suggested specifications of data for diabetes patients

| Data Required to classify an individual | • Fasting plasma glucose (FPG)  
|  | • OGTT (if FPG is normal)  
|  | • Insulin dependence, ketosis proneness  
|  | • Height/Weight  
|  | • Onset during pregnancy  
|  | • Previously demonstrated glucose intolerance  
|  | • Presence of higher statistical risk for glucose intolerance  
|  | • Presence of conditions/syndromes associated with diabetes mellitus and glucose intolerance  

| Recommendations for additional data to be collected to describe clinical patients and research subjects | • Age/date of birth  
|  | • Sex  
|  | • Race/ethnic origin  
|  | • Socioeconomic status  
|  | • Age of onset and date of diagnosis  
|  | • Duration of diabetes  
|  | • Type of therapy  
|  | • Immediate family history of diabetes  
|  | • Degree and type of obesity  
|  | • Metabolic perturbation (pregnancy, trauma, infection)  
|  | • Nature and extent of complications  
|  | • Source of medical care  

| Suggestions for other data desirable to more fully describe research subjects | • Adherence to therapy (compliance)  
|  | • Degree of control of hyperglycemia 

Given that the number of the complications and contributing factors is very big, it is impossible to include every single complication and factor in the analysis. Thus, some of them should be eliminated to decrease the complexity of the analysis. Yet, it is important to determine which complications and factors should be eliminated and which should be
kept for the sake of the scope and validity of the study. In Phase II, first, the most important complications and contributing factors are determined via statistical analysis of the data. Figure 3 illustrates the first step of this phase. In this step, the significance of the complications and factors are determined, and less significant ones are eliminated from the study. Second, the relationships between those complications and contributing factors are determined. After Phase II is completed, the tree structure for T2DM patients is created. Figure 4 represents an illustration for the tree structure. The 5-level tree structure, which is called ontology in the rest of the paper, summarizes the organs that are affected by T2DM, diseases at high level, complications at lower level, and contributing factors.

The importance of the definition of the relationships between complications and contributing factors is emphasized when the treatment plan is needed to be tailored and personalized for an individual patient. If the associations and the significance of the associations between the complications and contributing factors are clearly defined with statistical analysis, healthcare provider will easily be able to address the point where personalization required in an individual’s treatment plan with undue effort and time.

After the ontology of the significant complications and contributing factors is created, the next step is to cluster the T2DM patients. Clustering targets to define the natural group structures in the database with similar instances. In this study, the T2DM patients are clustered based on the type of complications that they experience, severity of the complications, and levels of the contributing factors.

It is essential to employ an appropriate clustering algorithm based on the desired achievements of the analysis, and size and properties of the collected data. There are many restrictions that apply to clustering algorithms such as the necessity for clusters to be convex sets or to follow a certain distribution, large number of variables, and big data sizes. [30]. The proposed clustering method for the current study is discussed in Section 5.
Since the data set includes the treatment plans for each patient, the treatment plans are clustered simultaneously once the T2DM patients are clustered. The goal of mapping the patient clusters to the treatment plans is to offer an optimal treatment plan, which can be tailored by healthcare providers, for each cluster of patients.

<table>
<thead>
<tr>
<th>T2DM 1st level</th>
<th>Organs (Functions) 2nd level</th>
<th>Diseases (Processes) 3rd level</th>
<th>Complications (Components) 4th level</th>
<th>Factors (Attributes) 5th level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3: Phase II determining the significant factors and complications (F represents significant factors and C represents significant complications)

5. Clustering Methods

The quality of the collected data in healthcare has always been an issue for statistical inferences. Most of the time, healthcare data have missing or incomplete information. This has been a challenge for researchers. Missing values may contain useful information. Omitting or improperly handling the missing data may lead to biased results and invalid results that reduces the power of the study. [31]. Thus, one of the main purposes in healthcare area is to improve the quality of the data by reducing the missing values and removing the noise in the data [32].

The large dimension of the data is also an issue in healthcare data. Besides mostly including missing values, healthcare data, often, have several variables (attributes), and the size of the data is usually big. Because of the huge variety of the variables, the complete data cannot be summarized with a single distribution or model. In healthcare data, there are many sub-populations within a larger population, and it may require a combination of discrete and continuous representation of the population [33]. To overcome these challenges, this research uses finite mixture models and the expectation maximization algorithm for clustering T2DM patients to overcome those problems aforementioned above.

Mixture models are mainly used for generating statistical inferences about the properties of sub-populations within a larger population, without sub-population identity information. The simplest and most natural derivation of the mixture model arises when one samples from a population that consists of several homogeneous subpopulations [34]. The finite mixture models (FMMs) allow flexible modeling of heterogeneous data because it incorporates a combination of discrete and continuous representation of population heterogeneity [33].
Implementation of finite mixture models requires the estimation of the model parameters, yet, it is not an easy task. Estimation procedure includes the estimation of the number of components within the model and of the values of model parameters. Expectation Maximization (EM) algorithm is one of the most commonly used estimation methods for mixture models. The EM algorithm is used to find the maximum likelihood estimator for finite mixture models when the model has latent (missing) variables [35]. It aims to find local maximums of various different starting points to find the global maximum [33].

The EM algorithm alternates between two steps: expectation (E) step, and maximization (M) step. In step E, algorithm aims to compute the expectation of the log likelihood of complete data with respect to latent variables, and step M targets to maximize the expected log likelihood of the complete data [36, 37]. Expectation maximization is also very powerful for variable estimation in probabilistic models with missing data [37], thus it can be considered as a probabilistic imputation method. It has been shown that the EM algorithm has a significant accomplishment on missing data and mixture density problems [38]. The EM algorithm can also accommodate categorical variables [32], and it is considered as an important advantage as healthcare data include numerous categorical variables, such as gender, race, and age.

5.1 Example: Clustering of “Fasting blood glucose levels”
Fasting blood glucose (FBG) level is the glucose levels in the blood 2 hours after fasting. It is an important factor for T2DM patients that require close monitoring of all times as it may cause fatal outcomes. Table 3 illustrates the FBG levels of T2DM patients. The top row represents the intervals for FBG levels. Rows display all the n T2DM patients. In the table, 0 implies the absence and 1 implies the presence of the corresponding FBG level for each patient.

<table>
<thead>
<tr>
<th>T2DM Patients</th>
<th>Fasting Blood Glucose Levels (mg/dL)</th>
<th>Excellent</th>
<th>Good</th>
<th>Dangerous</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>&lt;115</td>
<td>&gt;150</td>
<td>&gt;180</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;150</td>
<td>&gt;180</td>
<td>&gt;215</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;180</td>
<td>&gt;215</td>
<td>&gt;250</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt;250</td>
<td>&gt;280</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>n</td>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Total # of 1s</td>
<td>a</td>
<td>b</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

The last row shows the total number of patients for a given FBG category. For example, total number of the patients whose FBG levels are between 150mg/dL and 180mg/dL is c. The data given in the Table 3 includes missing values as seen in row 3. Because of the missing values in the data, the probability distributions cannot be determined. Thus, the maximum likelihood estimate should be calculated for model parameters. Table 4 represents the probability distributions and the total number of patients for each FBG category.

Table 4: Probability distributions for FBG levels variable and the total number of patients within the specified range

<table>
<thead>
<tr>
<th>Model Parameter</th>
<th># of patients within the specified range of FBG</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(FBG&lt;115mg/dL)=1/4</td>
<td>a</td>
</tr>
<tr>
<td>P(115≤FBG&lt;150mg/dL)=μ</td>
<td>b</td>
</tr>
<tr>
<td>P(150≤FBG&lt;180mg/dL)=1/4</td>
<td>c</td>
</tr>
<tr>
<td>P(180≤FBG&lt;215mg/dL)=2μ</td>
<td>d</td>
</tr>
<tr>
<td>P(215≤FBG&lt;250mg/dL)=1/4</td>
<td>e</td>
</tr>
</tbody>
</table>
\[ P(250 \leq \text{FBG} < 280 \text{mg/dL}) = \frac{3}{8} - 3\mu \]

At Iteration 1 \((t=1)\), the labeled data likelihood can be expressed as:

\[ P(a, b, c, d, e, f | \mu) = K \left( \frac{1}{4} \right)^a (\mu)^b \left( \frac{1}{8} \right)^c (2\mu)^d \left( \frac{1}{4} \right)^e \left( \frac{3}{8} - 3\mu \right)^f \]  

(1)

The log-likelihood of Equation (1) is:

\[ \log P(a, b, c, d, e, f | \mu) = \log K + a \log \left( \frac{1}{4} \right) + b \log \mu + c \log \left( \frac{1}{8} \right) + d \log 2\mu + e \log \left( \frac{1}{4} \right) + f \log \left( \frac{3}{8} - 3\mu \right) \]  

(2)

If the derivative of Equation (2) is set to 0:

\[ \frac{\partial}{\partial \mu} \log P(a, b, c, d, e, f | \mu) = \left( \frac{b}{\mu} \right) + \left( \frac{2d}{2\mu} \right) + \left( \frac{3f}{3 - 3\mu} \right) = 0 \]  

(3)

The maximum likelihood estimate for \( \mu \) can be found as:

\[ \mu = \frac{3(b + d)}{8(b + d + f)} \]  

(4)

The hidden data likelihood occurs if only it is known that there are \( G \) number of patients in the “Good” category, which includes the patients with FBG levels within the range of 115 mg/dL and 180 mg/dL. In such case, the maximum likelihood estimate of \( \mu \) can be found by EM algorithm.

**E Step:**

As it is known that

\[ \frac{b}{c} = \frac{\mu}{1}, G = b + c \]  

(5)

Then,

\[ b = \frac{8\mu}{1 + 8\mu}, c = \frac{1}{1 + 8\mu} G \]  

(6)

**M Step:**

The values found for \( b \) and \( c \) in the E-step is inserted into the Equation (4) to determine the new \( \mu \) value.

\[ \mu_{\text{new}} = \frac{3 \left( \frac{8\mu}{1 + 8\mu} G + d \right)}{8 \left( \frac{8\mu}{1 + 8\mu} G + d + f \right)} \]  

(7)

The new value for \( \mu \) is calculated at the end of each iteration and the E and M steps continue until the model converges to an optimum value for the model parameter. The graphical illustration of the convergence of the EM algorithm is shown in Figure 5 [37].
Figure 5: Convergence of the EM algorithm. Starting from $\mu(t)$, E-step constructs $h_t$ function that defines the lower limit for the objective function, $\log P(X; \mu)$. In the M-step, $\mu(t + 1)$ is computed as the max value of the $h_t$ function.

The algorithm continues to repeat E and M steps until the objective function is maximized.

This example provides a simple illustration about how the EM algorithm applies to T2DM case in one dimension. In this example, only one contributing factor, FBG, is considered for the sake of simplicity. In the actual problem, all the significant factors will be evaluated, and clusters will be formed accordingly.

6. Conclusions & Future Work

In this paper, a novel approach is presented to map product platform concept to T2DM to determine the optimal treatment plan for each cluster of patients to better manage their diabetes and related complications. The statistical analyses help to reveal the associations between the contributing factors and the complications, therefore physicians can be able to determine where the treatment plan needs personalization for an individual T2DM patient without undue effort. The personalization of the treatments can be achieved in an effective and efficient way.

The future work is to be aimed at applying the proposed approach to a real healthcare dataset that is collected from Massachusetts General Hospital’s patient records. The data pre-processing, statistical analysis, and clustering algorithms will be performed on SAS 9.3 software. An effective way to find the optimal treatment plan and to personalize the treatments for the entire T2DM patient community will also be studied.

References

20. Xie, L., KNN Classification and Regression using SAS. SD-09.
34. Lindsay, B.G. Mixture models: theory, geometry and applications. in NSF-CBMS regional conference series in probability and statistics. 1995. JSTOR.