Received: 7 October 2019 | Revised: 10 February 2020 | Accepted: 17 February 2020

DOI: 10.1111/coin.12313

SPECIAL ISSUE ARTICLE

Computational Intelligence WILEY

Opening the black box: Personalizing type 2 diabetes patients based on their latent phenotype and temporal associated complication rules

Leila Yousefi[1] | **Stephen Swift**[1] | **Mahir Arzoky**[1] | **Lucia Saachi**[2] | **Luca Chiovato**[3] | **Allan Tucker**[1]

[1]Department of Computer Science, Brunel University London, London, UK

[2]Department of Computer Science, University of Pavia, Pavia, Italy

[3]Unit of Endocrinology, University of Pavia, Pavia, Italy

Correspondence
Leila Yousefi, Department of Computer Science, Brunel University London, London, UK.
Email: leila.yousefi@brunel.ac.uk; lilyyousefi84@gmail.com

Abstract

It is widely considered that approximately 10% of the population suffers from type 2 diabetes. Unfortunately, the impact of this disease is underestimated. Patient's mortality often occurs due to complications caused by the disease and not the disease itself. Many techniques utilized in modeling diseases are often in the form of a "black box" where the internal workings and complexities are extremely difficult to understand, both from practitioners' and patients' perspective. In this work, we address this issue and present an informative model/pattern, known as a "latent phenotype," with an aim to capture the complexities of the associated complications' over time. We further extend this idea by using a combination of temporal association rule mining and unsupervised learning in order to find explainable subgroups of patients with more personalized prediction. Our extensive findings show how uncovering the latent phenotype aids in distinguishing the disparities among subgroups of patients based on their complications patterns. We gain insight into how best to enhance the

prediction performance and reduce bias in the models applied using uncertainty in the patients' data.

KEYWORDS

diabetes associated complication rules, latent variable discovery, patient personalization, temporal phenotype, time series clustering

1 | INTRODUCTION

Predicting complications associated with the disease is challenging. They can be numerous and can interact in complex nonlinear ways throughout the disease process. However, if we can better predict the onset of different complications in individual patients, then we can intervene more effectively. In addition, to gain patients trust and satisfaction, it is mandatory to understand/explain influencing factors of disease that guides decisions. Black box AI models in the clinical decision-making process are models that attempt to predict/diagnose/forecast/group patients using complex parameters that are not easily understood. For example, the complexity of countless hidden layers in a deep neural network and their interconnections makes it challenging to determine precisely how predictions are being made. Compare this to decision trees or graphical models where inference is more transparent and therefore explainable. Previously, we have explored the use of probabilistic graphical models to build more transparent methods of modeling disease progression. In particular, we used dynamic Bayesian networks to model clinical data and predict the onset of type 2 diabetes mellitus (T2DM) complications.[1] We developed methods to infer the location of hidden variables within these models in order to improve prediction.[2] The behavior of these hidden variables over the course of the disease process can be thought of as a "temporal phenotype" for an individual patient,[3] which is considered as a "latent phenotype."

Preliminary experiments obtained in Reference 4 showed that it is possible to find subgroups of patients only based on their latent phenotype. Nevertheless, the techniques used in these investigations were not validated for interpreting each subgroup to enhance the prediction of the associated complications. Therefore, this study facilitates a hybrid type approach that utilizes a variety of patients subgroups in which the prediction of the associated complications is improved for optimal performance. These techniques can also be combined for a better understanding of the latent variable as well as an underlying pattern of complications for the type of patients. In this article, temporal association rules (TARs) are utilized to identify the frequent co-occurrence of complications over time. An integration of TARs and pattern clustering attempts to build meaningful subgroups. The obtained clusters of the rules are compared with clusters of the latent phenotypes that are extracted from the hidden variable by using dissimilarities and dynamic time warping (DTW) distance among patients.[3] In Section 3, we discuss the data used to explore our approach along with details of the methods introduced. In Section 4, we document the results of these methods. The prediction accuracy of the complications is also validated the contribution of this study in terms of obtaining a higher performance by using the discovered subgroup comparing to the raw dataset which did not consider the presence of the latent phenotype and the proposed hybrid methodology. Section 5 discusses the challenges and our solution in more detail when tested on the diabetes data before concluding in Section 6.

1.1 | Related work

The World Health Organization (WHO) reported that T2DM accounts for at least 90% of all diabetes types. Another study in WHO revealed that T2DM patients are at increased risk of long-term vascular comorbidities, which is known as "underlying cause of death" and severe phenotype of the disease.[5] It has previously been observed that patients with T2DM are also at an increased risk of microvascular comorbidities, including nephropathy, neuropathy, and retinopathy.[5] Similar to diabetic type 1 patients, although genetic factors impact on developing T2DM, it is believed ignorance of developing complications harms patients' life because it may develop a different profile of complications and features, which changes over time per follow-up visit. However, these life-threatening complications remain undiagnosed for a long time because of the hidden patterns of their associated risk factors.[6] The underlying pattern of the complications is known as the major source of mortality and morbidity in T2DM and how their co-occurrence is followed/caused by other complications associated with the disease.[7] That is because predicting a target complication can be challenging without the consideration of the effects of its associated complications.

Understanding the associated pattern of complications has been used significantly in the clinical domain.[8] It provides an insight into the prediction and relative prevention of the associated complications, which are expected to occur in a patient follow-ups.[9] It generally can lead to less suffering time for patients while saving time and cost to healthcare. However, that is highly dependent on the stage of disease along with the prior occurring complications, which is associated with time series analysis. In time series analysis, every disease risk factor and complication is determined by various features in previous patient visits (time interval).

In this work, we attempt to address this issue and present an informative rules/ordering pattern of patient behavior, with an aim to capture the complexities of the associated complications' over time. The proposed descriptive strategy has been regarded as a useful tool known as association rules (ARs) to detect interesting relationships among T2DM complications. ARs strategy originated from learning patterns from supermarket transaction data and was introduced by Agrawal.[10] Temporal abstraction (TA) has also been employed for the segmentation and aggregation of time series data into a symbolic representation, suitable for decision making and data mining.[11] TARs[12] is an extension to ARs[10] to analyze basket data that include a temporal dimension to order related items. Many algorithms with temporal rules work by dividing the temporal transitions database into different partitions based on the time granularity. For example, different mining algorithms are reformulated and presented to reflect the new general TARs, and these include progressive partition minder (PPM), segmented progressive filter (SPF), and TAR algorithm.[10] Various algorithms are proposed for the incremental mining of TARs, especially for numerical attributes.[13] Allen's rules[14] abstracted time series data into a relation (PRECEDES) to find TARs in Reference 15. Various ways have been proposed to explore the problem of TARs discovery.[16] Nevertheless, previous studies employed ARs strategy on a given subset specified by the time,[17] while not considering the specific exhibition period of the elements.

Association rule mining (ARM) finds frequent patterns by mining ARs with the use of two basic parameters of support and confidence.[18] The majority of the previous ARM algorithms worked by dividing the temporal transitions database into different partitions based on the time granularity obliged. Then mining TARs were employed by locating frequent temporal item subsets within these partitions. However, the incremental mining of

TARs for numerical attributes cannot always be easily adapted to a transaction database. Despite all efforts, it appeared that no method exists today that can find meaningful subgroups of patients based on the underlying pattern of complications in the existence of the latent risk factors. With a similar objective as this thesis, Moskovitch and Shahar[11] conducted a study in which time-interval mining methods obtained informative temporal patterns for finding relationships in the transitivity inherent in time series diabetic patients. Also, they exploited TA for the segmentation and aggregation of a time series into a symbolic representation, suitable for decision making and data mining. Although Moskovitch's paper is consistent with this study by using supervised learning in time series diabetes data, it differs from this work in finding meaningful time series patterns only based on gender not complex temporal patterns from a longitudinal clinical dataset with the appearance of latent risk factors.

A considerable amount of literature has been published on TARs to discover interesting rules based on several quality filtering metrics known as constraints. Luna et al[19] conducted an empirical study in the optimization of the most interesting groups of metrics. In addition, recently, they provided a rich review on the commonly used frequent itemsets mining algorithms.[20] Part of this work is motivated by Hashler and Karpienko,[21] which introduced a distance-based clustering of ARs. It then is supported by Li et al,[22] which revealed that applying a postprocessing method to ARs to find the most frequent calendar patterns improves interpretability in the descriptive analysis. Unfortunately, the previous methods were not only limited in time granules but also increased the uncertainty in the relationship among rules, while there was overlap among clusters in k-means clustering. The frequent pattern mining research significantly affects data mining techniques in longitudinal data. A postprocessing approach in Reference 23 attempted to extract interesting subsets of temporal rules within T2DM data. However, it only considered characteristic patterns of administrative data without the appearance of latent variables. Other researchers have undertaken AR mining of clinical data. Lee et al attempted to address the issue in Reference 24, and these have led to the proposal of the concept of general TARs, where the items are allowed to have varying exhibition periods, and their support is made based on that. Another piece of research conducted by Plasse et al[25] looked at finding homogeneous groups of variables. They suggested that a variable clustering method could be applied to the data in order to achieve a better result in pattern discovering methodology. However, their strategy to mine ARs differs from this study in which the number of rules was reduced only based on hierarchical clustering applied to items, not to multiple identical binary attributes. Among these, some methods uncovered temporal patterns and relationships among clinical variables, including causal information[26] and numeric time series analysis.[27]

In longitudinal clinical data (eg, T2DM), one of the most important factors in the high number of dependencies among features and complications is the appearance of unmeasured risk factors. Surprisingly, the effect of understanding unmeasured variables, which play an important role in disease prediction, does not seem that closely examined. The reason behind this might be because of the recent focus on the AI models with a black box nature. What is more, there are several issues with TARs when there are some rare rules of particular interest.[28]

Given the strong association between the complications, another challenge is the existence of unknown (latent) factors in the data. It is crucial to understand better the latent variables and other associated risk factors to be able to predict their underlying patterns

earlier than their actual occurrence time. That can be done by exploring a well-chosen group of potentially all significant patients' patterns while identifying temporal phenotypes based on their unmeasured risk factors with reasonably minimal outliers. Having insight into the causal associations, among disease complications, we attempt to open a black box model to ease interpretation of the hidden patterns of complications in an accurate predictive model. We, therefore, need to take into consideration both descriptive and predictive data mining strategies.

Nevertheless, Lakkaraju et al[29] suggested that there is a trade-off between patient personalization (in a descriptive analysis) and prediction performance (in predictive analysis). In other words, aiming explainability (in an explainable/interpretable model) is often possible at a higher cost of the predictive accuracy (in a Black box model).[8] Therefore, in the black box models, it can be challenging to determine from just temporal clinical data what is coordinating the visible patterns, to separate the underlying causes into meaningful and spurious causes, which help patient stratification with understanding hidden variables. Black box AI models in decision making are mostly based on deep learning techniques with many latent variables. For example, these models map a patient's latent/risk factor into a class only based on the combinations of weights without exposing the reasons why. Black box models are problematic not only for lack of transparency but also for possible biases inherited by the algorithms from clinician's mistakes.[30] This issue is caused based on the human prejudices and underestimation of the impact of the risk factors underlying behavior/pattern as well as the existence of latent variables in the dataset, which may lead to incorrect and unfair decisions.

Nevertheless, considering all of this evidence, none of the above studies have clustered uneven time series clinical data based on a hidden variable for extracting temporal phenotype and behaviors of patients. There are quite few research studies on predicting T2DM complications and T2DM black box models. However, studies on explaining an unknown risk factor/latent phenotype by using a hybrid data mining methodology (including descriptive and predictive) are rare to find in literature.

In this work, we argue that binary complications could be predicted accurately by discovering the latent factors and adding them to the observed data.[1] Another study in Reference 3 have primarily concentrated on the clustering approach based on the latent variable to personalise the patients. That is consistent with the very current work in Reference 4, which also provided a comparison methodology to evaluate the discovered latent variable clusters by using a combination of supervised learning such as clustering and TARs among the binary complications. Hence, Reference 4 found similar clusters to those obtained in Reference 3. This article extends the previous work in Reference 4 in order to take into consideration both descriptive and predictive analysis when it comes to the basic idea of precise prediction through and explainable model.

To sum up, the motivation behind this work is conducting new research in order to suggest that the identification of a “latent phenotype” can be utilized to separate patients into meaningful subgroups with the consideration of the relation among T2DM complications. In general, as observed balancing strategies from the prior studies to deal with imbalanced data for one complication at a time, it is challenging to obtain the prediction performance enhancement for all complications. Therefore, another motivation for this study is to improve the performance for predicting associated complications considering the imbalance issue.

2 | METHODOLOGY

2.1 | Data

The data for this study consist of prediagnosed T2DM patients aged 25 to 65 years (inclusive) that were recruited from clinical followups at the "IRCCS Instituti Clinic Scientifici" (ICS) Maugeri of Pavia, Italy. The MOSAIC project funds the data under the Seventh Framework Program of the European Commission, Theme ICT—"2011.5.2 Virtual Physiological Human (600914)" from 2009 to 2013. The dataset consists of physical examinations such as cholesterol and blood pressure and laboratory data including HbA1c measurements and lipid profile. For this study, certain complications and risk factors (predictors) were selected based on existing literature on diabetes[31] and using recommendations from the clinicians at ICS. The selected T2DM complications are retinopathy (RET), hypertension (HYP), nephropathy (NEP), neuropathy (NEU), and liver disease (LIV). Here, the predictors are identified and selected from the dataset: body mass index (BMI), systolic blood pressure (SBP), high-density lipoprotein (HDL), glycated hemoglobin/HbA1c (HBA), diastolic blood pressure (DBP), cholesterol (COL), smoking habit (SMK), and creatinine (CRT).

2.2 | Preliminaries

From diabetes health status records, the T2DM dataset is accumulated (which is denoted here as DS) from prediagnosed diabetic patients. For each patient in T2DM dataset defined the following notations. $DS = \sum_{i=1}^{p} \pi_i = (\pi_1, \pi_2, \pi_3, \dots, \pi_p)$, where π demonstrates a distinct patient, i identifies the patient in which $i \leq p$, and p denotes the maximum number of patients in DS. V_i refers the visits of patient i (π_i), there is a maximum T_i of visits V_i, where $p = 356$ represents a maximum number of patients and T_i is a maximum of visits (V_i) for ith patient. The number of visits is not necessarily the same for different i and varies ($2 \leq T_i \leq 300$). Hence, there is a total of $T = 3959$ visits/instances/time series in DS, which contains temporal observations of the occurring complications. Let $\pi_i = \sum_{j=1}^{T_i=300} V_{ij} = (V_{i1}, \dots, V_{iv}, \dots, V_{iT_i})$ be a set of visits for i-th patient with T_i time series where V_{iv} represents the vth visit of π_i (as demonstrated in Equation (1)). For each of the patients in DS, over which linear order is defined, $v \leq z$ means V_{vi} occurs before or is earlier than V_{iz} in $[V_{iv}V_{iz}]$. In order to clarify the dataset, a vector of patients is demonstrated in Equation (1).

Tables 1 and 2 represent the selected T2DM complications (comorbidities), risk factors, and their clinical control values. Data are discretized into qualitative states (binary and nonbinary features) of ordinal clinical risk by using statistical parameters such as mean, median, and SD. The main goal of this thesis is to understand the underlying patterns of associated binary complications.

In this study, the association of nonbinary risk factors/symptoms has not been considered in order to extract rules among T2DM complications. The reason behind this is that by utilizing the discovered latent variable, the overall behavior of T2DM risk factors is captured by using the IC*LS algorithm in a DBN framework (which is called a "latent phenotype"). Therefore, this study only concentrates on five binary complications as predictive target classes in a binary classification problem (two categories of classes: "high" or "low"

TABLE 1 The description of T2DM complications, clinical control values, and the discretized states

Target complication	Diagnosis outcome	Clinical risk class
Retinopathy (RET)	{Negative,Positive}	{low,high}
Neuropathy (NEU)	{Negative,Positive}	{low,high}
Nephropathy (NEP)	{Negative,Positive}	{low,high}
Liver Disease (LIV)	{Negative,Positive}	{low,high}
Hypertension (HYP)	{Negative,Positive}	{low,high}

TABLE 2 The description of the T2DM clinical features, risk factors, control values, and the discretized states

T2DM risk factors	Control value (Mean±SD)	Discretized value
HbA1c (HBA)	6.6 ± 1.2 (%)	{low,medium,high}
BMI	26.4 ± 2.4 (kg/m^2)	{low,medium,high}
Creatinine (CRT)	0.9 ± 0.2 (mg/dL)	{low,medium,high}
Cholesterol (COL)	0.9 ± 0.2 (mg/dL)	{low,medium,high}
HDL	1.1 ± 0.3 (mmol/l)	{low,medium,high}
Diastolic blood pressure (DBP)	91 ± 12 (mmHg)	{low,medium,high}
Systolic blood pressure (SBP)	148 ± 19 (mmHg)	{low,medium,high}
Smoking habit (SMK)	{nonsmoker, ex-smoker, smoker}	{low,medium,high}

risk). Furthermore, a complication class value of low risk (zero) represents a patient visit in which the complication is not present; otherwise, it is at high risk (one). For instance, a complication class value of zero represents a patient visit in which the complication is not present; otherwise, it is one. Alternatively, other risk factors associated with a patient (symptoms/clinical tests) are abstracted in the multiclass classification problems with more than two targets including "high," "medium," and "low" risk patient, according to a diabetes expert's definitions.[32,33]

Let χ be a set of binary complications in DS, where $\chi = \sum_{i=1}^{5} \chi_i$. χ_i must be selected from one of HYP, NEU, NEP, LIV, RET, and χ_i only takes on clinical class values from $\{\text{low}, \text{high}\}$. For example, if ith complication (χ_i) of kth patient (π_k) is diagnosed negatively (not having the complication of χ_i), the class value becomes zero ($\pi_k(\chi_i) = \text{low}$); otherwise it sets to one ($\pi_k(\chi_i) = \text{high}$) in which it shows that the patient is diagnosed positively (having the ith complication).

For retrieving the conditional rules (if-then pattern) among the complications, we need to make use of some concepts within the associated complications rules. Here, preliminaries for ARs are defined according to a study conducted by Parvez et al.[34] ARs in this article aim to uncover all such relationships between complications from T2DM dataset. TAR of $\{antecedent \Rightarrow consequent\}$ is a representation of finding consequent on the patient visits (which is called basket) followed by the corresponding antecedent on it.

$$\pi = \begin{bmatrix} \text{Patients} & \text{Visits} & \text{Complications Pattern} \\ \pi_1 & \begin{bmatrix} V_{11} & \{\} \\ V_{21} & \{HYP\} \\ V_{31} & \{HYP\} \\ V_{41} & \{HYP, LIV\} \\ V_{51} & \{HYP, LIV, NEU\} \\ V_{61} & \{HYP, LIV, NEU, NEP\} \\ V_{71} & \{HYP, LIV, NEU, NEP\} \\ V_{81} & \{HYP, LIV, NEU, NEP\} \end{bmatrix} \\ \pi_2 & \begin{bmatrix} V_{12} & \{\} \\ V_{22} & \{RET\} \\ V_{32} & \{RET, HYP\} \\ V_{42} & \{RET, HYP, NEU\} \\ V_{52} & \{RET, HYP, NEU, LIV\} \\ V_{62} & \{RET, HYP, NEU, LIV\} \end{bmatrix} \\ \vdots & \\ \pi_i & \begin{bmatrix} V_{1i} \\ V_{2i} \\ \dots \\ V_{T_i i} \end{bmatrix} \\ \vdots & \\ \pi_p & \begin{bmatrix} V_{1p} \\ V_{2p} \\ \dots \\ V_{T_i p} \end{bmatrix} \end{bmatrix}. \quad (1)$$

2.3 | Latent phenotype discovery and time series clustering

The previous work by Yousefi and co-authors in Reference 4 stated that a discovered latent phenotype could be used to capture the temporal risk factors while monitoring the pattern changes in the disease. The latent phenotype for each patient is extracted from the most influential hidden variable identified using the IC* Stepwise algorithm,[4] which uses a DBN framework for inferring model structure and any potential hidden variables simultaneously. A latent variable H is defined to be the expected values for this hidden variable calculated using EM algorithm within the DBN framework. Time series clustering is used on these expected values of the latent variables with DTW to generate clusters of patients as well as identify the "mediod" patient at the center of each cluster. Having discovered the latent phenotype clusters (which is called "H clusters"), it assumes that patients within a cluster share a similar risk factor profile as well as a similar pattern of the occurring complications. In this study, this pattern for each H cluster represents the most frequent ordering pattern of complications, which is associated with the corresponding deep latent phenotype. However, the meaning of the H and its influence on the complications' pattern for each subgroup of patients has remained unclear. In order to understand how the latent phenotype helps to group patients, a combination of the TARs mining and time series clustering is performed in the next section.

2.4 | TAR and AR mining

In this study, ARM is a method that discovers all combination/sequence/set of items (complications), which is called itemsets with the frequency of transactions (referred to support) greater than a predefined minimum threshold based on large itemsets (in the case greater than 0.001). To generate interesting rules with having a confidence greater than the default threshold, it was important to find large itemsets. However, for the sake of simplicity and having a small-sized dataset with sensitive clinical data, a confidence constraint of 25% is chosen. In T2DM dataset, support is regarded as an explicit constraint to identify the outliers. Thus, the minimum constraints must be assigned at a low level. This is because complication rules with predefined constraints that vary from a patient to another patient. Moreover, in the small-sized dataset with the appearance of bias, it is necessary to ascertain that the frequent items do not affect the associations of other items rather than HYP.

T2DM binary complications are representing items of TARs in the shopping basket problem. Itemset of $\{antecedent, consequent\}$ is a representation of the sequence of complications occurs between two visits of $[V_{iv}V_{iz}]$. An itemset I is a transaction that represents a pattern of all associated complications over a patient time series (from the first recorded visit to the last visit). If I is a transaction in database R and a rule is an implication of the form $\{\chi_i \Rightarrow \chi_j\}$ where $\chi_i \subseteq I$, $\chi_j \subseteq I$, and $\chi_i \cap \chi_j \equiv \{\}$. The maximum number of items in I is five ($|I| = 5$), which is equal to the number of binary complications (items). In terms of explaining temporal notation, every two itemsets with a similar complication co-occurrences are treated equivalent and any redundant complication in their intersection is ignored.

In order to analyze antecedent and consequent itemsets, we declare the following definition: $\{\chi_i, \chi_j\}$ is an ordered pair of complications (two-tuple) in which representing the set consisting of both complications χ_i and χ_j with respect to their ordering pattern. $\{\chi_i, \chi_j, \chi_k\}$ is an ordered triple (three-tuple), while $\emptyset$ or $\{\}$ is the empty tuple (zero-tuple). The consequent itemsets may be consisted of more than one item per rules. In the process of pruning/analyzing the rules to pick the most interesting one, our main priority in predictive model for the decision making is based on consequents. Note that despite the fact that the empty set (no complication is diagnosed) is an empty type and one subtype of each of rules antecedent ($\{\} \perp \partial(\chi)$), this is not allowed to be located in consequents. Database $R = \sum_{l=1}^{m=87} R_l$ is retrieved based on the relationships among the complications for all patients within DS. An antecedent of R_l is the left-hand-side subrule ($LHS(R_l)$) in which R_l is a rule of the form $\{\chi_k, \chi_h\}$. Alternatively, consequent is the right-hand-side subrule ($RHS(R_l)$) where $RHS(R_l)$ is a rule of the form $\{\chi_k, \chi_m\}$.

In addition, if there is an "OR" (|) operation among complications in a rule, $\{\chi_i|\chi_j\}$ means that either of complications (χ_i or χ_j) can occur or neither of them ($\{\}$), as shown in Equation (2).

$$\{\chi_i|\chi_j\} \equiv \chi_i|\chi_j|\{\}. \tag{2}$$

In this case study, $\{HYP\}$, $\{RET\}$, and $\{HYP, RET\}$ are equivalent. However, $\{HYP, RET\}$ and $\{RET, HYP\}$ are not considered equivalent as the former ordering is important. Nevertheless, two rules such as $\{NEU, RET, RET\}$ and $\{NEU, RET\}$ are assumed equivalent, whereas $\{NEU|RET\}$ is different (in this study with a ",") as *NEU* should be developed before *RET*, while $\{NEU|RET\} \subseteq \{NEU, RET\}$ with a "|" without consideration of ordering indicates either *RET* or *NEU* or any ordering of both or none of them ($\{\}$) could occur. Two rules of $\{\{NEP, HYP\}, NEU\} \Rightarrow \{RET\}$

and $\{\{NEP, HYP\}, RET\} \Rightarrow \{NEU\}$ are the most interesting rules with the highest confidence and lift, respectively. In addition, an empty antecedent $\{\}$ and two empty antecedents $\{\}\{\}$ in the complication rules represent a patient/transaction with no complication during the first two visits.

Each of the patients in DS can develop any combination of items included in χ where $\mathbb{C}(\pi_i)$ represents all complications/items that patient i has developed during the visits record. A set k-combination of items is a subset of k distinct complications/items chosen from χ (k-itemsets, which is called a subrule). For each patient π_i with a set of visits V_i, the number of k-combinations is equal to the binomial coefficient. $\mathbb{C}(\pi_i) \bowtie \mathbb{C}(\pi_j)$ is the natural join of the relations $\mathbb{C}(\pi_i)$ and $\mathbb{C}(\pi_j)$ where all combinations of tuples in $\mathbb{C}(\pi_i)$ and $\mathbb{C}(\pi_j)$ are equal on their common complications.

Given a set of $\pi_k(\chi_i) \subseteq V_{T_i i} \subset DS$, the power set of I (which is represented by $\wp(V_{T_i i}, k)$) is the set of all complications denoted by $\binom{k}{V_{T_i i}}$, where $k \leq 5$. Thus, a sequence of complications co-occurrence of χ is assumed as a partial ordering on $\partial(\chi)$, where $\partial(\chi, \subseteq)$ is a poset considering χ. The inclusion relation $\subseteq$ defined as a partial ordering on the power set of χ with a reflexive, antisymmetric, and transitive nature. For example, π_i with a pattern of the complications co-occurrence $\mathbb{C}(\pi_i)$ is allocated a subset of $\partial(\chi, \subseteq)$ with $\mathbb{C}(\pi_i) \subseteq \partial(\chi)$. Since for every set $\{LIV, HYP\} \subseteq \{HYP, LIV\}$, hence, $\subseteq$ is reflexive. It is also antisymmetric while there is no repetition of the complications. $\mathbb{C}(\pi_i) \perp \mathbb{C}(\pi_j)$ means that $\mathbb{C}(\pi_i)$ is comparable to $\mathbb{C}(\pi_i)$ if $\mathbb{C}(\pi_i) \preccurlyeq \mathbb{C}(\pi_j)$ where $|\mathbb{C}(\pi_i)| \leq |\mathbb{C}(\pi_j)|$ under set containment. $\mathbb{C}(\pi_i) \vDash \mathbb{C}(\pi_j)$ indicates that $\mathbb{C}(\pi_i)$ entails $\mathbb{C}(\pi_j)$, that is, in every patient, in which $\mathbb{C}(\pi_i)$ and $\mathbb{C}(\pi_j)$ are true. There exists at least one π_j that matches π_i such that $\mathbb{C}(\pi_i) \Leftrightarrow \mathbb{C}(\pi_i)$. A mathematical expression of $\mathbb{C}(\pi_i) \ominus \mathbb{C}(\pi_j)$ means the set of items is exactly one of $\mathbb{C}(\pi_i)$ or $\mathbb{C}(\pi_j)$. Thus, patients ith and jth are developing a similar rules and belonging to a subgroup if their antecedents follows $\mathbb{C}(\pi_i) \ominus \mathbb{C}(\pi_j)$ or $\partial(\text{LHS}(\mathbb{C}(\pi_i) \ominus \mathbb{C}(\pi_j))) \subseteq \mathbb{C}(\pi_i) \cup \mathbb{C}(\pi_j)$. This means that the set of items in the consequent of is belonged to intersection of both rules $\partial(\text{RHS}(\mathbb{C}(\pi_i) \ominus \mathbb{C}(\pi_j))) \subseteq \mathbb{C}(\pi_i) \cap \mathbb{C}(\pi_j)$. As a result, $\{HYP, LIV\} \subseteq \{HYP, LIV, LIV\}$, $\{HYP, LIV, LIV\} \subseteq \{HYP, LIV\}$, and$\{HYP, LIV\} = \{HYP, LIV, LIV\}$. It is also transitive as $\{HYP\} \subseteq \{HYP, LIV\}$ and $\{HYP, LIV\} \subseteq \{HYP, LIV, RET\}$ imply $\{HYP\} \subseteq \{HYP, LIV, RET\}$. Therefore, there is equivalence itemsets of $\{HYP, LIV, LIV\}$ in which $\{HYP, LIV\} : \{HYP, LIV\} \equiv \{HYP, LIV\}$ ($\equiv$ is an equivalence relation). The symbol $\preccurlyeq$ illustrates the relation in any partial set and ordering pattern of occurrence in which each element in its left-hand side is a predecessor of element the right-hand side. The notation $\{HYP, LIV\} \prec \{HYP, LIV, RET\}$ is used to denote $\{HYP, LIV\} \subseteq \{HYP, LIV\}$ and but does not succeed $\{HYP, LIV, RET\} \subseteq \{HYP, LIV\}$ while $\{HYP, LIV\} \not\succ \{HYP, LIV, RET\}$.

2.4.1 | Quality metrics

Support is a fraction of patients containing the itemsets (which is called a transaction or a basket of items). Confidence calculates the probability of occurrence of $\{consequent\}$ given $\{antecedent\}$ is present. Lift is the ratio of confidence to baseline probability of occurrence of $\{consequent\}$. A frequent itemset is an itemset included in at least a significant number of patients. ARM involves the generation of itemsets and TARs. Maximal frequent itemsets represent an itemsets in which none of the corresponding supersets are frequent.

The support measure of itemsets $\mathbb{C}(\pi_i) * (\text{supp}(\mathbb{C}(\pi_i))$ is defined as the proportion of transactions in the dataset containing RHS($\mathbb{C}(\pi_i)$). In particular, an AR of $\partial(\mathbb{C}(\pi_i)) \Rightarrow \partial(\mathbb{C}(\pi_j))$ has a

support of $P(\mathbb{C}(\pi_i)\mathbb{C}(\pi_j))$. The confidence measure of a rule identifies the proportion of transactions with the most interesting/important relationships. In addition, the confidence of a rule is defined as confidence$(\mathbb{C}(\pi_i) \Rightarrow \mathbb{C}(\pi_j) \equiv$ support$(\mathbb{C}(\pi_i) \cup \mathbb{C}(\pi_j)) \equiv$ support$(\mathbb{C}(\pi_j))$, which satisfies Equation (8).

$$\text{support}(\mathbb{C}(\pi_i) \cup \mathbb{C}(\pi_j)) > \sigma, \text{confidence}(\mathbb{C}(\pi_i) \Rightarrow \mathbb{C}(\pi_j)) > \delta \text{lift}\left(\frac{P(\mathbb{C}(\pi_i) \cap \mathbb{C}(\pi_j))}{P(\mathbb{C}(\pi_i)) * P(\mathbb{C}(\pi_j))}\right). \quad (3)$$

Parameters such as σ and δ are the minimum support and confidence, respectively. Instead of using accuracy, efficiency is an appropriate way to evaluate ARs.[34] To obtain the frequents itemsets, first TARs are filtered by using support and confidence. However, they are not able to filter complication rules based on the different dependencies among the rules. For this purpose, a measurement of independence of $\mathbb{C}(\pi_i)$ and $\mathbb{C}(\pi_j)$, which is known as lift. Lift is the deviation of the whole rule support from the expected support under independence given both sides of the rule support. Higher lift values indicate strong associations. Lift of 1 represents $\mathbb{C}(\pi_i)$ and $\mathbb{C}(\pi_j)$ are independent as shown in Equation (9).

$$\text{lift}(\mathbb{C}(\pi_i) \Rightarrow \mathbb{C}(\pi_j)) = \text{support}(\mathbb{C}(\pi_i) \cup \mathbb{C}(\pi_j)) = \text{support}(\mathbb{C}(\pi_i)) * \text{support}(\mathbb{C}(\pi_j)). \quad (4)$$

For example, the probability of developing both *HYP* and *LIV* is associated with the likelihood of developing RET. Confidence of *HYP*, *LIV* implying *RET* is given as the likelihood of developing *HYP*, *LIV*, and also *RET* over the likelihood of developing only *HYP* and *LIV* (see Equation (10)).

$$\text{confidence}(\{HYP, LIV\} \Rightarrow \{RET\}) = \frac{\text{support}(\{HYP, LIV, RET\})}{\text{support}(\{HYP, LIV\})}. \quad (5)$$

The confidence measures whether $\{RET, HYP, NEU, RET\}$ implies *LIV*. This reveals that how likely a given patient develops $\{RET, HYP\}$, *NEU*, *RET*, and *LIV*. In order to find the most interesting itemsets, support ensures that all subrules of the frequent itemsets are also frequent, hence no superset of infrequent itemsets can be frequent. Confidence is very sensitive to the frequency of the consequent. It has been reported that consequents with higher support will produce higher confidence even though there is no association among the antecedent and consequent. Thus, it might not be useful in performing effectively with the existence of bias in dataset DS with a having small number of patients and relatively complications. Confidence measures the strength of the ARs in which the patients that have complication $\mathbb{C}(\pi_i)$ also developed $\mathbb{C}(\pi_j)$ together. There is a number of choices for selecting the filtering measures[35] such as lift, leverage, and coverage, where Lift$(\mathbb{C}(\pi_i) \Rightarrow \mathbb{C}(\pi_j))$ = confidence$(\mathbb{C}(\pi_i) \Rightarrow \mathbb{C}(\pi_j)) \times$ support(χ_j), leverage$(\mathbb{C}(\pi_i) \Rightarrow \mathbb{C}(\pi_j))$ = support$(\mathbb{C}(\pi_i) \Rightarrow \mathbb{C}(\pi_j))$ - (support$(\chi_i) \times$ support(χ_j)), coverage$(\mathbb{C}(\pi_i) \Rightarrow \mathbb{C}(\pi_j))$ = support$(\mathbb{C}(\pi_i))$. In T2DM dataset, there is a strong association (indicated by the highest lift) among the complications, which shows the likelihood of the complication being developed relative to its general developing rate, given that the patient developed other complications. For instance, the conditional probability of developing both HYP and LIV in are associated with the likelihood of the patient developing RET. There is a strong association (indicated by the highest lift) among the complications, which shows the likelihood of the complication being developed relative to its general developing rate, given that the patient developed other complications. For example, the conditional probability of a patient developing both HYP and LIV is associated with the likelihood of the patient developing RET. Whereas coverage filters the rules mostly based on

their antecedents. This is opposite to this article preferences where the consequents (the complications occur in the future visits) have been considered as the most revealing itemsets in the decision making and prediction process. Similar to lift, conviction metric assesses the likelihood of the appearance of an antecedent in which the corresponding consequent is not likely to occur.

Overall, a question still remains to answer whether it could be possible to trust these metrics by the user-defined thresholds. In particular, there are many challenges to find the most interesting rules[36] only by relying on TARs. Nevertheless, most of the previously mentioned metrics in this study are mainly depended on the support and frequency. In a small-sized dataset like DS, where there is a different imbalance ratio for each item (complication), bias, and latent factors, it may not be beneficial if is only trust on the obtained itemsets resulted by using support, confidence, and lift.

Moreover, there are some itemsets that are called frequent itemsets, while their occurrence exceeds the threshold in the database. In order to generate interesting rules, one could come across many frequent itemsets with minimal confidence. In the other words, by applying a rigid constraint with having bias in data, the final itemsets can be identified as interesting itemsets wrongly. This is because interestingness is only based on the association of HYP with the items, not the relationships among the items themselves. An item like HYP with a high occurrence rate can affect the way how other items are associated with each other. To avoid the above issue in a small-sized dataset, we tend to discover all types of associations regardless of effect of bias (eg, HYP) and focus mostly on the relaxed or flexible filtering metrics.

It does not seem to be possible to only rely on lift as it may not be trustworthy enough and unable to perform effectively with the existence of bias in the incomplete data. Lift suffers from having nonfixed range of variables. It only assesses the dependency and correlation of the items without taking into consideration the importance of the cause and effect relationships among antecedents and consequents. Similar to the issue related to support and confident, lift is susceptible to infrequent items with a relatively low probability complication rules that can be ranked wrongly as the most interesting itemsets. Although having a very low or minimal constraints to be applied on the quality metrics, it does not eliminate the above issue, which is caused by generating all possible permutations of complications for all transactions as an non-optimal option. This is because, Tables 3 and 4 contain many different antecedents and consequents, which increase the database size exponentially based on the number of items. It also leads to generating large number of uninteresting distances among many small rules despite the previously chosen optimal/minimal threshold for support and confidence. In this situation, neither clustering nor ARM methodology perform effectively and can be even worse and problematic in a sparse dataset (such as T2DM). In conclusion, for making a better decision, the uninteresting rules needs to be reduced at another level which is addressed in the next section.

2.5 | Methods

This section explains the methods to find explainable subgroups of patients. Our recent work in References 3 and 4 has suggested that the identification of a "phenotype" can be used to separate patients into meaningful subgroups with the consideration of the T2DM risk factor and complication relationships. Here we, first, identify an informative pattern based on latent

TABLE 3 Database R of the associated rules with the complications generated using TARs

Rule[a]	Antecedent		Consequent	Interesting itemsets (objects) from D[b]	Support	Confidence	Lift
1	{ }	⇒	{HYP,RET}	3,14,23,27,28,33,38,41	≥ 0.001	≥ 0.001	1.00
2	{ }	⇒	{RET,HYP}	3,14,23,27,28,33,38,41	0.01	0.01	1.00
3	{ }	⇒	{NEU,HYP}	5,13,21,26,28,31,38,41	0.01	0.01	1.00
4	{ }	⇒	{LIV,HYP}	6,24,30,31,33,35,36,37,39,40	0.02	0.02	1.00
5	{ }	⇒	{NEP,HYP}	2,13,14,20,26,27,30,38,41	0.03	0.03	1.00
6	{ }	⇒	{ }{ }	9	0.02	0.02	1.00
7	{ }	⇒	{NEP}	2,11,13,14,16,17,18,20,25,	0.11	0.11	1.00
				29,30,32,35,36,37,38,41			
8	{ }	⇒	{NEU}	5,7,13,15,16,19,21,25,26,28,	0.16	0.16	1.00
				29,31,34,35,37,38,39,40,41			
9	{ }	⇒	{RET}	3,4,8,14,15,17,22,23,25,27,28,	0.15	0.15	1.00
				32,33,34,38,41			
10	{ }	⇒	{LIV}	6,12,18,19,22,24,29,30,31,32,	0.15	0.15	1.00
				33,34,35,36,37,39,40			
11	{ }	⇒	{HYP}	2,3,4,5,6,10,13,14,20,21,23,	0.86	0.86	1.00
				24,26,27,28,30,31,33,38,41			
12	{NEU,HYP}	⇒	{NEU}	13,26,38,41	0.01	0.27	1.71
13	{NEU}	⇒	{NEP,HYP}	13,26,38,41	0.01	0.05	1.71
14	{NEP,HYP}	⇒	{RET}	14,27,38,41	0.01	0.27	1.79
15	{RET}	⇒	{NEP,HYP}	14,27,38,41	0.01	0.05	1.79
16	{ }{ }	⇒	{RET}	3,4,8,14,15,17,22,23,25,27,	0.01	0.22	1.46
				28,32,33,34,38,41			
17	{RET}	⇒	{ }{ }	3,4,8,14,15,17,22,23,25,27,	0.01	0.03	1.46
				28,32,33,34,38,41			
18	{ }{ }	⇒	{HYP}	2,3,4,5,6,10,13,14,20,21,23,	0.02	0.78	0.90
				24,26,27,28,30,31,33,38,41			
19	{HYP}	⇒	{ }{ }	2,3,4,5,6,10,13,14,20,21,23,	0.02	0.02	0.90
				24,26,27,28,30,31,33,38,41			
20	{NEP}	⇒	{NEU}	13,16,25,26,29,38,41	0.02	0.19	1.17
21	{NEU}	⇒	{NEU}	13,16,25,26,29,38,41	0.02	0.13	1.17
22	{NEP}	⇒	{RET}	14,17,25,27,32,38,41	0.02	0.14	0.92
23	{RET}	⇒	{NEP}	14,17,25,27,32,38,41	0.02	0.10	0.92
24	{NEP}	⇒	{LIV}	18,29,30,32,36,37	0.04	0.37	2.49

(Continues)

TABLE 3 (Continued)

Rule[a]	Antecedent		Consequent	Interesting itemsets (objects) from D[b]	Support	Confidence	Lift
25	{LIV}	⇒	{NEP}	18,29,30,32,36,37	0.04	0.27	2.49
26	{NEP}	⇒	{HYP}	2,13,14,20,26,27,30,38,41	0.10	0.93	1.08
27	{HYP}	⇒	{NEP}	2,13,14,20,26,27,30,38,41	0.10	0.12	1.08
28	{NEU}	⇒	{RET}	15,25,28,34,38,39,40,41	0.04	0.25	1.64
29	{RET}	⇒	{NEU}	15,25,28,34,38,39,40,41	0.04	0.26	1.64
30	{NEU}	⇒	{LIV}	19,29,31,34,35,37,39,40	0.02	0.11	0.73
31	{LIV}	⇒	{NEU}	19,29,31,34,35,37,39,40	0.02	0.12	0.73
32	{NEU}	⇒	{HYP}	5,13,21,26,28,31,35,	0.12	0.78	0.91
				37,38,39,40,41			
33	{HYP}	⇒	{NEU}	5,13,21,26,28,31,35,37,	0.12	0.14	0.91
				38,39,40,41			
34	{RET}	⇒	{LIV}	22,32,34,36,39,40	0.03	0.20	1.31
35	{LIV}	⇒	{RET}	22,32,34,36,39,40	0.03	0.20	1.31
36	{RET}	⇒	{HYP}	3,14,23,27,28,33,36,38,41	0.12	0.79	0.91
37	{HYP}	⇒	{RET}	3,14,23,27,28,33,36,38,41	0.12	0.14	0.91
38	{LIV}	⇒	{HYP}	6,24,30,31,33,35,36,37,39,40	0.14	0.92	1.06

[a]This table shows rule number 1 to 37.

[b]Identification of a set of objects which follows any combinations of the itemsets in D.

variables, which we call a "latent phenotype." This is then used to group patients and captured the complexities/homogeneity of the risk factors/complications over time.

Studies relating to enhancing the interpretability of latent variables along with a significant improvement in the prediction performance have been relatively scant. There is no study focusing on utilizing ARM in the underlying patterns of temporal complications rules (which we note as "complication rules") in order to explain the latent variable behavior. Since the clinical model can have serious consequences, it is imperative to better understand the associated complication rules in trustable/interpretable patients models. These models are relatively complex; however, it can be accurately modeled by using data mining techniques (including both descriptive and predictive strategies). We further extend this idea by using a hybrid methodology of TARs, ARM, time series clustering, statistical, Bayesian structure modeling, and predictive analysis in order to find explainable subgroups of patients with more personalised prediction. To implement the model, the associated complication rules are mined to assess the occurrence likelihood of binary complications in relation to the rest of complications associated with a prediagnosed T2DM patient. For example, to find out whether the increasing prevalence of HYP has been accompanied by an increase in the NEU or patients with NEP are also diagnosed by LIV. Then, TARs are chosen according to the needs of the study to discover underlying relationships among the complications.

Similarly, pattern mining and sequence discovery are performed to explain and highlight the potential usefulness of the complication rules with a deeper understanding of their causal structure within the clinical data. With ARM we are interested in the absolute number of patients that

TABLE 4 Database R of the associated rules with the complications generated using TARs

Rule[a]	LHS		RHS	Interesting itemsets (objects) from D[b]	Support	Confidence	Lift
39	{HYP}	⇒	{LIV}	6,24,30,31,33,35,36,37,39,40	0.14	0.16	1.06
40	{{NEP,HYP},NEU}	⇒	{RET}	28,38,41	0.01	1.00	6.57
41	{{NEP,HYP},RET}	⇒	{NEU}	28,38,41	0.01	1.00	6.27
42	{NEU,RET}	⇒	{NEP,HYP}	28,38,41	0.01	0.19	6.84
43	{NEP,NEU}	⇒	{RET}	25,38,41	0.01	0.25	1.64
44	{NEP,RET}	⇒	{NEU}	25,38,41	0.01	0.33	2.09
45	{NEU,RET}	⇒	{NEP}	25,38,41	0.01	0.13	1.17
46	{NEP,NEU}	⇒	{LIV}	29,35,37	0.01	0.25	1.67
47	{LIV,NEP}	⇒	{NEU}	29,35,37	0.01	0.13	0.78
48	{LIV,NEU}	⇒	{NEP}	29,35,37	0.01	0.29	2.66
49	{NEP,NEU}	⇒	{HYP}	26,35,37,38,41	0.02	0.88	1.01
50	{HYP,NEP}	⇒	{NEU}	26,35,37,38,41	0.02	0.18	1.10
51	{HYP,NEU}	⇒	{NEP}	26,35,37,38,41	0.02	0.14	1.31
52	{NEP,RET}	⇒	{LIV}	32,36	≥ 0.001	0.17	1.11
53	{LIV,NEP}	⇒	{RET}	32,36	≥ 0.001	0.06	0.41
54	{LIV,RET}	⇒	{NEP}	32,36	≥ 0.001	0.08	0.78
55	{NEP,RET}	⇒	{HYP}	27,36,38,41	0.01	0.67	0.77
56	{HYP,NEP}	⇒	{RET}	27,36,38,41	0.01	0.10	0.66
57	{HYP,RET}	⇒	{NEP}	27,36,38,41	0.01	0.08	0.78
58	{LIV,NEP}	⇒	{HYP}	30,35,36,37	0.04	0.94	1.09
59	{HYP,NEP}	⇒	{LIV}	30,35,36,37	0.04	0.38	2.51
60	{HYP,LIV}	⇒	{NEP}	30,35,36,37	0.04	0.27	2.54
61	{NEU,RET}	⇒	{LIV}	34,39,40	0.01	0.13	0.84
62	{LIV,NEU}	⇒	{RET}	34,39,40	0.01	0.29	1.88
63	{LIV,RET}	⇒	{NEU}	34,39,40	0.01	0.17	1.04
64	{NEU,RET}	⇒	{HYP}	28,38,39,40,41	0.03	0.75	0.87
65	{HYP,NEU}	⇒	{RET}	28,38,39,40,41	0.03	0.24	1.58
66	{HYP,RET}	⇒	{NEU}	28,38,39,40,41	0.03	0.25	1.57
67	{LIV,NEU}	⇒	{HYP}	31,35,37,39,40	0.02	0.86	0.99
68	{HYP,NEU}	⇒	{LIV}	31,35,37,39,40	0.02	0.12	0.80
69	{HYP,LIV}	⇒	{NEU}	31,35,37,39,40	0.02	0.11	0.68
70	{LIV,RET}	⇒	{HYP}	33,36,39,40	0.03	1.00	1.16
71	{HYP,RET}	⇒	{LIV}	33,36,39,40	0.03	0.25	1.67

(Continues)

TABLE 4 (Continued)

Rule[a]	LHS		RHS	Interesting itemsets (objects) from D[b]	Support	Confidence	Lift
72	{HYP,LIV}	⇒	{RET}	33,36,39,40	0.03	0.22	1.43
73	{NEP,NEU,RET}	⇒	{HYP}	38,41	≥ 0.001	0.50	0.58
74	{HYP,NEP,NEU}	⇒	{RET}	38,41	≥ 0.001	0.14	0.94
75	{HYP,NEP,RET}	⇒	{NEU}	38,41	≥ 0.001	0.25	1.57
76	{HYP,NEU,RET}	⇒	{NEP}	38,41	≥ 0.001	0.08	0.78
77	{LIV,NEP,NEU}	⇒	{HYP}	37	0.01	1.00	1.16
78	{HYP,NEP,NEU}	⇒	{LIV}	37	0.01	0.29	1.91
79	{HYP,LIV,NEP}	⇒	{NEU}	37	0.01	0.13	0.84
80	{HYP,LIV,NEU}	⇒	{NEP}	37	0.01	0.33	3.11
81	{LIV,NEP,RET}	⇒	{HYP}	36	≥ 0.001	1.00	1.16
82	{HYP,NEP,RET}	⇒	{LIV}	36	≥ 0.001	0.25	1.67
83	{HYP,LIV,NEP}	⇒	{RET}	36	≥ 0.001	0.07	0.44
84	{HYP,LIV,RET}	⇒	{NEP}	36	≥ 0.001	0.08	0.78
85	{LIV,NEU,RET}	⇒	{HYP}	39	0.01	1.00	1.16
86	{HYP,NEU,RET}	⇒	{LIV}	39	0.01	0.17	1.11
87	{HYP,LIV,NEU}	⇒	{RET}	39	0.01	0.33	2.19

[a]This table shows rule number 39 to 87.
[b]Identification of a set of objects which follows any combinations of the itemsets in D.

contain a particular set of complications. By utilizing TARs, given many patterns of complication rules (itemsets), we attempt to find which itemsets, that belong to a patient, predict another complication for the patient. Thus, we use a postprocessing approach (which is called minimum coverage itemsets [MCI]) to prune the rules to the most important ones and to find the most useful distances in order to obtain meaningful clustering outcomes. We then attempt to explain and validate these groups through the integration of TARs combined with time series clustering. Figure 1 illustrates the overall process that includes: hidden variable discovery that is used to identify the latent phenotype and, in turn, generates the latent phenotype clusters (H cluster), TARs clusters, and finally comparison and validation strategies (involving Jaccard distance metrics and sensitivity analysis).

The proposed hybrid methodology to find explainable subgroups of patient and interpret the latent variable by personalizing diabetic patients in precision medicine is demonstrated as a multiple-stage process in Figure 1, which is labeled and explained as follows:

1. Data discretization and preparation are employed to generate the original T2DM dataset (DS) in the preprocessing approach.
2. For each patient, an informative pattern (latent phenotype) is identified based on the latent variable discovery approach using DBNs and IC* Stepwise algorithm, latent phenotype.
3. DTW finds dissimilarities between the discovered latent phenotypes and captures the complexities/homogeneity of the risk factors/complications over time.

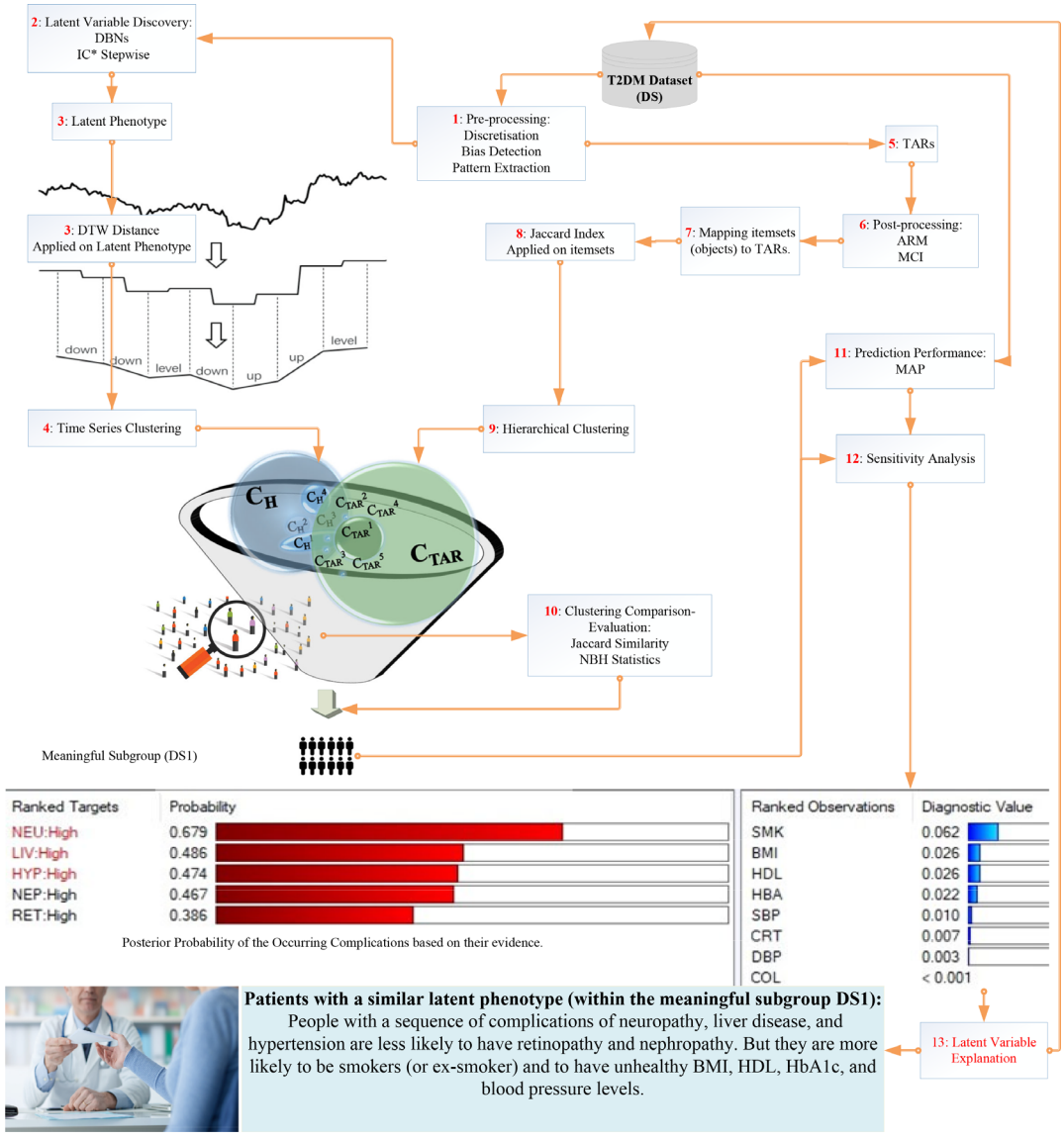

FIGURE 1 The proposed hybrid methodology to find explainable subgroups of patient to interpret the latent variable by personalizing diabetic patients in precision medicine [Color figure can be viewed at wileyonlinelibrary.com]

4. Time series clustering based on DTW distance is applied to stratify patients into four latent phenotype clusters.
5. The multiple binary complications, as items from the preprocessed dataset DS, are extracted and mined to retrieve the temporal patterns of items for all patients.
6. TARs are applied on the obtained patterns from DS and generate Tables 3 and 4. These rules consists of (87×2) subrules (including 87 antecedents and 87 consequents).
7. A postprocessing ARM methodology is applied to the complication rules where metrics such as support and confidence with predefined soft thresholds filtered frequent rules. These

constraints are strengthened in which lift of the frequent rules must come through the highest lift boundary. Then another algorithm (which we called MCI) generates least itemsets in D covering the interesting rules from R. MCI locates alternative optimal combinations of the subrules in which the number of repetitive items can be reduced. As a result, dataset D is generated based on the most important rules.

8. All rules in R are mapped to the relevant objects/itemsets in D based on the implications of the antecedents and consequents. Jaccard index measured the objects to clusters the complication rules (TAR clusters).
9. By using agglomerative clustering, objects are grouped in five groups. Patients are assigned to the corresponding cluster based upon their associated unique pattern of complications.
10. Jaccard Similarity and more statistical methods is applied to compare and validate the discovered clusters to find meaningful subgroups of patients from the intersection of H and TAR clusters.
11. Prediction performance of the discovered meaningful subgroup (DS1) as a subset is compared to DS.
12. Sensitivity analysis is utilized to assess DS1 and analyze its prediction performance comparing to DS.
13. The latent variable is explained for patients with a similar pattern of TARs and latent phenotype.

2.6 | Latent phenotype discovery and time series clustering

Previously, we stated that a discovered latent phenotype could be used to capture the temporal risk factors while monitoring the pattern changes in the disease. The latent phenotype for each patient is extracted from the most influential hidden variable identified using the IC* Stepwise algorithm,[4] which uses a DBN framework for inferring model structure and any potential hidden variables simultaneously. We define H to be the expected values for this hidden variable calculated using expectation-maximization (EM) algorithm[37] within the DBN framework. Time series clustering is used on these expected values of the latent variables with DTW to generate clusters of patients as well as identify the "mediod" patient at the center of each cluster. Having discovered the latent phenotype clusters (which we call "H clusters"), we assume that patients within a cluster share a similar risk factor profile as well as a similar pattern of the occurring complications. In this study, this pattern for each H cluster represents the most frequent ordering pattern of complications, which is associated with the corresponding deep latent phenotype. However, the meaning of the H and its influence on the complications' pattern for each subgroup of patients has remained unclear. In order to understand how the latent phenotype helps to group patients, a combination of the TARs mining and time series clustering is performed in the next section.

2.7 | TARs and AR mining

In this study, ARM is a method that discovers all combination/sequence/set of items (complications), which is called itemsets with the frequency of transactions (referred to support) greater than a predefined minimum threshold based on large itemsets (in our case greater than 0.001). To generate interesting rules with having a confidence greater than the default threshold, it was

important to find large itemsets. However, for the sake of simplicity and having a small-sized dataset with sensitive clinical data, we choose a confidence constraint of 25%. In T2DM dataset, support is regarded as an explicit constraint to identify the outliers. Thus, the minimum constraints must be assigned at a low level. This is because complication rules with predefined constraints which vary from a patient to another patient. Moreover, in the small-sized dataset with the appearance of bias, we need to ascertain that the frequent items do not affect the associations of other items rather than HYP.

In order to find the most interesting itemsets, support ensures that all subrules of the frequent itemsets are also frequent, hence no superset of infrequent itemsets can be frequent. Confidence is very sensitive to the frequency of the consequent. It has been reported that consequents with higher support will produce higher confidence even though there is no association among the antecedent and consequent. Thus, it might not be useful in performing effectively with the existence of bias in dataset DS with a having small number of patients and relatively complications. Confidence measures the strength of the ARs in which the patients that have complication χ_i also developed χ_j together. We have a number of choices for selecting the filtering measures[35] such as lift, leverage, and coverage, where Lift($\chi_i \Rightarrow \chi_j$) = confidence($\chi_i \Rightarrow \chi_j$) × support($\chi_j$), Leverage($\chi_i \Rightarrow \chi_j$) = support($\chi_i \Rightarrow \chi_j$) - (support($\chi_i$) × support($\chi_j$)), Coverage($\chi_i \Rightarrow \chi_j$) = support($\chi_i$). In T2DM dataset, there is a strong association (indicated by the highest lift) among the complications, which shows the likelihood of the complication being developed relative to its general developing rate, given that the patient developed other complications. For instance, the conditional probability of a patient developing both HYP and LIV is associated with the likelihood of the patient developing RET. There is a strong association (indicated by the highest lift) among the complications, which shows the likelihood of the complication being developed relative to its general developing rate, given that the patient developed other complications. For example, the conditional probability of a patient developing both HYP and LIV is associated with the likelihood of the patient developing RET. Whereas coverage filters the rules mostly based on their antecedents. This opposite the present paper preferences where the consequents (the complications occur in the future visits) have been considered as the most revealing itemsets in the decision making and prediction process. Similar to lift, conviction metric assesses the likelihood of the appearance of an antecedent without the corresponding consequent.

Nevertheless, a question still remains to answer as if we can trust these metrics by the user-defined thresholds. In particular, there are many challenges to find the most interesting rules[36] only based on the TARs and its constraints. For example, all of the previously mentioned metrics in this article only depend on the support and frequency. In a small-sized dataset like DS, where there is a different imbalance ratio for each item (complication), bias, and latent factors, it may not be beneficial if we only rely on the obtained itemsets resulted by using support, confidence, and lift. Unfortunately, there are some itemsets that are called frequent itemsets while their occurrence exceeds the threshold in the database.

Moreover, in order to generate interesting rules, we come across many frequent itemsets with minimal confidence. In the other words, by applying a rigid constraint with having bias in data, the final itemsets can be identified as interesting itemsets wrongly. This is because interestingness is only based on the association of HYP with the items, not the relationships among the items themselves. An item like HYP with a high occurrence rate can affect the way how other items are associated with each other. To avoid this issue in a small-sized dataset, we need to find all types of associations regardless of effect of HYP and relaxed or flexible filtering metrics.

Having said that, if we only rely on lift, it might not be trustworthy enough and unable to perform effectively with the existence of bias in the incomplete data. Lift suffers from having nonfixed

range of variables. It only assesses the dependency and correlation of the items without taking into consideration the importance of the cause and effect relationships among antecedents and consequents. Similar to the issue related to support and confident, lift is susceptible to infrequent items with a relatively low probability complication rules that can be ranked wrongly as the most interesting itemsets.

Although having a very low minimum could eliminate the above issue, generating all possible permutations of complications for all transactions is not an optimal option. This is because, Tables 3 and 4 contain many different antecedents and consequents, which increase the database size exponentially based on the number of items. It also leads to generating large number of uninteresting distances among many small rules despite the previously chosen optimal minimum threshold for support and confidence. In this situation, neither clustering nor ARM methodology perform effectively and can be even worse and problematic in a sparse dataset (such as T2DM). In conclusion, for making a better decision, we need to reduce uninteresting rules at another level which is addressed in the next section.

2.8 | Interesting itemsets in complication rules using minimal coverage itemsets algorithm

Thus far, metrics such as support, confidence, and lift were used to identify the most interesting rules. However, we argued that there might still be many uninteresting/uninformative rules remained, which would be challenging to interpret due to the complex nature of the associated complications. To overcome this, we intend to discover the minimum coverage of rules by using MCI, which is motivated by a variation on the proposed methodology conducted by Liu et al to enhance k-means clustering in Reference 38. The identified sequence of complications is mined to extract the useful rules and detect an appropriate ordering of the complications as a minimum coverage of set, which is called itemsets. As can be seen in Figure 2 in the left hand side, temporal patterns of the complications co-occurrences are retrieved from DS. The database is mined to include the temporal relationships among the multiple complications into their associated rules. We used TARs on the temporal co-occurrence pattern of the complications to obtain 87 rules. Then, MCI analyze subrules (antecedents and consequents) as input and produces the minimum coverage itemsets (41 objects found) as output in Table 5. A minimum number of aggregated subrules are produced based on their uniqueness/intersection while covering the most frequent/interesting rules. We then refer to database R to find the related objects of the relevant associated rules once all of the objects are identified and mapped to the rules in Tables 3 and 4. By choosing the objects in the instead of rules, a minimum overlap among the data points is produced, this cannot be achieved using only lift. Thus, distance among the objects represents higher quality data points with less repetition of unimportant rules as the clustering input. In addition to this, MCI helps in achieving the optimal number of meaningful subgroups in the clustering method.

2.9 | Combined methodology of ARM and clustering

In this article, a hybrid methodology of TARs mining and clustering attempts to validate and give meaning to the H clusters. We also proposed MCI algorithm to find minimum rules set as the most interesting itemsets from the temporal complications within the T2DM. Furthermore, the

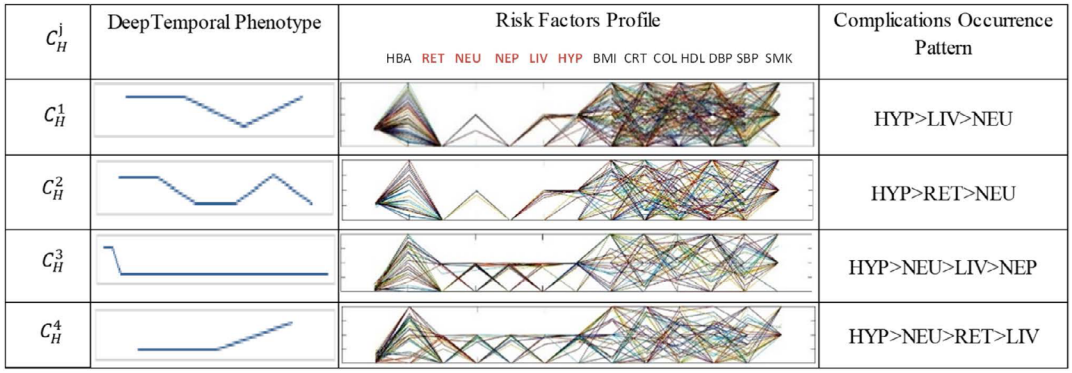

FIGURE 2 The deep latent phenotype for the Hidden variable clusters, the corresponding risk factor profiles, and the most frequent ordering pattern of the complications [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 5 The frequent itemsets are generated in dataset D based on the rules in generated using TARs

Objects ID	Interesting itemsets	Objects ID	Interesting itemsets
1	{ }	21	{NEU|HYP}
2	{NEP,HYP}	22	{RET}
3	{HYP,RET}	23	{HYP|RET}
4	{RET,HYP}	24	{HYP|LIV}
5	{NEU,HYP}	25	{NEP|NEU|RET}
6	{LIV,HYP}	26	{HYP|NEP|NEU}
7	{NEU}	27	{HYP|NEP|RET}
8	{RET}	28	{HYP|NEU|RET}
9	{ }{ }	29	{LIV|NEP|NEU}
10	{HYP}	30	{HYP|LIV|NEP}
11	{NEP}	31	{HYP|LIV|NEU}
12	{LIV}	32	{LIV|NEP|RET}
13	{{NEP,HYP}|NEU}	33	{HYP|LIV|RET}
14	{{NEP,HYP}|RET}	34	{LIV|NEU|RET}
15	{NEU|RET}	35	{LIV|HYP|NEU}
16	{NEU}	36	{HYP|LIV|RET|NEP}
17	{NEP|RET}	37	{HYP|LIV|NEU|NEP}
18	{LIV|NEP}	38	{NEU|RET|NEP|HYP}
19	{LIV|NEU}	39	{NEU|HYP|RET|LIV}
20	{HYP|NEP}	40	{LIV|HYP|NEU|RET}
21	{HYP|NEU}	41	{NEP,HYP}RET,NEU}

Note: Each of resulted itemsets of the applied MCI is identified by unique objects.

meaningful rules after applying the MCI based on the aggregation of only the most frequent and important antecedents and consequents are utilized in Table 5. The issue of discovering the frequent itemsets (ARM) differs from the similarity search in the clustering method. Instead of using all rules as a clustering input, we only use the significant itemsets (objects) in the hierarchical clustering method. The clustering method allocates objects as itemsets in such a way that objects in the same subgroup coincide with each other subgroups, based upon Jaccard index. The Jaccard distance between two itemsets (objects) (I_i and I_j) is calculated by the number of similar itemsets between I_i and I_j over all unique itemsets in both itemsets. For a set of m itemsets, there is overall of $m(m-1)/2$ distances that can be used to cluster the objects and further patient subgroups. Therefore, clustering tries to find objects that have a significant fraction of their associated pattern of complications in common; the absolute number of those objects is not of interest. Thus, patients are assigned to a cluster if their patterns of complications match the most frequent object/itemsets in the corresponding cluster. In other words, patients that have been diagnosed with a similar occurring pattern of complications over time (corresponding frequent itemsets) are gathered in one cluster. In the next part, we attempt to measure the distance among the objects. The proposed MCI procedure to discover the most interesting itemsets (which we call objects/clustering data points) is illustrated below and shown in Figure 2.

1. Input: R in Tables 3 and 4, considering minimum support σ and minimum confidence δ thresholds.
2. Output: Interesting itemsets (objects) in Table 5.
3. Rule R is of the form (X, Y) which represents $\{X\} \Rightarrow \{Y\}$.
4. lift$(R_l) \geq$ MAX lift(R) conf$(R_l) \geq \delta$
5. OverlapRate $\leftarrow 0$, MinOverlapRate $\leftarrow 0$
6. $r \leftarrow \emptyset$, $r_k \leftarrow \emptyset$
7. For each R_l SUBSET R
8. If lift$(R_l) \geq$ MAX lift(R)
9. If supp$(R_l) \geq \sigma$ AND conf$(R_l) \geq \delta$
10. $R \leftarrow R$EXCLUDES R_l
11. $R \leftarrow R$EXCLUDES R_l
12. $D \leftarrow \{LHS(R_l), RHS(R_l)\} \cup D$
13. $r(R_l) \leftarrow \{LHS(R_l), RHS(R_l)\}$
14. $D \leftarrow \sum_{l=1}^{m=87} r(R_l)$
15. For each itemsets $r(R_l)$ SUBSET of D
16. $r(R_l) \leftarrow \{LHS(R_l)\} \cup r(R_l)$
17. $r(R_l) \leftarrow \{RHS(R_l)\} \cup r(R_l)$
18. $r(R_l) \leftarrow R \subseteq$ POWERSET $r(R_l)$
19. $Objects(r(R_l)) \leftarrow D \subseteq$ POWERSET $r(R_l)$
20. For Each $r(k)$ and $r(h)$
21. If $r(k) \subseteq r$ and $r(h) \subseteq r$ and $r(k) \neq r(h)$
22. OverlapRate $\leftarrow$ COUNT $r(k) \cap r(h)$
23. If OverlapRate $\leq$ *MinOverlap*
24. MCI $\leftarrow \{(r(k) \cup r(h)) \cup\}$ MCI
25. Else
26. MCI $\leftarrow \{(r(k) \cap r(h)) \cup\}$ MCI
27. RETURN MCI.

2.10 | Jaccard index and TAR clusters

To handle a large number of rules, we grouped the rules using agglomerative hierarchical clustering.[39] The combined use of unsupervised learning is motivated by Hahsler et al' research conducted in Reference 39, which introduced a distance-based clustering of ARs. However, we adopted a different method for a more in-depth analysis of the correlation between rules to find dissimilarities (distances). In the clustering literature, the frequent rule sets as a fundamental concept of TARs have enhanced the overall clustering methodology.[38] Agglomerative hierarchical clustering is employed to group the associated rules into more informative rules or the so-called itemsets. Accordingly, the Jaccard index is applied to create distances between itemsets. Comparisons between the two patients from two different clusters are made using unrelated rules on their associated complications. Table 6 represents the elements of clusters, which represented as objects. Finally, patients are allocated to a cluster based on the object meeting the rules belong to the itemsets within the corresponding cluster. We cluster patients based on their TARs clusters (C_{TAR}), where each cluster shares a similar complications sequence (co-occurrence pattern of complications). For comparing two different sequences of the complications (i and j) in the hierarchical clustering of the itemsets of I_i and I_j, we use Jaccard index (Jaccard(I_i, I_j)) and Jaccard distance ($d_{i,j}$) in Equation (6).

$$\text{Jaccard}(I_i, I_j) = \frac{|I_i \cap I_j|}{|I_i \cup I_j|}, \qquad d_{i,j} = 1 - \text{Jaccard}(I_i, I_j). \tag{6}$$

2.11 | Clustering comparison and validation strategies

We intend to ascertain the usefulness trustworthiness of the TAR cluster in understanding the underlying disease as well as being a reliable source to validate the latent phenotype. Internal validation is applied to assess the validity of the C_{TAR} through the use of the information contained within the given database of complication rules. In order to remove uninformative and rare rules from the database, the most infrequent itemsets are ignored. Then the dissimilarities (distances) among TAR clusters filter out the discovered meaningful rules. For example, rules with a high lift and confidence score are selected. Thus, the number of TARs is reduced to a manageable number while concentrating the most interesting rules. For external validation, the H clusters are assessed based upon another data source (TAR clusters). Jaccard similarity is applied to calculate the proportion of the overlapped patients for each pair of the latent phenotype and TAR

TABLE 6 Clusters of the frequent itemsets identified by groups of objects in the associated interesting itemsets from Tables 3-5

TAR clusters	Elements of cluster (interesting itemsets/objects)
C^1_{TAR}	10,13,2,20,21,24,26,30,31,5,6,38
C^2_{TAR}	11,12,16,18,19,29,7,9
C^3_{TAR}	14,23,27,28,3,33,38,4,41
C^4_{TAR}	15,17,22,25,32,34,8
C^5_{TAR}	35,36,37,39,40

clusters. Although the Jaccard similarity seems useful to measure the overlap between two clusters, the resulting value is not able to indicate the likelihood of the observed overlap. As a result, normal approximation for the binomial approximation of the hypergeometric distribution (NBH) metric[40] is utilized to evaluate the probability of observing an overlap between each pair of clusters from C_H and C_{TAR}. A low value (probability) indicated that the chance of observing a given overlap was highly unlikely to occur by random chance. For a given C^i_{TAR} of size s_i (where (i indicates the cluster number) compared to a C^j_H of size k_j (where (j indicates the cluster number), the probable score of the overlap occurring randomly can be modeled using a binomial distribution, as shown in Equation (7).[40]

$$Pr(\text{observing x from group}j) = \binom{k_j}{x} p^x q^{k_j}, \tag{7}$$

$$x = \text{Jaccard}(C^j_H, C^i_{\text{TAR}}), \quad n = |C_H \cup C_{\text{TAR}}|, \quad s_i = |C^i_{\text{TAR}}|, \quad k_j = |C^j_H|, \quad p_i = \frac{s_i}{n}, q = 1 - p,$$

where n is the number of patients in the union of all of the C^i_{TAR} and all of the C^j_H. If both n and npq are large, a binomial distribution can be approximated by a normal distribution. For example, obtaining a very low NBH probability represents there is a considerable/significant overlapped rate between two clusters from different data sources. We illustrate the finding and more explanation regarding the NBH probability in the following section.

3 | EXPERIMENTAL RESULTS

In this section, we validate TAR clusters and compare them with the latent phenotype to understand whether the latent phenotype reduces some uncertainty caused by the complex relationships among the temporal complications. In Table 5, the most frequent and interesting itemsets (ordering pattern of complications) are identified by an object. In order to quantify a distance between two heterogeneous rules, one solution could be to use cluster rules based on their features (support, confidence, and lift). However, these measures can only capture the interaction of rules on the data and characterize only a single rule. Thus, more in-depth analysis of the correlation/causation between rules is possible when we find dissimilarities among the itemsets. Agglomerative hierarchical clustering is employed in order to build homogeneous groups of objects.

ARs are grouped according to the descriptors (itemsets or objects), as shown in Table 6. On the other hand, they are not grouped according to their coverage, as explained in MCI algorithm. Each of the patients within DS that have been diagnosed with the a similar occurring pattern of complications (the corresponding frequent itemsets) are gathered in one cluster. The distances among the frequent itemsets are aggregated for two patients within a cluster by using Jaccard distance, which are applied to the group of the object associated with the corresponding pattern.

3.1 | Discovered clusters

We obtained the initial five clusters of the TARs as $C_{\text{TAR}} = \{ C^1_{\text{TAR}}, C^2_{\text{TAR}}, C^3_{\text{TAR}}, C^4_{\text{TAR}}, C^5_{\text{TAR}}\}$, according to the dissimilarities between associated rules (itemsets) using Jaccard dissimilarity.

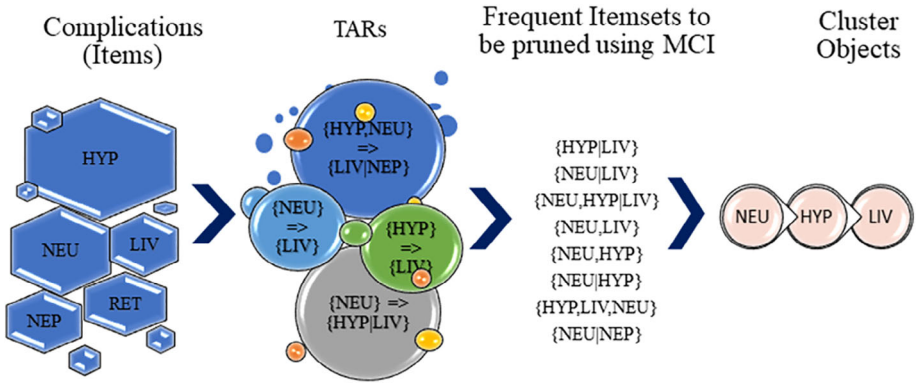

FIGURE 3 The proposed complication pattern mining methodology by using ARM and MCI to obtain the interesting itemsets as clustering objects [Color figure can be viewed at wileyonlinelibrary.com]

The optimal number of clusters, in here five, is established and validated by using the elbow method.[41] T2DM patients are grouped based upon C_{TAR}. If two rules do not share patients, we assume that they are not in the same cluster. In Figure 3, there were four T2DM patient clusters as the discovered hidden variable $C_H = \{C_H^1, C_H^2, C_H^3, C_H^4\}$ in which obtained using dissimilarity (1-correlation). Each one had a unique deep temporal phenotype (latent phenotype) and risk factor profile. In Figure 3, in the right-hand column (the most frequent ordering pattern of the complications), a symbol of $>$ between two complications demonstrates whether a complication in the left-hand of the symbol occurred before the right-hand one with the higher occurrence rate.

3.2 | Clustering comparison and validation findings

In this section, the latent phenotype clusters are compared with the TAR clusters by applying a number comparison and validation strategies to the identified clusters. These strategies assess the similarities among subgroups of patients, whereas they are clustered based upon different data sources. The comparison also aims to ensure a more appropriate decision for discovering the most meaningful subgroup of patients as well as explaining the behavior of the latent phenotype. For example, the intersection of C_H^4 and C_{TAR}^3 (the right-hand column in Table 7) revealed that a significant number of patients (with an overlap of $>$50%) shared a similar complications co-occurrence pattern. C_H^4 with the complications pattern of $\{HYP, NEU, RET, LIV\}$ and C_{TAR}^3 with the occurrence order of $\{RET, HYP\}, NEU, LIV$ have also coincided. The intersection of C_{TAR}^3 and C_H^4 showed that they greatly resembled each other, and it revealed an important link between the two clustering methods. Overall, we believed that there was a strong link between C_H^1 and C_{TAR}^1 where both clusters were sharing a similar complications co-occurrence pattern of $\{HYP, LIV, NEU\}$. In order to ascertain precisely that the overlap was not random, we used the NBH metric as illustrated in Table 7. C_{TAR}^3 was the most significant cluster with the lowest NBH values ($<$0.001) among C_{TAR}^i. It also had the highest percentage of C_H^j (52%) of overlapping patients. Patients within C_{TAR}^3 were more likely to develop RET, HYP, NEU, and LIV with the occurrence percentages of 96, 90, 25, and 13, respectively. Similarly, C_H^4 was more

TABLE 7 Probabilities of the Jaccard similarity, overlapped rate (ð), and NBH across C_H and C_{TAR}

	C^1_{TAR}		C^2_{TAR}		C^3_{TAR}		C^4_{TAR}		C^5_{TAR}	
	NBH	ð	NBH	ð	NBH	ð	NBH	ð	NBH	ð
C^1_H	<0.001	**90%**	0.580	45%	≥ 0.001	4%	0.480	38%	0.490	40%
C^2_H	0.064	66%	0.072	0%	0.290	16%	0.440	13%	0.092	0%
C^3_H	0.032	60%	0.630	9%	<0.001	28%	0.610	13%	<0.001	40%
C^4_H	<0.001	55%	0.045	45%	<0.001	52%	0.170	38%	0.530	20%

TABLE 8 Proportion of patients with the complication co-occurrence pattern for C_{TAR} and C_H

TAR Cluster	RET	NEU	NEP	LIV	HYP	Interesting Itemsets
C^1_{TAR}	0	15	10	16	100	{HYP}{LIV,NEU}
C^2_{TAR}	0	80	10	40	0	{NEU,LIV}
C^3_{TAR}	96	25	8	13	90	{RET,HYP},{NEU}{LIV}
C^4_{TAR}	67	0	33	17	50	{RET,HYP,NEP,LIV}
C^5_{TAR}	30	40	40	60	100	{HYP,LIV}{NEP,NEU,RET}
C^1_H	7	11	8	13	61	{HYP,LIV,NEU}
C^2_H	13	10	6	7	63	{HYP,RET,NEU}
C^3_H	4	16	11	13	56	{HYP,NEU,LIV,NEP}
C^4_H	12	13	6	11	57	{HYP,NEU,RET,LIV}

Note: On the right-hand, there are comparison results of the complication rates occurring in each cluster.

likely to develop $\{RET, HYP\}, NEU$, and LIV (see Table 8), revealing a significant as well as a meaningful relationship between those two clusters (C^4_H and C^3_{TAR}). Moreover, a C^i_{TAR} pattern, for example, $\{RET, HYP\}, \{NEU\}, \{LIV\}$ revealed that $\{RET, HYP\}$ was more likely to be seen than NEU, and NEU was more likely to be developed compared to LIV and the rest of complications were not likely to be developed in patients within the corresponding cluster C^3_{TAR} (as shown in Table 8). In particular, our hypothesis was checked whether C^2_H resembled C^4_{TAR}. As can be seen in Table 8, for the patients within C^4_{TAR}, the chances of having RET, HYP, and NEP were approximated by percentages of 67, 50, and 33, respectively. Similarly, the chance of having a consequence of RET, HYP, and NEP for patients in C^2_H was high (see evidence in Table 8). Additionally, as shown in Table 7, $C^2_H \cap C^4_{\text{TAR}}$ with the lowest NBH probability of <7.9E-90 and second highest overlapped number of patients of 25% revealed a significant and meaningful relationship between those two clusters (C^2_H and C^4_{TAR}). In this article, the dissimilarities (distances) between clusters are analyzed as the interestingness to filter discovered rules, which was optimized after filtering out uninteresting rules effectively. These results will attract a domain expert to choose interesting patterns from the remaining small set of rules. For instance, the itemsets consisting of similar items are uninteresting, despite the fact that the frequent itemsets with different items are interesting. Figure 4 represents a dendrogram of the TAR clusters based upon the objects.

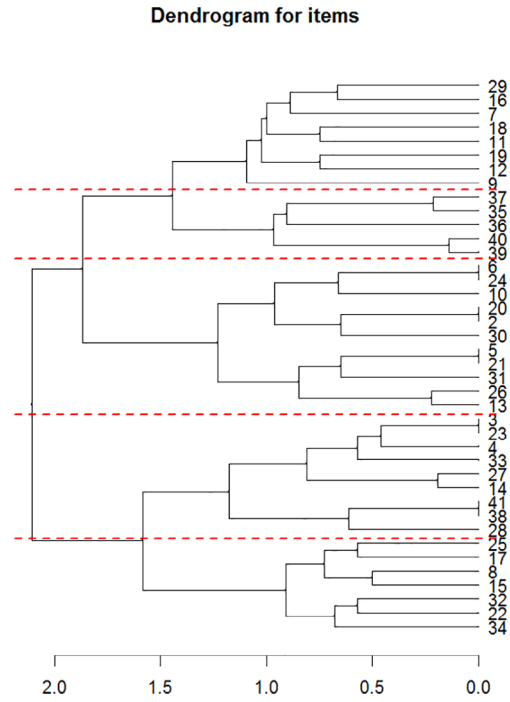

FIGURE 4 Hierarchical clustering for objects items in association rules, using dissimilarity Jaccard distance. *x*-axis and *y*-axis illustrate Jaccard Distance among objects and objects id obtained in Table 5, respectively [Color figure can be viewed at wileyonlinelibrary.com]

3.3 | The meaningful subgroup of the personalized patients

In this section, we attempted to investigate how the similarities between the C^i_{TAR} and C^j_H could validate and give meaning to the latent phenotype. Figure 3 represented patients in C^1_H, with a decreasing and an increasing pattern in their deep temporal phenotype, shared similar trajectories over the observed risk factor profiles. Almost 90% of patients within C^1_H was found in C^1_{TAR}. More importantly, it was significantly validated from a statistical point of view as the likelihood of randomly observing this overlap was very low with an NBH probability of <0.001, as shown in Table 7. Thus, there was sufficient evidence to suggest that nearly all patients belonged to a similar TAR cluster (C^1_{TAR}). It also appeared that the most frequent ordering pattern of complications of HYP, LIV, and NEU belonged to C^1_H matched $\{HYP, LIV, NEU\}$ belonged to C^1_{TAR}. Having known that patients within C^1_H and (C^1_{TAR}) were selected from two different data sources, not only statistically validated our clusters but also revealed the meaningfulness of the latent phenotype. Therefore, patients in the intersection of C^i_{TAR} and C^j_H ($C^i_{\text{TAR}} \cap C^j_H$) with the highest similarities among other clusters might represent a link between their latent phenotype and the temporal associated complications.

The most significant intersection of the TARs and latent phenotype clusters ($C^1_{\text{TAR}} \cap C^1_H$) was considered as the most informative (meaningful) subgroup and thought as DS1 (see Figure 1. We are interested in prediction the complications, personalizing patients based on their latent phenotype as well as the underlying pattern of complications. The latent variable is discovered based on the whole set of features (using IC* stepwise approach in DBNs framework). We trained the data, including all risk factors and complications for comparing two datasets (the original dataset (DS) to the meaningful subgroup (DS1)). In the next section, the prediction results are

TABLE 9 The prediction accuracy of a target complication (MAP), posterior likelihood level (clinical level), patients' group (dataset), evidence (E), P(MAP|E), P(E), and P(MAP,E) are compared between DS and DS1

	Level		Data					
MAP	**Low**	**High**	**DS**	**DS1**	**E**	**P(MAP|E)**	**P(E)**	**P(MAP,E)**
NEU	☑		☑		HYP,LIV	0.57	0.23	0.13
NEU		☑		☑	HYP,LIV	0.83	0.29	0.24
NEU	☑		☑		HYP,LIV,RET	0.57	0.23	0.13
NEU		☑		☑	HYP,LIV,RET	0.85	0.03	0.02
RET	☑		☑		HYP,LIV	0.71	0.23	0.16
RET	☑			☑	HYP,LIV	0.87	0.29	0.27
NEP		☑		☑	HYP,LIV	0.76	0.29	0.22
NEP		☑	☑		HYP,LIV,RET,NEU	0.76	0.02	0.02
NEP	☑			☑	HYP,LIV,RET,NEU	0.86	0.03	0.02
SMK		☑	☑		NEP	0.33	0.49	0.16
SMK		☑		☑	NEP	1.00	0.49	0.50
			Accuracy in DS	Accuracy in DS1				
NEP		☑	0.81	0.93				
LIV		☑	0.77	0.88				
HYP		☑	0.90	1.00				
NEU		☑	0.76	0.81				
RET	☑		0.81	0.79				
All			0.81	0.88				

analyzed to investigate the differentiation of DS1 and DS in terms of how accurate the hybrid complications prediction is in the personalised dataset compared to the raw dataset.

4 | EVALUATING THE PREDICTION PERFORMANCE

The evaluation strategy in this section argued that uncertainties in the cause and effects relationship among T2DM data could affect the prediction performance negatively. It also suggested that DS1 (by personalizing patients) could be considered as a dataset with less uncertainty compared to DS. This section has not concentrated only on the descriptive study. Therefore, by utilizing a predictive strategy (as a contribution for this chapter), the underlying patterns of complications were predicted for each of patients within DS1 (which were discovered in the descriptive strategy of the proposed hybrid methodology). These results then were compared to the prediction performance of the whole group of patients (which also includes DS1). This comparison attempted to reveal an explainability of the state-of-the-art method in order to uncover the meaning behind the latent AI model and gain insight into opening the black box. Thus, the prediction results were

analyzed to investigate the differentiation of DS1 and DS in terms of how accurate the hybrid complications were predicted in the personalised dataset (DS1) compared to the raw dataset of DS. Table 9 illustrated the prediction accuracy of the hybrid complications, which was compared between DS and DS1, where an optimal posterior likelihood of a high or low clinical level was the question of the interest.

4.1 | Improvement in the overall prediction accuracy

The prediction accuracy for each target complication was assessed for both DS and DS1 in Table 9. For example, the prediction accuracy of DS1 being diagnosed with HYP is 1, while for LIV and NEU are 0.88 and 0.81, respectively. Additionally, as shown in Table 9, the overall prediction accuracy across all complications for DS1 was 0.88 compared to a lower overall accuracy of 0.81 for DS. Similarly, the prediction accuracy for DS of individual complications was significantly smaller than in DS1, for HYP, LIV, and NEU by 0.90, 0.77, and 0.76, respectively. These results indicated that by applying the proposed methodology and discovering the meaningful subgroup, the prediction accuracy was increased for each complication within the most frequent ordering pattern of complications belonging to DS1. Accordingly, the overall prediction accuracy across all complications with a different pattern has been improved significantly.

4.2 | Optimal posterior likelihood

Predicting a target complication and deciding whether a diagnostic test result was positive or negative were challenging. One possible solution could be provided by computing the expected utility as a likelihood of each decision alternative. The clinical decision alternative with the highest expected gain must be an optimal option, which was chosen by the clinicians. Thus, an approach was utilised to approximate the posterior likelihood of developing complications when optimizing the Bayesian parameters. An integration of maximum entropy and Bayesian optimization methodology was applied to the parameters. For this purpose, the posterior likelihood of the developing complications was approximated by using "Maximum A posteriori Probability" (MAP) algorithm,[42,43] which converged toward the set of parameters.

In the proposed model with latent variables, MAP was utilized as an iterative strategy to discover maximum a posteriori of parameters. Then, an optimization procedure such as simulated annealing algorithm[44] was obtained to produce optimal posterior results along with the evidence. The simulated annealing algorithm was aggregated to a stochastic simulation of the hidden Markov chain which relied on data augmentation in the same way as EM algorithm.

The causal relationships are explained in terms of static and dynamic correlations between T2DM risk factors (attributes) to describe the inference problem. The causal inference has a greater focus on distinguishing causes from other associations than on uncovering detailed temporal relationships. It also facilitates a hybrid type approach that would yield useful information to find the inference used in probabilistic graphical models (Bayesian networks). These aim to distinguish and understand different categories while exploring knowledge in discovering causes. In this chapter, the prediction is obtained based on prior knowledge as well as the current stage of the risk factors and complications.

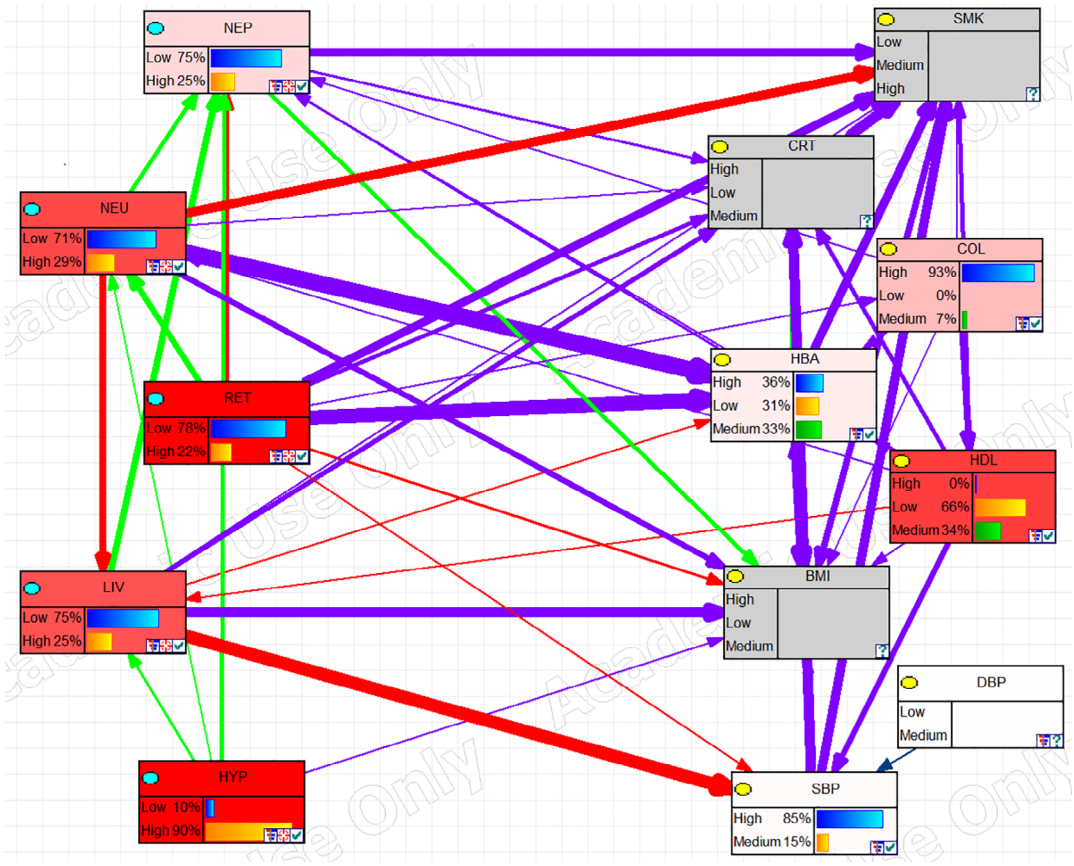

FIGURE 5 An influence diagram to represent Bayesian Structure applied to DS [Color figure can be viewed at wileyonlinelibrary.com]

In Table 9, an optimal posterior likelihood of developing RET, NEU, LIV, and SMK is compared with DS1 and DS in terms of a prior/evidence (already developed complications such as HYP and LIV). Having illustrated the extensive findings in Table 9, the cause and effect relationship were investigated in influence diagrams (as illustrated in Figures 5 and 6), which demonstrated Bayesian structures for DS and DS1, respectively. In these figures, class values for HYP and LIV are set to their highest/lowest clinical level, as the evidence, to observe changes in the clinical level of a targeted complication in the Bayesian structure modeling. In order to ascertain the obtained posterior likelihood of being at the high risk of having LIV and NEU, both should be coincided with demonstrating Bayesian structures. For example, once the patients in DS1 has been diagnosed with having HYP and LIV, the likelihood of developing NEU is increased to 0.84. Alternatively, with assuming that patients in DS to be diagnosed with NEU by knowing that the patients have already developed HYP and LIV, where the optimal level was low with the posterior likelihood of 0.57.

As can be seen in Table 9, if HYP and LIV class values were set to their high clinical level, probability of developing NEU ($P(NEU|\{HYP, LIV\})$ of 0.83 was higher than likelihood of not developing RET ($P(RET|\{HYP, LIV\})$ of 0.96 thus, the evidence showed that DS1 patients with HYP and LIV were at a high risk of being diagnosed with NEU compared to a high

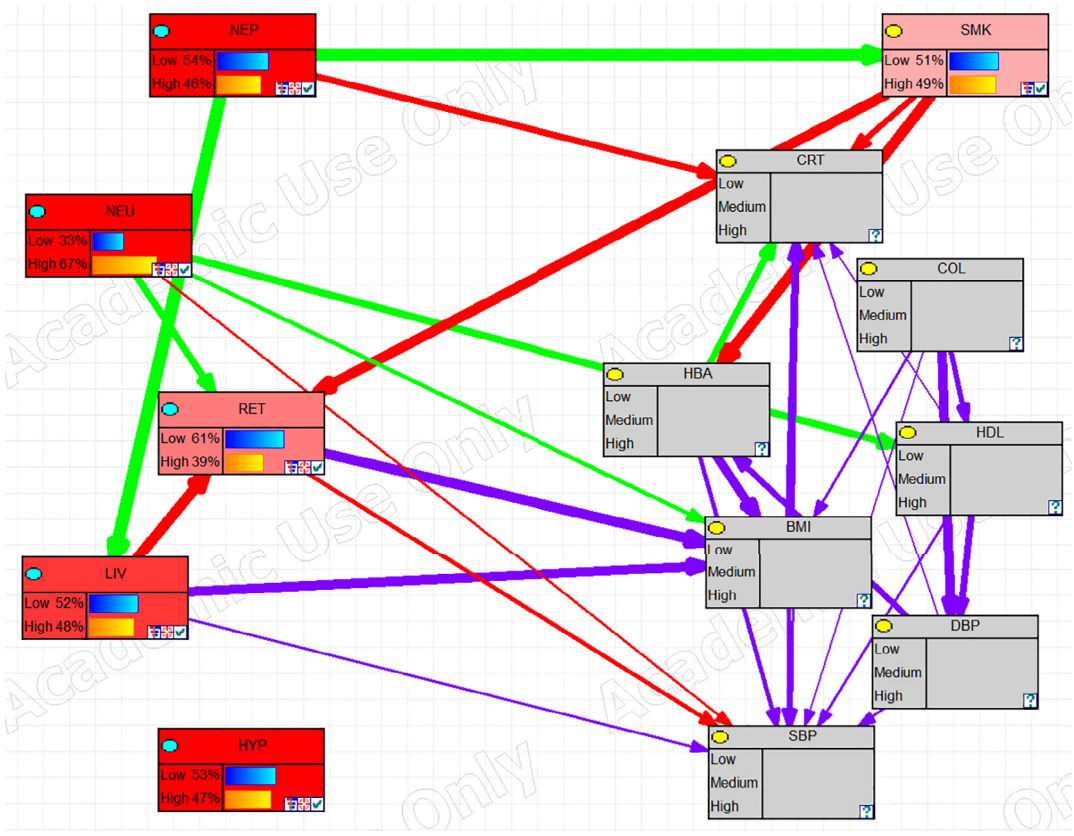

FIGURE 6 An influence diagram to represent Bayesian Structure applied to the subgroup of patients in DS1 [Color figure can be viewed at wileyonlinelibrary.com]

probability not being diagnosed with RET. Again, in Table 9, in DS1, when the posterior likelihood of LIV raised above 88%, growth of damaged eye cells in developing RET decreased to 96%. RET was negatively affected by the occurrence of LIV shown with a thick red arrow in DS1, which was revealed in Figure 6. Then this was compared to the no influence arrow in DS, as shown in Figure 5. Similarly, NEP in DS1 seemed less likely to be developed since HYP, LIV, RET and NEU occured with the optimal likelihood posterior of 0.86 at the low clinical level. However, for DS1, the posterior with the same evidence was 0.76 at a high clinical level (see Table 9). The influence of HYP (with a diagnosis likelihood of 1) on the rest of complications is neglected as it is a macrovascular complication, which is often developed in T2DM data with the same likelihood with or without adding it to the evidence of $(P(NEU|LIV) = P(NEU|\{HYP, LIV\}))$. This observation is ensured with a thicker red-coloured arrow pointing to RET from LIV in DS1 and no arc in DS as demonstrated in Figures 5 and 6.

In Figure 5, a thick purple edge from NEP to SMK illustrated the development of NEP causes SMK. Additionally, in Figure 6, positive causation was represented by a green edge from NEP to SMK. These findings suggested that once a patient has been diagnosed with NEP, the probability of being a smoker was increased significantly from 0.33 to 1.00 by comparing $P(SMK|NEP)$ values between DS and DS1 in Table 9.

TABLE 10 A subset of database $R(r_1)$ of the associated rules with the complications

Rule[a]	LHS		RHS	Objects[b]	Support	Confidence	Lift
7	{ }	⇒	{NEP}	2,11,13,14,16,17,18,20,25,	0.11	0.11	1.00
				29,30,32,35,36,37,38,41			
8	{ }	⇒	{NEU}	5,7,13,15,16,19,21,25,26,28,	0.16	0.16	1.00
				29,31,34,35,37,38,39,40,41			
9	{ }	⇒	{RET}	3,4,8,14,15,17,22,23,25,27,28,	0.15	0.15	1.00
				32,33,34,38,41			
10	{ }	⇒	{LIV}	6,12,18,19,22,24,29,30,31,32,	0.15	0.15	1.00
				33,34,35,36,37,39,40			
11	{ }	⇒	{HYP}	2,3,4,5,6,10,13,14,20,21,23,	0.86	0.86	1.00
				24,26,27,28,30,31,33,38,41			
12	{NEU|HYP}	⇒	{NEU}	13,26,38,41	0.01	0.27	1.71
39	{HYP}	⇒	{LIV}	6,24,30,31,33,35,36,37,39,40	0.14	0.16	1.06
40	{{NEP|HYP}|NEU}	⇒	{RET}	28,38,41	0.01	1.00	6.57
42	{NEU|RET}	⇒	{NEP|HYP}	28,38,41	0.01	0.19	6.84
53	{LIV|NEP}	⇒	{RET}	32,36	≥ 0.001	0.06	0.41
62	{LIV|NEU}	⇒	{RET}	34,39,40	0.01	0.29	1.88
80	{HYP|LIV|NEU}	⇒	{NEP}	37	0.01	0.33	3.11

[a]This table shows a subset of rules in r_1.

[b]Identification of a set of objects which follows any combinations of the items in D.

5 | DISCUSSION

Lack of prediction of the onset of associated diseases/complications can negatively affect a patient's health in many ways. They can be numerous and interact in complex nonlinear ways throughout the disease process. Patients must switch to another medication if more complications have been developed, for example, when a patient uses a treatment that may not be suitable for another complication. It leads to unsuccessful treatment, where clinicians are pushed to follow an unreliable and suboptimal approach in prescribing treatment options. In this situation, the medicine that is prescribed to help a patient in a particular complication might lead to patient dissatisfaction and more severe health outcomes. On the other hand, T2DM is potentially reversible, treatable, manageable, and, if caught, early enough. Early diagnosis and management of the disease have reduced the risk of complication development.[45]

The state-of-the-art modeling techniques for analyzing T2DM progression is either focused on descriptive or predictive strategies. Despite this, the present research in order to personalise patients in a precise prediction is based on both descriptive methodology and predictive analysis. For this purpose, the thesis conducted a new methodology based on a framework that combines notions of causality in medicine with algorithmic approaches built on Bayesian model as well as statistical techniques for analyzing the causal relationship. Additionally, having greater insight into the discovered subgroups and relatively, the prior understanding of the interesting

TABLE 11 A subset of database R (r_2) of the associated rules with the complications

Rule[a]	LHS		RHS	Objects[b]	Support	Confidence	Lift
10	{ }	⇒	{LIV}	6,12,18,19,22,24,29,30,31,32,	0.15	0.15	1.00
				33,34,35,36,37,39,40			
11	{ }	⇒	{HYP}	2,3,4,5,6,10,13,14,20,21,23,	0.86	0.86	1.00
				24,26,27,28,30,31,33,38,41			
16	{ }{ }	⇒	{RET}	3,4,8,14,15,17,22,23,25,27,	0.01	0.22	1.46
				28,32,33,34,38,41			
18	{ }{ }	⇒	{HYP}	2,3,4,5,6,10,13,14,20,21,23,	0.02	0.78	0.90
				24,26,27,28,30,31,33,38,41			
20	{NEP}	⇒	{NEU}	13,16,25,26,29,38,41	0.02	0.19	1.17
25	{LIV}	⇒	{NEP}	18,29,30,32,36,37	0.04	0.27	2.49
39	{HYP}	⇒	{LIV}	6,24,30,31,33,35,36,37,39,40	0.14	0.16	1.06
40	{{NEP|HYP}|NEU}	⇒	{RET}	28,38,41	0.01	1.00	6.57
42	{NEU|RET}	⇒	{NEP,HYP}	28,38,41	0.01	0.19	6.84
48	{LIV|NEU}	⇒	{NEP}	29,35,37	0.01	0.29	2.66
60	{HYP|LIV}	⇒	{NEP}	30,35,36,37	0.04	0.27	2.54
69	{HYP|LIV}	⇒	{NEU}	31,35,37,39,40	0.02	0.11	0.68
80	{HYP|LIV|NEU}	⇒	{NEP}	37	0.01	0.33	3.11

[a]This table shows a subset of rules which is called r_2.

[b]Identification of a set of objects which follows any combinations of the items in D.

rules helps interpreting the predictive results correctly. Therefore, the discovered hidden variable/latent phenotype can be combined with the meaningfully associated complication rules for optimal performance of the patient personalization.

Despite the importance of prediction of an expected complication at a time, finding a patient model that simultaneously takes into account the chance of occurrence of other associated complications can produce a more precise predictive model. In order to investigate whether a particular patient is at a high risk of developing a target complication, we need to analyze multiple factors. That may depend on the patient's clinical history, stage of the disease, and fluctuations of the related risk factors. More importantly, it can be affected by the associations of the prior complications with the expected complications (likely to be diagnosed and yet to be developed). In T2DM data, the worsening level of the microvascular diseases and HYP is known as a significant cause of death.[46] Even though microvascular complications such as RET, NEP, and NEU are less frequent comparing to HYP, an inadequate estimation of them causes long-term suffering and life-threatening comorbidities.[7] Fowler et al[9] researched type 2 diabetic American patients. This research utilized T2DM key risk factors such as H21Ac, SBP, and DBP to investigate relationships among complications such as HYP, NEP, RET, and NEU. In addition, LIV is a severe phenotype of diabetes and associated with T2DM complications, especially NEU.[47] Litwak et al analyzed Russian diabetic patients in Reference 48, which referred to the influence of macrovascular and microvascular disease on one anther. For example, important features in T2DM dataset such as blood pressure, HDL, lipid, BMI, and H2A1c influence diabetic patients' complications.

TABLE 12 The power set (*MCI*) obtained based on the MCI algorithm of the most interesting rules in MCI representing two subsets (r_1 and r_2) of the intersected associated rules with the complications

Rule[a]	LHS	RHS	Objects[b]	Support	Confidence	Lift	
7	{ }	⇒	{NEP}	2,11,13,14,16,17,18,20,25	0.11	0.11	1.00
				,29,30,32,35,36,37,38,41			
8	{ }	⇒	{NEU}	5,7,13,15,16,19,21,25,26,28	0.16	0.16	1.00
				,29,31,34,35,37,38,39,40,41			
9	{ }	⇒	{RET}	3,4,8,14,15,17,22,23,25,27,28	0.15	0.15	1.00
				,32,33,34,38,41			
10	{ }	⇒	{LIV}	6,12,18,19,22,24,29,30,31,32	0.15	0.15	1.00
				,33,34,35,36,37,39,40			
11	{ }	⇒	{HYP}	2,3,4,5,6,10,13,14,20,21,23	0.86	0.86	1.00
				,24,26,27,28,30,31,33,38,41			
12	{{NEU|HYP}}	⇒	{NEU}	13,26,38,41	0.01	0.27	1.71
11	{ }	⇒	{HYP}	2,3,4,5,6,10,13,14,20,21,23	0.86	0.86	1.00
				,24,26,27,28,30,31,33,38,41			
42	{NEU|RET}	⇒	{NEP,HYP}	28,38,41	0.01	0.19	6.84
60	{HYP|LIV}	⇒	{NEP}	30,35,36,37	0.04	0.27	2.54
62	{LIV|NEU}	⇒	{RET}	34,39,40	0.01	0.29	1.88
80	{HYP|LIV|NEU}	⇒	{NEP}	37	0.01	0.33	3.11

[a]This table shows an intersection of the most interesting rules from r_1 and r_2.
[b]Identification of a set of objects which follows any combinations of the items in D.

They also revealed that HDL has a negative effect on HYP, NEP, NEU, and RET, whereas H2A1c negatively associated with HYP. Again, a study conducted by Ramachandran et al[49] referred to the high prevalence of NEU and RET in Type 2 diabetes in India. Similarly, research in Reference 50 suggested that most of the diabetic patients have objective evidence for some variety of NEU, but only a few of them have identified by symptoms. This research also showed that there is a strong association among NEP, NEU, and RET.

All together, it seems pertinent to remember that understanding the underlying pattern of the complications is based on the correlation and causation of their co-occurrences (both positively and negatively). Here we give an illustration of what we mean. In the first case, an occurring complication is caused or followed by other complications. Alternatively, in the second case, if any combination of two complications is less likely to be followed or caused by another one. That is to say, the occurrence of some complications may negatively affects the occurrence of another complication. Here, we provide one case study example to clarify the contribution of this article. We have been able to come to the conclusion that if the levels of HYP and LIV of the patient population rises, the risk of developing RET decreases while the chance of developing NEU increases (based upon "causality backwards"). Moreover, since DS has appeared to be more complicated than DS1, there could be some unmeasured/hidden risk factors, which may affect both LIV levels and the likelihood of having RET in DS. In this situation with a considerable amount of uncertainty, one could argue that RET is caused by other underlying risk factors

(latent phenotype), such as exercise, genes, and diet. Thus, we attempt to open the clinical black box model by utilizing an appropriate methodology in order to discover correlation and causation among temporal risk factors and complications in the presence of hidden factors. We utilized DBNs, which allow the description of each time between cause and effect and the likelihood of this relationship being discovered. We obtained this causal phenotype with the associated probabilities as we had a tendency to calculate the joint impact a cause made to its influence and then observed statistically significant causes through the ideas of multiple hypothesis testing (treating each causal relationship as a hypothesis) and false discovery control. Having known the mentioned investigation, it seems reasonable to assume that the ordering of the complications co-occurrence and their temporal transactions produces remarkable/informative knowledge in order to interpret the patient model. This is not the only evidence that supports this study's claim, there is evidence to suggest that informative patterns of complications significantly improve the prediction performance for the personalised subgroup comparing to the original dataset.

6 | CONCLUSIONS

Discovering a latent phenotype by identifying the underlying sequence of temporally associated complications to explain AI black box model is notably absent from Scholar. It even becomes more problematic when a significant improvement in the predictive model is vital. Our main contribution in this article is based on the challenge of how to construct meaningful explanations of patients' subgroups in a precise prediction by uncovering the hidden factors. As a matter of fact, due to the difficulties of the explanation of the constraints and latent phenotype, we proposed a combination of data mining techniques while exploring knowledge in discovery cause and effect. For being able to explain the black-box model and hidden variables, we attempted to explore a well-known group of patients. We applied TARs to the data followed by MCI algorithm that filter out the most interesting itemsets only based on the underlying patterns of complications. Then, the discovered interesting patterns were considered as the input of the descriptive methods. Alternatively, the resulted subgroups of patients in the descriptive study became a new dataset for analyzing the predictive model. Then, we used the predictive model to capture the behavior of the latent variable and then in descriptive data mining techniques like unsupervised learning, patients were allocated to four clusters only based on their latent phenotype.

In this work, a combined data mining methodology was adopted to help understand and validate the latent phenotype in order to find a meaningful subgroup of patients. It also intended to assist the clinicians in the decision-making process to help with the early and precise diagnosis of complications. Existing approaches in pattern discovery from time series clinical data have not yet exploited the representational power of the integrated data mining techniques such as hidden variable discovery, TAR mining, time series clustering, patient personalization, and enhanced prediction methodology.

To sum up, in this research, we addressed three goals. First, we demonstrated a rich clinical data to provide fine temporal phenotype in associations. Second, we aimed to illustrate cluster analysis of time series data with an underlying causal structure in T2DM phenomenon. Furthermore, considering the hybrid complications as a class, in the classification/prediction problem, we addressed the unbalanced issue. Our promising experimental results showed that the patient personalization by using the proposed integrated data mining techniques could provide better prediction accuracy and interpretability in discovering the temporal associated complication rules and understanding the latent phenotype. More importantly, these findings revealed that the

proposed hybrid techniques could handle uncertainty in the clinical decision-making process. It also aided the clinicians to prepare future prognosis of the most likely occurring complications.

Nevertheless, several questions remain to be answered as we have just attempted to open the black box AI models. In future work, we are attempting to provide more interpretation of the results from the clinician's point of view. The generalizability of these results is limited to the T2DM dataset. Thus, we intend to apply the proposed methodology to a new dataset with more risk factors and patient visits with the aim to understand the black box latent DBNs model. This new understanding should help to improve predictions of the impact of the latent phenotype on associated complications. In the future works, a few possible solutions can be of interest to the authors of this paper. For instance, causal confidence and support could be combined with the other metrics in order to uncover these types of uncertainties. We will also consider employing Fisher's *p*-value that is ranked as the most robust measure in which ensuring the interesting itemsets acts as an informative input in the predictive model.

SUPPORTING INFORMATION

Implementation tools

We exploit AR mining based on an extension package "arules" in R.[39] The original (imbalanced) dataset is considered to find a pattern of developing different complications throughout patients visits. Additionally, we use the R-extension package "arulesViz" and "Gephi" for visualization techniques to explore ARs clearly. The visualization techniques are utilized to determine a considerable number of rules, allowing interesting information to be discovered from the transaction data. Finally, "Genie" is used to infer the BN as well as illustrate an influence diagram and applied diagnosis test and sensitivity analysis.

TARs

The support measure of itemsets X (supp(X)) is defined as the proportion of transactions in the dataset containing X. In particular, an ARs of $X \Rightarrow Y$ has a support of $P(XY)$. The confidence measure of a rule ($conf(X \Rightarrow Y)=\frac{P(XY)}{P(X)}$) identifies the proportion of transactions with the most interesting or important relationships. In addition, the confidence of a rule is defined as conf($X \Rightarrow Y$) = supp($X \cup Y$) = supp(X) in which it satisfies Equation (8).

$$\text{supp}(X \cup Y) > \sigma, \text{conf}(X \Rightarrow Y) > \delta. \tag{8}$$

Parameters such as σ and δ are the minimum support and confidence, respectively. Instead of using accuracy, efficiency is an appropriate way to evaluate ARs.[34] To obtain the frequents itemsets, first, we filter TARs by using support and confidence. However, they are not able to filter complication rules based on the different dependencies among the rules. For this purpose, we used a measurement of independence of X and Y (known as lift and defined as $\text{lift}(X \Rightarrow Y) = \frac{P(XY)}{P(X)P(Y)}$. Lift of 1 represents two itemsets X and Y are independent as shown in Equation (9).

$$\text{lift}(X \Rightarrow Y) = \text{supp}(X \cup Y) = \text{supp}(X)\text{supp}(Y). \tag{9}$$

Lift is the deviation of the whole rule support from the expected support under independence given both sides of the rule support. Higher lift values indicate strong associations. For instance, the conditional probability of a patient developing both HYP and LIV is associated with the likelihood of the patient developing RET. For example, the confidence of *HYP*, *LIV* implying *RET* is given as the likelihood of the patient developing *HYP*, *LIV* and also *RET* over the likelihood of developing only *HYP* and *LIV* (see Equation (10)).

$$\text{Conf}(\{HYP, LIV\} \Rightarrow \{RET\}) = \frac{\text{Supp}(\{HYP, LIV, RET\})}{\text{Supp}(\{HYP, LIV\})}. \tag{10}$$

The confidence measure in $\{RET, HYP, NEU, RET\}$ implying *LIV* reveals how likely a given patient developed $\{RET, HYP\}, NEU, RET$, and also *LIV*.

MCI

To ascertain whether patients in each cluster are developing a similar pattern of complications (the most frequent itemsets which are also unique for the corresponding cluster) as well as a different pattern from other patients within another cluster. For instance, if *HYP* happens before *LIV* and *RET* and *NEU* or *NEP* or no complication occur after them, there is a co-occurrence pattern of $\{HYP, \{LIV, RET\}, \{NEU|NEP\}\}$. For example, we select two subsets of rules with maximum lift and reasonable support and lift (meeting the constraints) as follows (as shown in Tables 10 and 11):

$$r_1 = \{R_7, R_8, R_9, R_{10}, R_{12}, R_{27}, R_{39}, R_{40}, R_{42}, R_{53}, R_{62}, R_{80}\},$$

$$r_2 = \{R_{10}, R_{11}, R_{16}, R_{18}, R_{20}, R_{25}, R_{39}, R_{40}, R_{42}, R_{48}, R_{60}, R_{69}, R_{80}\}.$$

The union of objects in these subsets is meeting the most items in D. We need to find out whether the rules set are covering the optimal/minimal number of the associated objects. There is an ideal itemsets MCI of the intersection of r_1 and r_2, which is defined as $\text{MCI} = \{R_{10}, R_{11}, R_{42}, R_{60}, R_{62}\}$ (as illustrated in Table 12). These itemsets are generated based upon the intersection of objects in MCI representing a unique/minimum coverage set of items in D and are illustrated in Figure 2.

ORCID

Leila Yousefi https://orcid.org/0000-0003-1952-0674

REFERENCES

1. Yousefi L, Saachi L, Bellazzi R, Chiovato L, Tucker A. Predicting comorbidities using resampling and dynamic Bayesian networks with latent variables. Paper presented at: 2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS); IEEE; 2017:205-206.
2. Yousefi L, Tucker A, Al-luhaybi M, Saachi L, Bellazzi R, Chiovato L. Predicting disease complications using a stepwise hidden variable approach for learning dynamic Bayesian networks. Paper presented at: 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS); IEEE; 2018:106-111.
3. Yousefi L, Swift S, Arzoky M, Saachi L, Chiovato L, Tucker A. Opening the black box: discovering and explaining hidden variables in type 2 diabetic patient modelling. Paper presented at: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); IEEE; 2018:1040-1044.
4. Yousefi L, Swift S, Arzoky M, Sacchi L, Chiovato L, Tucker A. Opening the black box: exploring temporal pattern of type 2 diabetes complications in patient clustering using association rules and hidden variable discovery. Paper presented at: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS); IEEE; 2019:198-203.

5. UK Prospective Diabetes Study Group. UK prospective diabetes study 16: overview of 6 years' therapy of type II diabetes: a progressive disease. *Diabetes*. 1995;44(11):1249-1258.
6. Yoon KH, Lee JH, Kim JW, et al. Epidemic obesity and type 2 diabetes in Asia. *Lancet*. 2006;368(9548):1681-1688.
7. Munana KR. Long-term complications of diabetes mellitus, Part I: Retinopathy, nephropathy, neuropathy. *Vet Clin Small Animal Pract*. 1995;25(3):715-730.
8. Wang T, Lin Q. Hybrid predictive model: when an interpretable model collaborates with a black-box model. 2019; arXiv preprint arXiv:1905.04241.
9. Fowler MJ. Microvascular and macrovascular complications of diabetes. *Clin Diabetes*. 2008;26(2):77-82.
10. Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. Paper presented at: ACM SIGMOD Record; ACM; 1993:207-216.
11. Moskovitch R, Shahar Y. Medical temporal-knowledge discovery via temporal abstraction. Paper presented at: AMIA Annual Symposium Proceedings; American Medical Informatics Association; 2009:452.
12. Wang W, Yang J, Muntz R. TAR: Temporal association rules on evolving numerical attributes. Paper presented at: Proceedings 17th International Conference on Data Engineering; IEEE; 2001:283-292.
13. Gharib TF, Nassar H, Taha M, Abraham A. An efficient algorithm for incremental mining of temporal association rules. *Data Knowl Eng*. 2010;69(8):800-815.
14. Allen JF. Towards a general theory of action and time. *Artif Intell*. 1984;23(2):123-154.
15. Sacchi L, Larizza C, Combi C, Bellazzi R. Data mining with temporal abstractions: learning rules from time series. *Data Min Knowl Disc*. 2007;15(2):217-247.
16. Ale JM, Rossi GH. An approach to discovering temporal association rules. Paper presented at: Proceedings of the 2000 ACM symposium on Applied computing; Vol 1; 2000:294-300.
17. Huang JW, Dai BR, Chen MS. Twain: two-end association miner with precise frequent exhibition periods. *ACM Trans Knowl Discov Data*. 2007;8(2):800-815.
18. Zhao Q, Bhowmick SS. *Association Rule Mining: A Survey*. Singapore: *Nanyang Technological University*; 2003:135.
19. Luna JM, Ondra M, Fardoun HM, Ventura S. Optimization of quality measures in association rule mining: an empirical study. *Int J Comput Intel Syst*. 2018;12(1):59-78.
20. Luna JM, Fournier-Viger P, Ventura S. Frequent itemset mining: a 25 years review. *WIRES Data Min Knowl Discov*. 2019;9(6):e1329.
21. Hahsler M, Karpienko R. Visualizing association rules in hierarchical groups. *J Bus Econ*. 2017;87(3):317-335.
22. Li Y, Ning P, Wang XS, Jajodia S. Discovering calendar-based temporal association rules. *Data Knowl Eng*. 2003;44(2):193-218.
23. Concaro S, Sacchi L, Cerra C, Fratino P, Bellazzi R. Mining health care administrative data with temporal association rules on hybrid events. *Methods Inf Med*. 2011;50(02):166-179.
24. Lee C-H, Ming-Syan C, Lin C-R. Progressive partition miner: an efficient algorithm for mining general temporal association rules. *IEEE Trans Knowl Data Eng*. 2003;4:1004-1017.
25. Plasse M, Niang N, Saporta G, Villeminot A, Leblond L. Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. *Comput Stat Data Anal*. 2007;52(1):596-613.
26. Mani S, Cooper GF. Causal discovery using a Bayesian local causal discovery algorithm. Paper presented at: Medinfo; 2004:731-735.
27. Sparacino G, Facchinetti A, Maran A, Cobelli C. Continuous glucose monitoring time series and hypo/hyperglycemia prevention: requirements, methods, open problems. *Curr Diabetes Rev*. 2008;4(3):181-192.
28. Mennis J, Liu JW. Mining association rules in spatio-temporal data: an analysis of urban socioeconomic and land cover change. *Trans GIS*. 2005;9(1):5-17.
29. Lakkaraju H, Kamar E, Caruana R, Leskovec J. Faithful and customizable explanations of black box models. Paper presented at: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society; 2019:131-138.
30. Pedreschi D, Giannotti F, Guidotti R, Monreale A, Ruggieri S, Turini F. Meaningful explanations of Black Box AI decision systems. Paper presented at: Proceedings of the AAAI Conference on Artificial Intelligence; 2019:9780-9784.

31. Turner RC, Millns H, Neil HA, et al. Risk factors for coronary artery disease in non-insulin dependent diabetes mellitus: United Kingdom Prospective Diabetes Study (UKPDS: 23). *BMJ*. 1998;316(7134):823-828.
32. Bellazzi R, Sacchi L, Concaro S. Methods and tools for mining multivariate temporal data in clinical and biomedical applications. Paper presented at: 2009 Annual Inter national Conference of the IEEE Engineering in Medicine and Biology Society; IEEE; 2009:5629-5632.
33. Bellazzi R, Ferrazzi F, Sacchi L. Predictive data mining in clinical medicine: a focus on selected methods and applications. *WIRES Data Min Knowl Discov*. 2011;1(5):416-430.
34. Ahmad P, Qamar S, Rizvi SQ. Techniques of data mining in healthcare: a review. *Int J Comput Appl*. 2015;120(15):38-50.
35. Mining What Is Data. Data mining: Concepts and techniques. Morgan Kaufinann. 2006;10:559-569.
36. Djenouri Y, Gheraibia Y, Mehdi M, Bendjoudi A, Nouali-Taboudjemat N. An efficient measure for evaluating association rules. Paper presented at: 2014 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR); IEEE; 2014:406-410.
37. Moon TK. The expectation-maximization algorithm. *IEEE Signal Process Mag*. 1996;13(6):47-60.
38. Liu G, Huang S, Lu C, Du Y. An improved K-Means Algorithm Based on Association Rules. *Int J Comp Theory Eng*. 2014;6(2):146.
39. Hahsler M, Chelluboina S, Hornik K, Buchta C. The a rules R-Package Ecosystem: Analyzing Interesting Patterns from Large Transaction Datasets. *J Machine Learn Res*. 2011;12:1977-1981.
40. Stephen S, Allan T, Veronica V, et al. Consensus clustering and functional interpretation of gene-expression data. *Genome Biol*. 2004;5(11):R94.
41. Zambelli AE. A data-driven approach to estimating the number of clusters in hierarchical clustering. *F1000Research*. 2016;5. https://doi.org/10.12688/f1000research.10103.1.
42. Hurwitz H Jr. Entropy reduction in Bayesian analysis of measurements. *Phys Rev A*. 1975;12(2):698.
43. Herman GT. Application of maximum entropy and Bayesian optimization methods to image reconstruction from projections. *Maximum-Entropy and Bayesian Methods in Inverse Problems*. Dordrecht, Netherlands: Springer; 1985:319-338.
44. Granville V, Krivánek M, Rasson JP. Simulated annealing: a proof of convergence. *IEEE Trans Pattern Anal Mach Intell*. 1994;16(6):652-656.
45. Bellamy L, Casas JP, Hingorani AD, Williams D. Type 2 diabetes mellitus after gestational diabetes: a systematic review and meta-analysis. *Lancet*. 2009;373(9677):1773-1779.
46. Cusick M, Meleth AD, Agrón E, et al. Associations of mortality and diabetes complications in patients with type 1 and type 2 diabetes: early treatment diabetic retinopathy study report no. 27. *Diabetes Care*. 2005;28(3):617-625.
47. Thuluvath PJ, Triger DR. Autonomic neuropathy and chronic liver disease. *QJM-Int J Med*. 1989;72(2):737-747.
48. Litwak L, Goh SY, Hussein Z, Malek R, Prusty V, Khamseh ME. Prevalence of diabetes complications in people with type 2 diabetes mellitus and its association with baseline characteristics in the multinational A_1chieve study. *Diabetol Metab Syndr*. 2013;5(1):57.
49. Ramachandran A, Snehalatha C, Satyavani K, Latha E, Sasikala R, Vijay V. Prevalence of vascular complications and their risk factors in type 2 diabetes. *J Assoc Physicians India*. 1999;47(12):1152-1156.
50. Dyck PJ, Kratz KM, Karnes JL, et al. The prevalence by staged severity of various types of diabetic neuropathy, retinopathy, and nephropathy in a population-based cohort: the Rochester Diabetic Neuropathy Study. *Neurology*. 1993;43(4):817-817.

How to cite this article: Yousefi L, Swift S, Arzoky M, Saachi L, Chiovato L, Tucker A. Opening the black box: Personalizing type 2 diabetes patients based on their latent phenotype and temporal associated complication rules. *Computational Intelligence*. 2020;1–39. https://doi.org/10.1111/coin.12313