Time-Series Bootstrapping For Identifying Latent Variables in Dynamic Bayesian Networks Applied To Type 2 Diabetes Complication Prediction

Journal:	*IEEE Journal of Biomedical and Health Informatics*
Manuscript ID	JBHI-01212-2018
Manuscript Type:	Paper
Date Submitted by the Author:	09-Nov-2018
Complete List of Authors:	Yousefi, Leila; Brunel University College of Engineering Design and Physical Sciences; Leila Yousefi Al-Luhaybi, Mashael; Brunel University College of Engineering Design and Physical Sciences Sacchi, Lucia; University of Pavia, Dept. of Computer Science and Systems; Chiovato, Luca; Endocrinology Unit of Instituti Clinici Salvatore Maugeri Tucker, Allan; Brunel University, School of Information Systems, Computing and Mathematics

Time-Series Bootstrapping For Identifying Latent Variables in Dynamic Bayesian Networks Applied To Type 2 Diabetes Complication Prediction

Leila Yousefi, Mashael Al-Luhaybi, Lucia Sacchi, Luca Chiovato, and Allan Tucker

Abstract—**Type 2 Diabetes Mellitus (T2DM) is a rising public health concern worldwide. It is a chronic disease with an onset that is commonly complicated by associated comorbidities. T2DM patients are at increased danger of micro vascular complications (comorbidities), such as hypertension, liver disease, and retinopathy. With each visit, a patient has a unique profile of symptoms and complications, regardless of the phase of the disease. Models of time-series data are needed to manage diabetic comorbidities that can deal with this type of data, where time-series data is imbalanced and there are complex interactions. Although extensive research has been carried out on the prediction of diabetic progression, no single study exists which has attempted, in detail, to interpret the impact of latent (hidden) variables in the presence of diabetic disorders. The aim of this work is to balance Time-Series clinical data with a Bootstrapping approach to enhance prediction of the future phases of diabetic complications for patients at various stages in the disease. In addition, we determine the precise position of the latent variable and discover the dependencies between the latent variable and the observed variables. Specifically, our key contribution is the combination of the IC* algorithm to identify latent variables within DBNs, as well as inference methods to assess the influences of these latent variables. We show how this targeting of latent variable influence improves prediction accuracy, specificity, sensitivity over standard latent variables approaches as well as aids understanding disease complications and related risk factors.**

***Index Terms*—Diabetes, Disease prediction, Dynamic Bayesian Networks, Latent variable and Time series Bootstrapping.**

I. Introduction

DIABETES UK revealed Type 2 Diabetes Mellitus (T2DM) as a silent killer, which is increasingly seen as a serious, worldwide public health concern. T2DM is the most usual common form of diabetes, accounting for at least 90 percent of all instances. The World Health Organization reported that in the next 12 years there will be about 550 million people suffering from this disease [1]. This disease occurs because of impaired insulin secretion or opposition to insulin action or both, which is associated with severe long-term morbidities and large health maintenance costs to providers. Moreover, T2DM is commonly complicated by other medical conditions. Previous research has established that only 14 percent of patients with type 2 diabetes have no other comorbidities [2]. For instance, hypertension is a major cardiovascular disease risk factor [3]. Up to 75 percent of adults with diabetes also have hypertension, and patients with hypertension often show evidence of insulin resistance. It has previously been observed that patients with T2DM are also at an increased risk of microvascular comorbidities, including nephropathy, neuropathy, and retinopathy [4].

Many people with T2DM are not aware that associated complications have developed. Thus, the prediction of diabetes complications may take a long time and leave patients at a huge cost for the treatment. The non-stationary characteristics of clinical data, collected as part of the monitoring of T2DM, create a difficult context for effective forecasting [5]. At every medical visit, all diabetic patients have a unique profile of symptoms and complications that change over time, regardless of the phase of the disease. Taking into account the state of the patient during each visit and over time changes can be an important challenge for physicians preparing for future visits. Clinical data needs to be considered as time-series data to provide a description of the progression of a disease over time. Dealing with time-series patient records is known to be a major issue in the prognosis of comorbidities [6]. In particular, mining time-series in the prognosis of disease with rare positive results is one of the challenging problems. Furthermore, the issue of imbalanced data learning has received considerable critical attention in data mining [7]. It has been reported that a class imbalance in the training data caused by one class (here positive cases) heavily outnumbering the examples in another class (negative class) [8]. Additionally, there is still some uncertainty around understanding the relationship between observed clinical data, comorbidities and unmeasured variables. Discovering latent variables aims to capture unmeasured effects from clinical data, reveals a better understanding of application domains, and simplifies complex networks of interactions. It can improve classification accuracy and boost user confidence in the classification models [9]. In [10], the authors emphasised the importance of the presence of hidden variables and then determined a hidden variable that interacts with observed variables and tried to locate it in the right place in the Bayesian network structure. In addition, they

L. Yousefi, M. Al-Luhaybi and A. Tucker were with the Department of Computer Science, Brunel University London, United Kingdom.
E-mail: Leila.yousefi, Mashael.Al-Luhaybi and Allan.Tucker@brunel.ac.uk
L. Saachi was with the Department of Computer Science and Systems, University of Pavia, Pavia, Italy.
L. Chiovato was with the Endocrinology Unit of Instituti Clinici Salvatore Maugeri, Pavia, Italy.
E-mail: lucia.sacchi, Luca.chiovato@unipv.it

noted that networks without hidden variables are clearly less useful because of the increased number of edges needed to model all interactions. There are various methodologies for Diabetes prediction, e.g, risk-prediction equation and Markov models [11]. However, the former (risk-prediction equation) suffers from uncertainty as well as performs only one-step-ahead prediction, while the latter (Markov model) is limited to a small number of discrete risk factors for a prediction. It has been reported that Dynamic Bayesian Networks (DBNs) are suitable for modelling uncertain noisy time-series such as T2DM clinical data [12]. DBNs are probabilistic graphical models that can classify longitudinal data considering noise, missing data, and unmeasured risk factors. They can be used to combine existing knowledge with data and be interpreted by non-statisticians. Previous works on learning DBNs have inferred both network structures and parameters from (sometimes incomplete) clinical data sets [12]. There has been some research investigating the prognosis of diabetes type 2 complications [13]–[15]. Some researchers used logistic regression and Nave Bayes with diverse modelling strategies, including methods for unbalanced data [14]–[16]. There have been few investigations into diagnosing diabetes type 1 exploiting Bayesian Networks [15]. In Marinis paper, variables are connected within two time-series and within the same time slice assumes that the temporal dependencies are time-invariant. In [13] external and internal heterogeneity was explored in T2DM patients for predicting comorbidities in cross-sectional data with three horizons of time. Marini et al. [15] simulated the health state and complications of type 1 diabetes patients by using partially and entirely Bayesian inferred models. However, we use a different approach to the representation of the relationship between different comorbidities. Dagliati et al. [13] presented a Hierarchical Bayesian Logistic Regression model to anticipate patients changes when the individual model parameters are estimated. However, the major limitation of their work derives from time discretisation in temporal time slices per year. In Addition, parameter estimation used in the MCMC approach is not suitable for large datasets. Moreover, time-series modelling was not employed in the individual measurements. A recent study presented a Dynamic Bayesian Network method with a latent variable for modelling fisheries data [17]. Many diseases involve structural changes based upon key stages in the progression, but many models do not take this into account. For instance, there has been some work in extending DBNs to model underlying processes that are non-stationary [18] [19]. Authors in [20] used a Markov Chain Monte Carlo (MCMC) approach to estimate the variance in the data structure, but the search space was limited to a fixed number of segments and indirect edges only. The ability to both assess weak and strong changes in variable distributions and explicitly model the evolution of their relationships would be extremely informative, especially in unknown processes such as complex disease. For all approaches, the search space is usually limited by constraints on one or more degrees of freedom; the segmentation points of the time-series, the parameters of the variables, the dependencies between the variables and the number of segments. [21] formalised non-stationary DBN models and proposed an MCMC sampling algorithm for learning the structure of the model from time-series biological data. In [22] instead retained the stationary nature of the structure in favour of parameter flexibility, arguing that structure changes lead almost certainly to over-flexibility of the model in short time-series.

TABLE I
MAIN CLINICAL RISK FACTORS OF T2DM AND CONTROL (MEAN ± SD) IS SHOWN IN FIGURE 2.

Node		Clinical feature	Complication	Hidden variable	
1	HbA1c (%)	6.6 ± 1.2	YES	NO	NO
2	Retinopathy	0,1	NO	YES	NO
3	Neuropathy	0,1	NO	YES	NO
4	Nepropathy	0,1	NO	YES	NO
5	Liver Disease	0,1	NO	YES	NO
6	Hypertension	0,1	NO	YES	NO
7	BMI (kg/m2)	26.4 ± 2.4	YES	NO	NO
8	Creatinine (mg/dL)	0.9 ± 0.2	YES	NO	NO
9	Cholestrol (mg/dL)	0.9 ± 0.2	YES	NO	NO
10	HDL cholesterol (mmol/l)	1.1 ± 0.3	YES	NO	NO
11	Diastolic blood pressure(DBP) (mmHg)	91± 12	YES	NO	NO
12	Systolic blood pressure(SBP) (mmHg)	148 ± 19	YES	NO	NO
13	Smoking Habit	0,1,2	YES	NO	NO
Hidden1	First Hidden variable	[0,1]	NO	NO	YES

Various studies on longitudinal data mining literature suggest an association between comorbidities and T2DM risk factors e.g., [23]. Clinicians cannot measure all risk factors and carry out all kinds of tests, so there are some unmeasured factors that clinicians fail to measure, which need to be discovered at the early stage of diabetes. Early prediction of T2DM complications i.e., finding the behaviour of associated aggressive risk factors, can help to improve patients' quality of life [24]. Yousefi et al. [25] developed an intuitive stepwise method to learn these latent effects based upon the IC* algorithm, using Pair-sampling for balancing data. However, there is still a need to investigate the impact of latent variables on the prediction of T2DM comorbidities. Here, we expand our previous work [26] on disease progression as a time-series process where the dynamic variables are affected depending on the stage of the disease. In this paper, prediction model for T2DM is learned for obtaining medical insight, taking into account the issues with overcoming class imbalance. Around half of the paper is dedicated to explaining how to balance time series clinical data and how to learn DBNs with and without latent variables. A set of models are learned from the data to evaluate the impact of adding latent variables and re-balancing the data (via bootstrapping). The other half of the paper is dedicated to analyse the results, in terms of classification (predicting two comorbidities associated to T2DM) and temporal prediction.

II. DATA PRE-PROCESSING AND IMPUTATION

We explored the use of latent variable models for prediction and the early detection of these comorbidities from clinical follow-ups of diabetes patients at the IRCCS instituti clinici scientifici (ICS) Maugeri of Pavia, Italy. A key aim of this paper is to rebalance clinical data to improve our latent variable learning model. In this study, a previously collected dataset for clinical and management purposes is used from the MOSAIC project funded by the European Commission under the 7th Framework Program, Theme ICT–2011.5.2 Virtual Physiological Human (600914) from 2009 to 2013.

Diabetes health status records are accumulated from 356 prediagnosed diabetic patients with an unequal number of visits per year, which was 3959 instances (the total number of observations considering all the patients). The T2DM risk factors were shown as associated variables over time for various microvascular comorbidities (diabetic nephropathy, neuropathy, and retinopathy), non-alcoholic liver disease, and hypertension) [27] [28] [23] [4]. Predictors of T2DM complications include Body Mass Index (BMI), Systolic Blood Pressure (SBP), High-Density Lipoprotein (HDL), Glycated Hemoglobin (HbA1c), Diastolic Blood Pressure (DBP), total Cholesterol (Cholesterol), Smoking habit and Creatinine (see Table I). We stratified the patient visits based on their risk factors and classified T2DM risk factors and comorbidities using the model and prioritised features based on existing literature on diabetes. We selected these risk factors and features based on literature from [29].

III. Method

A. Dynamic Bayesian Networks

We designed a probabilistic model, which represents the integration of comorbidities and clinical data of T2DM patients. Our DBN model exploited independence assumptions for modelling the joint distribution of the domain and represented probabilistic relationships between comorbidities and risk factors. Given a set of symptoms, DBNs were used to compute the probabilities of the presence of time series comorbidities. In our DBN structure, nodes represent variables at distinct time slots and there are links between nodes over time, so they can be used to forecast into the future. Our explanation of Bayes theorem for our model in prediction T2DM complication using time series T2DM risk factors is represented in Equation 1.

$$P(complication|riskfactors) = \frac{P(riskfactors|complication) * P(complication)}{(riskfactors)} \qquad (1)$$

B. Balancing strategy (Visit-based time series Bootstrapping-TS bootstrapping)

The primary objectives of this paper were to balance clinical time series data and then discover relationships between latent variables and clinical features within a DBN framework. Figure1 represented the whole flow chart of our prediction model, including balancing strategy. We discretised risk factors to be applied to the discrete space model in DBNs. Therefore, each risk factor with a dynamic nature has qualitative states, e.g, low, medium, and high. In contrast, the comorbidities have binary states which are very imbalanced with an unbalanced ratio of 3.2, comparing positive cases to negative cases. Imbalanced data, which is common in clinical data, includes an uncommon but important event, leading to difficulties in learning models related to a minority class. The problem of class imbalance is related to learning with too few positive cases, where patients have been diagnosed positively having a specific comorbidity at each time point. Therefore, a minority class in this paper represents a patients visit during which a complication is diagnosed. Once a complication is diagnosed, it is recorded for the rest of the patients time series data. For analysing complicated predictors, bootstrap [30] produces more accurate and reliable results rather than other traditional statistical and resampling methods [31]. Therefore, firstly, we exploited a bootstrap approach (TS Bootstrapping) to resample our observed time-series visits per patient whereby the original training data was sampled in pairs of consecutive time points, *t*-1 and *t*.

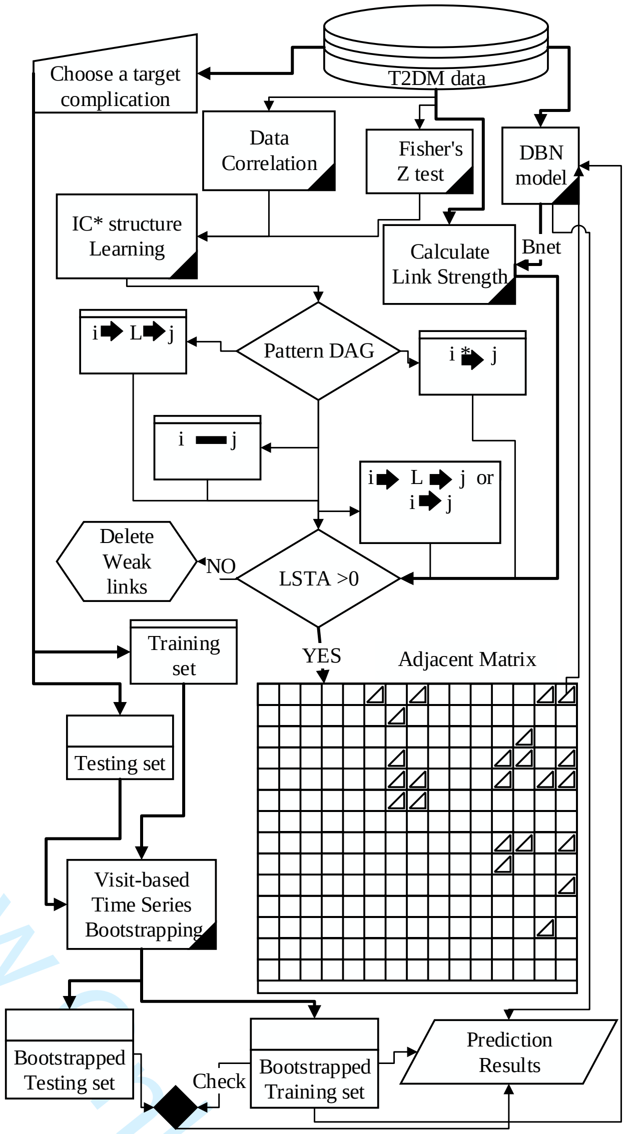

Fig. 1. This diagram demonstrates our Balancing strategy (Visit-based time series Bootstrapping) and structure learning approach (IC*LS).

For analysing the performance measurements from our observed patients dataset, we construct artificial patient samples by resampling with replacement from the observed sample. We consider these artificial samples as new observed samples.

We repeatedly resampling in this way (here 250 times), based on selecting a positive case to negative case ratio for a specific complication. To this end, we used a variant on the resampling approach in [17], whereby the bootstrapped pairs of time points are biased in the training data to have equal states where the comorbidities were present in time *t*. DBNs were trained on this bootstrap data and examined on their power to predict a complication at the next time point before latent variables were explored.

C. Model and structure

This paper proposes a model for the prognosis of major comorbidities of T2DM patients using a latent variable approach in a DBN framework. Data mining and analysis were performed using MATLAB and Bayes Net toolbox [32] and for visualisation we used Graphviz. The process of parameter estimation uses expectation-maximization (EM) algorithm [33]. For learning the structure of our model, firstly, we used the K2 [34] and REVerse Engineering ALgorithm (REVEAL) [35] to create non-temporal and temporal links, respectively. The networks with temporal associations inferred from time series T2DM historical patients data are represented in two DBNs (*t* and *t*-1). Therefore, the time-series data were analysed by DBNs learning models and balanced with a TS bootstrapping approach. Furthermore, we assumed that patient status at time *t* depends on the corresponding hidden variable at a previous time *t-1* (the general structure is shown in Figure 3). In Figure 3, we exploited a combination of the IC* algorithm [36] [37] and LinkStrength (LS) [38] to identify the location of hidden variables. The LinkStrength is a metric we have exploited in this paper to calculate the overall strength of dependent links so we can focus on the most powerful dependencies between T2DM risk factors. By setting the value of True Average Link Strength greater than 20 percent, though, overfitting in the DAG can be reduced. In our discrete-space/discrete-time DBNs, two-time steps is considered to show the relationship between risk factors. For instance, Figure 3 (right side) shows the first comorbidity at time *t*-1, which affects states of all other comorbidities and risk factors at time *t*.

*D. The IC*LS methodology*

1) Link Strength methodology: It finds a structure for locating latent variables within a Bayesian structure. We employed local and global sensitivity analysis [28] that consists of Mutual Information (MI), and Link Strength (LS). We exploited the following measure for measuring the uncertainty in the discovered models from the data:

- Entropy, which introduced in [39] to measure the uncertainty in a single node.

$$U(X) = \sum_{x_i} P(x_i) \log_2 \frac{1}{P(x_i)}. \tag{2}$$

- Mutual Information is a way of inferring links in data and measures connection strength [39] [40]. The MI between node X and node Y is uncertainties in Y that we decrease by knowing the state of X (and vice verse):

$$MI(X,Y) = U(Y)U(Y|X), \tag{3}$$

where $U(Y|X)$ is calculated by averaging $U(Y|x_i)$ over all possible states x_i of X, taking $P(x_i)$ into account:

$$MI(X,Y) = \sum_{x,y} P(x,y) \log_2 (\frac{P(x,y)}{P(x)P(y)}). \tag{4}$$

- The Link Strength [41] measure enables us to observe the specific impact of each discovered edge. Moreover, the percentage points of uncertainty reduction in Y are utilised by knowing the state of X if the states of all other parent variables are known. There are two types of LS in measuring uncertainties, True Average Link Strength (LSTA), and Blind Average Link Strength (LSBA).
- The LSTA calculates LS based on the average over the parent states using their actual joint probability. For a node with only one parent, MI Percentage and LSTA Percentage yield the same value. LSTA of edge $X \to Y$ is defined as the MI of (X,Y) conditioned on all other parents of Y, Which shown as:

$$LSTA(X \to Y) = \text{requires} \quad \text{P(all parents of Y)} \tag{5}$$

$$= MI(X,Y|Z) = U(Y|Z)U(Y|X,Z),$$

where $U(Y|X,Z)$ is the average over the states of all parents and $U(Y|Z)$ is the average over all other parents.
- The LSBA is derived from LSTA, but ignores the actual frequency of occurrence of the parent states. Thus, in LSBA measure, all parents are assumed to be independent of each other and uniformly distributed.

$$LSBA(X \to Y) = \text{requires} \quad \text{no inference at all} \tag{6}$$

The same probabilities as the corresponding absolute measure above is converted to each percentage measure. For removing all uncertainty, we require deterministic functions, in which the state of a child is completely known if the states of all of its parents are known.
Representing all parents from Y in $MI(X,Y|Z)$ in Equation(5) essentially blocks all information flow through the other parents, Z. According to [38], we are confident that there are no other indirect open links between Y and X, e.g. through descendants of Y, once all different parents are instantiated is the direct link from X to Y. Theorem: Consider a BN (G,P) consisting of DAG G and joint probability P. Let X Y be an edge in G and denote the set of all other parents of Y as Z. Let G% be the modified DAG generated by deleting edge X Y in G. Then X and Y are conditionally independent given Z in BN (G%,P%) for any joint probability P%. As indicated by the LSTA on the diagram, most links are quite strong except for those with LSTA less or equal to zero (removed from the final structure) can be classified as significant.

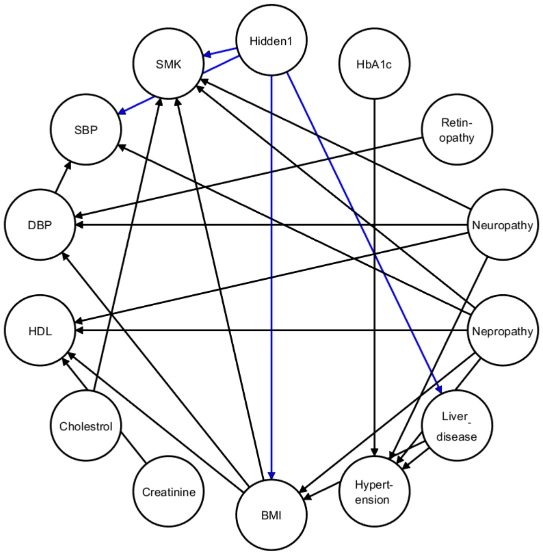

Fig. 2. IC*LS graph obtained from a combination of an IC* adjacent matrix and filtered edges from LS approach.

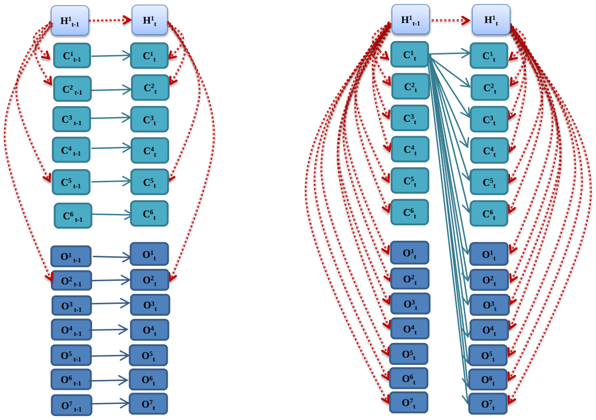

Fig. 3. Two time-series (t and t-1) structures using IC*LS approach and Fully Auto-Regressive dynamic links on the left hand side. On the right hand side, the latent variable is pointing to all comorbidities and observed nodes. Dynamic links are learned from the REVEAL algorithm. The H, C, and O illustrate Hidden nodes, Comorbidities, and Observed nodes, respectively.

2) Induction Causation (IC) and PC algorithms:* The IC* is a constraint-based method which calculates several conditional independence tests and combine them to create a Directed Acyclic Graph (DAG) to characterise the entire Markov equivalence class [37]. The probability of a high state of any learned hidden variables is then inferred using a standard Bayesian network inference. The IC* algorithm is similar to the PC algorithm, except that the former can detect the presence of latent variables. Furthermore, it learns a latent variable structure associated with a set of observed variables. The latent variable structure is a projection in which every latent variable is either a root node or a link to observed variables.

3) Latent Structures: Some variables are unmeasured but have potential to be developed, called hidden or latent variables. Based on Pearl's causality a latent structure is a pair $L =< D, O >$, where D is a causal structure over variables that O is a set of observed variables. IC* algorithm returns a marked pattern, a partially directed acyclic graph that contains four types of edges:

1) a marked arrow $O_1 \overset{*}{\rightarrow} O_2$, signifying a directed path from two observed node (O_1 and O_2) in the underlying latent structure (and there is no latent common cause for these two nodes).
2) a bi-directed edge $O_1 \longleftrightarrow O_2$, signifying a latent common cause $O_1 \leftarrow L \rightarrow O_2$ in the underlying latent structure or there is an inducing path between two variables, thus there is no directed path between them.
3) an unmarked arrow $O_1 \rightarrow O_2$, signifying either a directed path from O_1 to O_2 or a bi-directed edge; and
4) an undirected edge $O_1 -- O_2$, standing for either $O_1 \rightarrow O_2$ or $O_1 \leftarrow O_2$ or $O_1 \longleftrightarrow O_2$.

In our approach, Figure 1, we used a combination of the IC* algorithm and the LS method to discover locations of the latent variable and static links within the DBN structure (which we call IC*LS). First a correlation matrix from the dataset (via IC* algorithm) was found. Then, we filter existing edges in the IC* structure by assessing the LSTA percentages obtained from LS. To this end, the links in the IC* were removed where the LSTA percentage contributed less than a threshold of 20% and lower. If there was a link in the IC* adjacent matrix and its LSTA was greater or equal to 20 percent, then a link in the final structure was kept, otherwise, it was deleted. We chose this threshold to avoid providing overly connected networks and loops in the DAG, as well as to decrease the risk of edge overfitting. Moreover, by changing the threshold to less than 20 percent e.g., to 5 percent, no links were removed.

Furthermore we believe that Link Strength measures may be useful in the context of constraint-based structure learning algorithms to derive hypotheses of a systems primary causal pathways from data. We believe that link strengths measures could also be used to evaluate the quality of structure learning algorithms. Currently structure learning algorithms are evaluated by counting the number of incorrect arrows when identifying known systems. We believe that it may be more appropriate to weigh those counts by the link strength of the incorrect arrows.

IV. Experimental results

We evaluated the proposed structure by performing the sensitivity analysis on the cohort of patients who are diagnosed as having the T2DM disease. Moreover, we demonstrate the model using our balancing approach to remove the bias from the data, which leads to a huge improvement in the classification accuracy. Two key elements are used in our proposed approach; first, the bootstrap balancing, and second, consideration of a latent variable. We calculate The experiments presented in this section evaluate the effectiveness of

TABLE II
VISIT-BASED ASSESSMENT BASED ON COMPARATIVE CHARACTERISTICS OF THE DIFFERENT APPROACHES: 1) IMBALANCED USING THE K2 AND REVEAL; 2) BALANCED DATA USING THE K2 AND REVEAL AND ONE LATENT VARIABLE; 3) BALANCED DATA AND USING IC*; 4) BALANCED DATA USING IC*LS.

	UNB-K2-REVEAL	B-K2-REVEAL	IC*	IC*LS
Balanced data	NO	YES	YES	YES
Fully Auto-Regressive	NO	NO	YES	YES
Latent variable	YES	YES	YES	YES
K2 and REVEAL algorithm	YES	YES	NO	NO
IC* algorithm	NO	NO	YES	NO
IC*LS algorithm	NO	NO	NO	YES
AUC of retinopathy	0.35	0.50	0.89	0.99
AUC of liver disease	0.38	0.51	0.92	0.99
AUC of Hypertension	0.60	0.51	0.83	0.99

TABLE III
PATIENT-BASED PREDICTION ACCURACY FOR TWO COMPLICATIONS (RETINOPATHY AND LIVER DISEASE) IS SHOWN WITH PERCENTAGE OF PATIENTS THAT ARE DIAGNOSED EARLY, LATE, CORRECTLY, INCORRECTLY BASED ON DIFFERENT STRUCTURE LEANING METHODS (I.E., THE K2 AND IC*LS). IN ADDITION, WE REVEAL THE NUMBER OF FALSE POSITIVES, FALSE NEGATIVES, TRUE POSITIVES IN THE EARLY PREDICTION, TRUE POSITIVES IN THE LATE PREDICTION, TRUE NEGATIVES IN THE EARLY PREDICTION, AND TRUE NEGATIVES IN THE LATE PREDICTION.

Prediction	Retinopathy	Retinopathy	Liver disease	Liver disease
Accuracy	K2	IC*LS	K2	IC*LS
Early	90	95	92	97
Late	2	2	0	0
Hit	0	2	2	1
Miss	8	1	8	2
FP	0	1	0	0
FN	8	2	8	2
TP-Early	1	0	2	2
TP-Late	24	25	18	18
TN-Early	68	73	75	81
TN-Late	2	2	0	0

the proposed method that has these two aspects. In addition, we evaluated our model based on two different perspectives; 1) visit-based validation test and 2) Patient-based validation test.

A. *Visit-based Validation*

In this section, we assume all visits per each patient should be assess to get the classification accuracy. The results are documented for the following structure comparisons:

- UNB-K2-REVEAL: Using a latent variable and a fully learned structure from the K2 and the REVEAL algorithms with imbalanced data. (The latent variable pointed to all the observed nodes.)
- B-K2-REVEAL: Using a latent variable and a fully learned structure from the K2 and REVEAL algorithm with balanced data. (The latent variable pointed to all the observed variables.)
- NO-latent-K2-REVEAL: Using no latent variable, but data is balanced and model fully learned from K2 and REVEAL algorithm.
- IC*: Using a latent variable and the structure obtained from the IC* algorithm for Intra links and Fully Auto-Regressive for Inter links.
- IC*LS: Using a latent variable and the structure obtained from the IC*LS approach for Intra links and Fully Auto-Regressive for Inter links (Figure 3).

Table II summarizes characteristics of four structure learning methodology, which were exploited for testing our proposed approach. In this table, each column represents a different structure for modelling data and each row provides details in balancing data, choosing static / dynamic edges, and the use of the latent variable. The evaluation of the structures is included in terms of the AUC (Area Under the Curve) [42] in the last three rows of Table II. The dataset is split into two halves: one-half is used for balanced and trained for classification of a specific comorbidity, another half is considered as an independent testing data set to assess the performance of the classifier. For example, the testing dataset is not used in the structure introduced in Table II. Table II shows that bootstrapping for data balancing provides more accurate prediction results than unbalanced models (comparing the AUC of UNB-K2-REVEAL and B-K2-REVEAL). This is also clear from the true positive and true negative percentages as well as the classification accuracy represented in confusion matrices in Figures 3-5. In addition, in UNB-K2-REVEAL and B-K2-REVEAL, the structure of risk factors was learned using the K2 and the REVEAL algorithm. In these structures, the latent variable pointed to the entire set of risk factors within the same time slice. The latent variable in slice t depended on its latent variable in slice t-1 for all structures. The confusion matrices for the Our results in prediction of comorbidities reveal an increase in the classification accuracy of the IC*LS approach for locating the latent variable within the risk factor set. In contrast, without using a hidden variable, classification results and the AUC have dropped considerably (B-K2-REVEAL versus IC*LS). The IC*LS approach to the prediction of the comorbidities is more precise than simply using the IC* algorithm itself, comparing the AUC for IC* and IC*LS structures in Table II. This is likely to be because the LS is filtering the less robust links, thus avoiding overfitting.

B. *Patient-based Validation*

In Table III, we explore the effect of early/late time prediction of comorbidities. We indicate the diagnosis point as a switch from a state that the specific comorbidity has not been diagnosed to a state where the comorbidity is developed. The results of class confusion matrices and the prediction accuracy is shown in terms of percentage of patients, which are diagnosed early (Early), late (Late), correctly in the same time as the diagnosis time (Hit), incorrectly (Miss). Moreover, results from confusion matrices is retrieved as number of False Positives (FP), False Negatives (FN), True positives in the early prediction (TP-Early), True Positives in the late prediction (TP-Late), True Negatives in the early prediction (TN-Early), True Negatives in the late prediction (TN-Late). These performance measures are obtained for two complications (retinopathy and liver disease) and comparing classification accuracy based on different structure learning methods i.e, K2 and IC*LS. We compare classification performance measures of using all links obtained from the IC* correlation matrix with IC*LS to see the effect of this reduced link strength that is reported in Table III. To this end, we compare a vector of retinopathy status in the first four visits of patients to a vector of the first four prediction outcomes. The results in Table III show that a better prediction accuracy of comorbidities is generally achieved when using IC*LS (the total number of early time predictions is higher than the late time prediction and the classification accuracy is increased from 90 to 95 percent.

C. Latent variable validation Pattern

We exploit this work to figure out how fluctuations in the state of a latent variable are impacted by various comorbidities. We now explore the same case studies for individual patients. Figure 4 reveals the prediction pattern for liver disease before the clinician classified the diagnosed patient. In Figure 4, the x-axis represents the follow-up visits and the y-axis represents the probability of a patient is diagnosed with the comorbidity (Liver disease). The red dashed vertical line shows the exact time that a comorbidity is diagnosed from actual clinical results. In Figure 4 a, observed liver disease shows the clinician diagnosed the patient with the T2DM comorbidity at the fourth follow-up visit. Figure 4 b (Latent-B-K2-REVEAL-liver), representing the probability of the latent variable in its high state, does not seem to coincide with the appearance of liver disease. This indicates that B-K2-REVEAL is not capturing a behaviour associated with this comorbidity. Figure 4 c, however, shows the latent variable found using IC*LS STRUCTURE demonstrates a ”switching behaviour” just before the comorbidity appearance as seen in Figure 4 a, indicating that the IC*LS approach has found a set of latent variable relationships that enable an improved prediction as seen in Figure 4 d (representing the correct point in predicting the comorbidity diagnosis). Another interesting result that was observed occasionally was where the $K2-REVEAL$ and IC*LS classified the patient as being diagnosed. Then it began to fluctuate from one visit to the next before settling on a high probability of T2DM, shortly before the clinician classifies the patient as diagnosed (for example, see Figure 5 d). This implied that some aspects in the behaviour of the latent variable might capture an early warning sign of comorbidity onset. However, IC*LS approach did not always improve prediction. For example, in Figure 6 for many visits (Figure 6 c) the probability, for hypertension, remained low and then began to fluctuate after the clinician has classified the patient.

D. Latent variable as an evidence

In order to explore the impact of the targeted latent variable from IC*LS, we considered the distribution of comorbidities in the DBN conditioned as observed latent variable in different

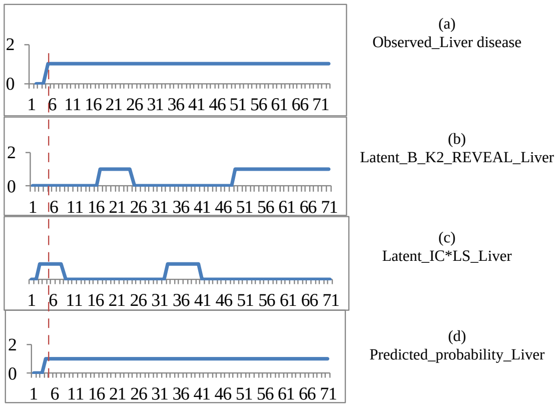

Fig. 4. Latent variable prediction pattern of liver disease for the patient visits (Early time prediction by using the IC*LS approach).

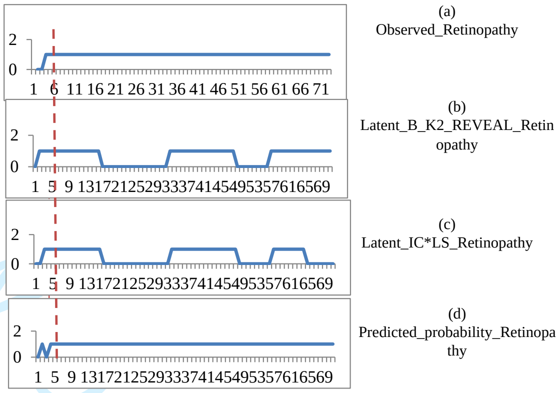

Fig. 5. Latent variable prediction pattern of retinopathy for the patient visits (Early time prediction using IC*LS).

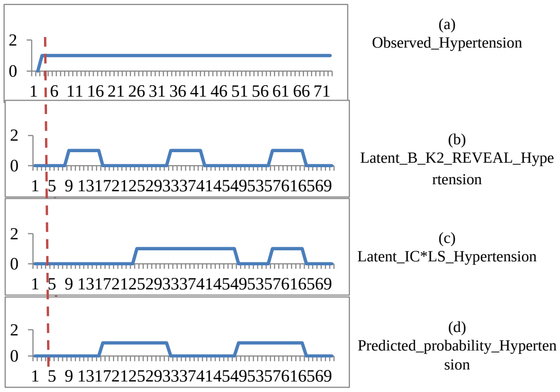

Fig. 6. Latent variable prediction pattern of hypertension for the patient visits (Late time prediction using IC*LS).

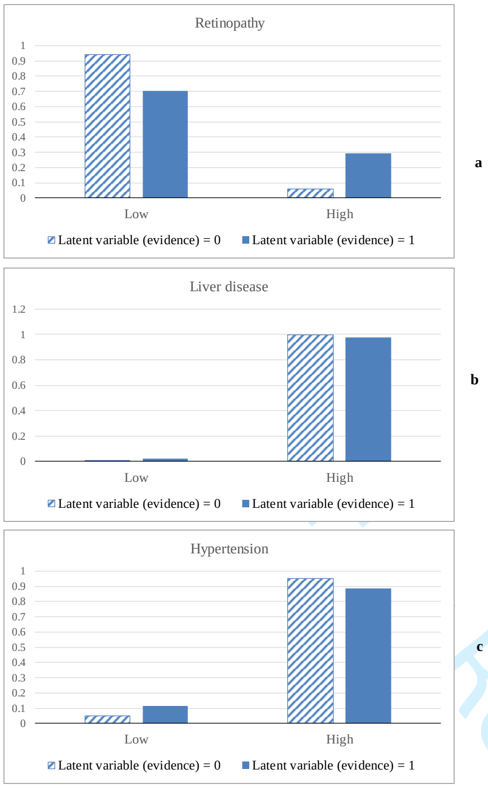

Fig. 7. Prediction probabilities for retinopathy, liver disease and hypertension with using Latent variable as the evidence.

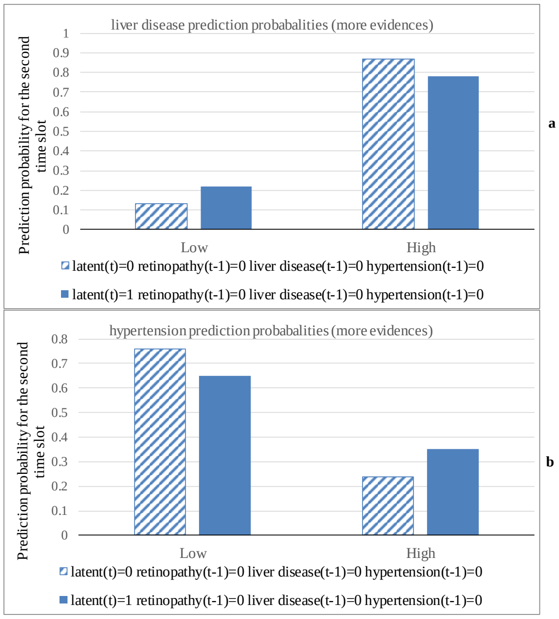

Fig. 8. Prediction probabilities for liver disease and hypertension by exploiting more evidences.

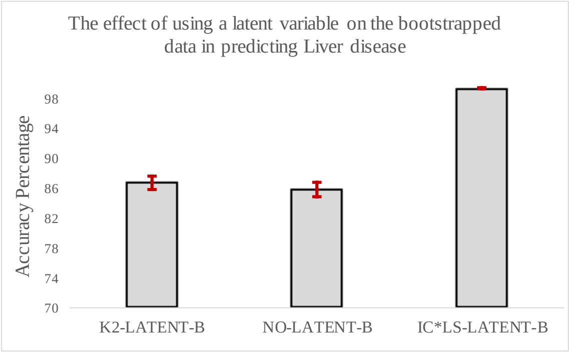

Fig. 9. Bootstrap Confidence Interval representing cluster chart and standard error for calculating accuracy in liver disease.

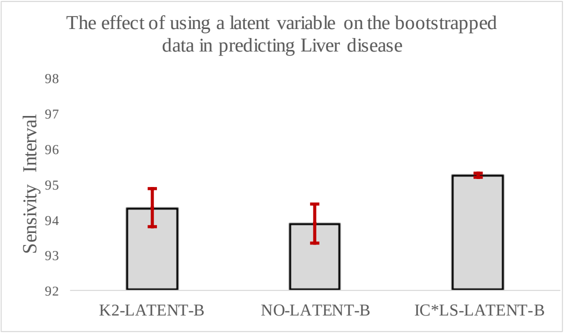

Fig. 10. Bootstrap Confidence Interval representing cluster chart and standard error for calculating Sensitivity in liver disease.

states as evidence. This was made possible due to the nature of Bayesian network inference where any node could be queried. The bar chart in Figure 7-a illustrates the prediction probability of comorbidities regarding the second-time slot while varying the latent variable value as the evidence.

Here, the result of the analysis in Figure 7 at first indicates that the relationship between the latent variable to the T2DM risk factors is not straightforward as changing the hidden state only has a small impact on the distributions. However, when focusing on the switch from pre-diagnosis to post-diagnosis (by entering no comorbidity at time t-1 in the DBN model), it becomes clearer that the hidden variable impacts the distributions (see Figure 8) with a decrease in the probability of liver disease when the hidden state is high (Figure 8 a) and an increase in the probability of hypertension (Figure 8 b). These results suggest that whilst there is an association between the latent variable and common complications in the prognosis of the T2DM patients, this relationship is complex. Interpreting these latent variables is complicated as they may represent different types of predictors, such as life expectancy, quality of life, or some combination of these risk factors with comorbidities.

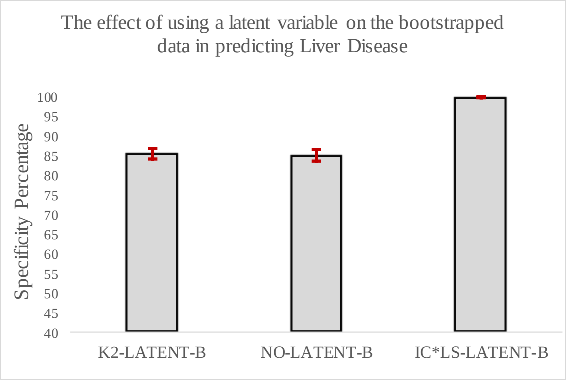

Fig. 11. Bootstrap Confidence Interval representing cluster chart and standard error for calculating specificity in liver disease.

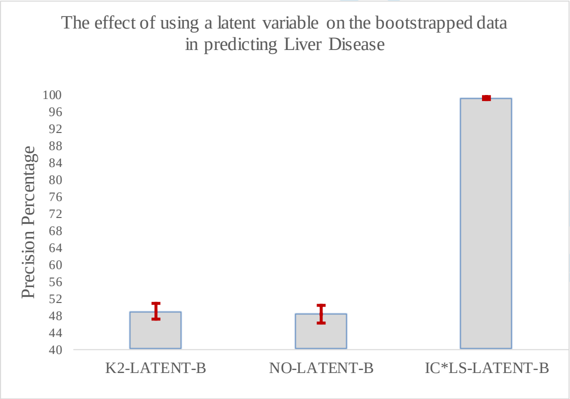

Fig. 12. Bootstrap Confidence Interval representing cluster chart and standard error for calculating precision in liver disease.

E. Confidence Interval Results

In the study of the bootstrap in related literature, there are some articles that deal with bias in the data and enhance convergence accuracy [43] [44] [45]. In situations in which asymptotic confidence intervals are known to apply and are correct, bias-corrected and accelerated (BCa) confidence intervals have been demonstrated to show faster convergence and increased accuracy over ordinary percentile-based methods, while retaining the desirable property of robustness (see, e.g., [46]). This section reveals the influence of using a latent variable on the bootstrapped data in predicting two common complications of T2DM. A clustered column chart in Figure9 compares the accuracy percentages average, among 250 times bootstrap, for three methods (K2-LATENT-B, IC*LS-NO-LATENT-B, and IC*LS-LATENT-B), which are 88, 86 and 99 percent, respectively. We illustrates error bars on the top of bar charts in Figure9. These results reveal that IC*LS-LATENT-B has higher accuracy comparing to IC*LS-NO-LATENT-B and K2-LATENT-B, while IC*LS-NO-LATENT-B error bar is bigger than others. The error bar for K2-LATENT-B is quite big due to a bigger confidence interval of IC*LS-NO-LATENT-B. However, a smaller confidence interval in IC*LS-LATENT-B makes the corresponding error bar consistent. Whenever the error bars overlap e.g., in the sensitivity analysis shown in Figure10, which lower bottom of the error bar in K2-LATENT-B is lower than the top of error bar in IC*LS-LATENT-B then statistically speaking these two averages are not different, even though the means themselves are totally different. In contrast, error bars in IC*LS-LATENT-B and IC*LS-NO-LATENT-B do not overlap, whereas the results of means shows they are pretty much the same average , then we can say they are statistically different. We are confident more than 95 percent that IC*LS-NO-LATENT-B has lower accuracy, sensitivity, specificity than K2-LATENT-B. We are at least 95 percent sure that exploiting IC*LS-LATENT-B caused a huge and significant improvement in classification precision, accuracy, sensitivity and precision comparing to K2-LATENT-B (see Figure12,9,10 and 11).

V. Conclusion

Diabetes specialists predict disease and comorbidities based on their knowledge of the disease and an individual patient's clinical history. This is a complex task because of the existence of unmeasured risk factors in the data, various responses to the disease and heterogeneity in monitoring patients. Here, we make a first step in modelling unmeasured factors by considering an approach to model progression using latent variables with a focus on trying to understand their behaviour and meaning. We exploited Dynamic Bayesian Networks due to their transparent way of modelling data as well as their flexibility in incorporating latent variables. We incorporate the IC* algorithm and a Mutual Information based scoring metric to identify the strength of relationships between the latent variable and clinical factors. Firstly, our results showed that re-balancing data demonstrates an improvement in classification. Secondly, this paper attempted to gain insight by interpreting the latent states (looking at the associated distributions of comorbidities), which leaded to a better understanding of risk factors and patient-specific interventions.

VI. Future work

Ultimately, we intend to use these latent statuses to facilitate us to identify different cohorts of patients who have various dynamics and therefore stratify them so that more can be inferred around the different expressions of the disease and its advancement. Information that demonstrates the way to improve the quality of life of patients using predictive models is limited. A further study should concentrate on the continuous investigation of T2DM features while modelling DBNs rather than a discrete space/discrete time model. Another natural progression of this work involves exploring the extension of these models with more latent variables, to capture a greater variety of factors that characterize key changes in the clinical and complications data. Our proposed approach will be useful for stratifying patients according to their probability of developing complications and clinician advice.

Leila Yousefi a PhD researcher and Graduate Teacher Assistant in computer science and engineering at Brunel University London, UK. She hold a bachelor of software engineering and Master of IT engineering from Iran where she was a lecturer at computer science and IT engineering department. She is a member of the Intelligent Data Analytics (IDA) Research Group at Brunel University London. Her research interest is in Artificial Intelligence, Biomedical, disease prediction and Data Mining. Currently, she works as a research fellow at Brunel University London.

Mashael Al Luhaybi a Ph.D. candidate in Machine learning in particular Educational Data Mining at Brunel University, London, U.K. She obtained her MSc from the University of Brighton, UK in 2011. Her research interests lie in modelling student academic performance using classification algorithms. She is also interested in temporal clustering of students' online trajectories in Learning Management Systems in order to detect students' engagement levels and learning patterns.

Lucia Sacchi an Assistant Professor (RTDB), Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy.

Luca Chiovato a MD, PhDProfessor of Endocrinology Internal Medicine and Medical Therapy University of Pavia. He is head, Unit of Internal Medicine and Endocrinology Fondazione Salvatore Maugeri I.R.C.C.S Pavia, Italy

Allan Tucker a Senior Lecturer at Brunel University London, United Kingdom. He is the Head of Intelligent Data Analytics (IDA) Research Group at Brunel University. His research interests lie in modelling of brain function, human and animal behaviour. He obtained his PhD from Birkbeck, University of London.

References

[1] C. D. Mathers and D. Loncar, "Projections of global mortality and burden of disease from 2002 to 2030," *PLoS medicine*, vol. 3, no. 11, p. e442, 2006.

[2] D.-C. Suh, I.-S. Choi, C. Plauschinat, J. Kwon, and M. Baron, "Impact of comorbid conditions and race/ethnicity on glycemic control among the us population with type 2 diabetes, 1988–1994 to 1999–2004," *Journal of Diabetes and its Complications*, vol. 24, no. 6, pp. 382–391, 2010.

[3] A. N. Long and S. Dagogo-Jack, "Comorbidities of diabetes and hypertension: mechanisms and approach to target organ protection," *The journal of clinical hypertension*, vol. 13, no. 4, pp. 244–251, 2011.

[4] R. Raman, A. Gupta, S. Krishna, V. Kulothungan, and T. Sharma, "Prevalence and risk factors for diabetic microvascular complications in newly diagnosed type ii diabetes mellitus. sankara nethralaya diabetic retinopathy epidemiology and molecular genetic study (sn-dreams, report 27)," *Journal of Diabetes and its Complications*, vol. 26, no. 2, pp. 123–128, 2012.

[5] M. A. Van Gerven, B. G. Taal, and P. J. Lucas, "Dynamic bayesian networks as prognostic models for clinical patient management," *Journal of biomedical informatics*, vol. 41, no. 4, pp. 515–529, 2008.

[6] M. Heijden, M. Velikova, and P. Lucas, "Learning bayesian networks for clinical time series analysis," 2014.

[7] Q. Yang and X. Wu, "10 challenging problems in data mining research," *International Journal of Information Technology & Decision Making*, vol. 5, no. 04, pp. 597–604, 2006.

[8] G. Liang, "An effective method for imbalanced time series classification: Hybrid sampling," in *Australasian Joint Conference on Artificial Intelligence*. Springer, 2013, pp. 374–385.

[9] N. Friedman, K. Murphy, and S. Russell, "Learning the structure of dynamic probabilistic networks," in *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1998, pp. 139–147.

[10] G. Elidan, N. Lotner, N. Friedman, and D. Koller, "Discovering hidden variables: A structure-based approach," in *Advances in Neural Information Processing Systems*, 2001, pp. 479–485.

[11] E. Mueller, S. Maxion-Bergemann, D. Gultyaev, S. Walzer, N. Freemantle, C. Mathieu, B. Bolinder, R. Gerber, M. Kvasz, and R. Bergemann, "Development and validation of the economic assessment of glycemic control and long-term effects of diabetes (eagle) model," *Diabetes technology & therapeutics*, vol. 8, no. 2, pp. 219–236, 2006.

[12] K. P. Murphy and S. Russell, "Dynamic bayesian networks: representation, inference and learning," 2002.

[13] A. Dagliati, A. Malovini, P. Decata, G. Cogni, M. Teliti, L. Sacchi, C. Cerra, L. Chiovato, and R. Bellazzi, "Hierarchical bayesian logistic regression to forecast metabolic control in type 2 dm patients," in *AMIA Annual Symposium Proceedings*, vol. 2016. American Medical Informatics Association, 2016, p. 470.

[14] A. Dagliati, A. Marinoni, C. Cerra, P. Decata, L. Chiovato, P. Gamba, and R. Bellazzi, "Integration of administrative, clinical, and environmental data to support the management of type 2 diabetes mellitus: From satellites to clinical care," *Journal of diabetes science and technology*, vol. 10, no. 1, pp. 19–26, 2016.

[15] S. Marini, E. Trifoglio, N. Barbarini, F. Sambo, B. Di Camillo, A. Malovini, M. Manfrini, C. Cobelli, and R. Bellazzi, "A dynamic bayesian network model for long-term simulation of clinical complications in type 1 diabetes," *Journal of biomedical informatics*, vol. 57, pp. 369–376, 2015.

[16] L. Sacchi, A. Dagliati, D. Segagni, P. Leporati, L. Chiovato, and R. Bellazzi, "Improving risk-stratification of diabetes complications using temporal data mining," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE, 2015, pp. 2131–2134.

[17] N. Trifonova, A. Kenny, D. Maxwell, D. Duplisea, J. Fernandes, and A. Tucker, “Spatio-temporal bayesian network models with latent variables for revealing trophic dynamics and functional networks in fisheries ecology,” *Ecological Informatics*, vol. 30, pp. 142–158, 2015.

[18] A. Tucker, X. Liu, and D. Garway-Heath, “Spatial operators for evolving dynamic bayesian networks from spatio-temporal data,” in *Genetic and Evolutionary ComputationGECCO 2003*. Springer, 2003, pp. 205–205.

[19] S. Ceccon, D. Garway-Heath, D. Crabb, and A. Tucker, “The dynamic stage bayesian network: identifying and modelling key stages in a temporal process,” *Advances in Intelligent Data Analysis X*, pp. 101–112, 2011.

[20] M. Talih and N. Hengartner, “Structural learning with time-varying components: tracking the cross-section of financial time series,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 3, pp. 321–341, 2005.

[21] J. W. Robinson and A. J. Hartemink, “Learning non-stationary dynamic bayesian networks,” *Journal of Machine Learning Research*, vol. 11, no. Dec, pp. 3647–3680, 2010.

[22] M. Grzegorczyk and D. Husmeier, “Non-stationary continuous dynamic bayesian networks,” in *Advances in Neural Information Processing Systems*, 2009, pp. 682–690.

[23] A. I. Adler, I. M. Stratton, H. A. W. Neil, J. S. Yudkin, D. R. Matthews, C. A. Cull, A. D. Wright, R. C. Turner, and R. R. Holman, “Association of systolic blood pressure with macrovascular and microvascular complications of type 2 diabetes (ukpds 36): prospective observational study,” *Bmj*, vol. 321, no. 7258, pp. 412–419, 2000.

[24] A. Lloyd, W. Sawyer, and P. Hopkinson, “Impact of long-term complications on quality of life in patients with type 2 diabetes not using insulin,” *Value in Health*, vol. 4, no. 5, pp. 392–400, 2001.

[25] L. Yousefi, A. Tucker, M. Al-luhaybi, L. Saachi, R. Bellazzi, and L. Chiovato, “Predicting disease complications using a stepwise hidden variable approach for learning dynamic bayesian networks,” in *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2018.

[26] L. Yousefi, L. Saachi, R. Bellazzi, and L. C. A. Tucker, “Predicting comorbidities using resampling and dynamic bayesian networks with latent variables.”

[27] H. Rosolova, B. Petrlova, J. Simon, P. Sifalda, I. Sipova, and F. Sefrna, “Macrovascular and microvascular complications in type 2 diabetes patients,” *Vnitrni lekarstvi*, vol. 54, no. 3, pp. 229–237, 2008.

[28] J. Khoo, T.-L. Tay, J.-P. Foo, E. Tan, S.-B. Soh, R. Chen, V. Au, B. J.-M. Ng, and L.-W. Cho, “Sensitivity of a1c to diagnose diabetes is decreased in high-risk older southeast asians,” *Journal of Diabetes and its Complications*, vol. 26, no. 2, pp. 99–101, 2012.

[29] R. Turner, H. Millns, H. Neil, I. Stratton, S. Manley, D. Matthews, and R. Holman, “Risk factors for coronary artery disease in non-insulin dependent diabetes mellitus: United kingdom prospective diabetes study (ukpds: 23),” *Bmj*, vol. 316, no. 7134, pp. 823–828, 1998.

[30] N. Moniz, P. Branco, and L. Torgo, “Resampling strategies for imbalanced time series,” in *Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on*. IEEE, 2016, pp. 282–291.

[31] L. SIMAR, “An invitation to the bootstrap: Panacea for statistical inference?” *Institut de Statistique, Universite Catholique de Louvain, Louvain*, 2008.

[32] K. Murphy *et al.*, “The bayes net toolbox for matlab,” *Computing science and statistics*, vol. 33, no. 2, pp. 1024–1034, 2001.

[33] T. K. Moon, “The expectation-maximization algorithm,” *IEEE Signal processing magazine*, vol. 13, no. 6, pp. 47–60, 1996.

[34] G. F. Cooper and E. Herskovits, “A bayesian method for the induction of probabilistic networks from data,” *Machine learning*, vol. 9, no. 4, pp. 309–347, 1992.

[35] S. Liang, S. Fuhrman, and R. Somogyi, “Reveal, a general reverse engineering algorithm for inference of genetic network architectures,” 1998.

[36] J. Pearl, *Causality*. Cambridge university press, 2009.

[37] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, prediction, and search*. MIT press, 2000.

[38] I. Ebert-Uphoff, “Measuring connection strengths and link strengths in discrete bayesian networks,” Georgia Institute of Technology, Tech. Rep., 2007.

[39] C. E. Shannon, W. Weaver, and A. W. Burks, “The mathematical theory of communication,” 1951.

[40] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

[41] N. Jitnah, *Using Mutual Information for Approximate Evalutation of Bayesian Networks*. Monash University, 1999.

[42] A. P. Bradley, “The use of the area under the roc curve in the evaluation of machine learning algorithms,” *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[43] T. J. DiCiccio and B. Efron, “Bootstrap confidence intervals,” *Statistical science*, pp. 189–212, 1996.

[44] F. A. Wichmann and N. J. Hill, “The psychometric function: Ii. bootstrap-based confidence intervals and sampling,” *Perception & psychophysics*, vol. 63, no. 8, pp. 1314–1329, 2001.

[45] H. DAVID, “Bootstrap estimates of the statistical accuracy of thresholds obtained from psychometric functions*ˆ,” *Spatial Vision*, vol. 11, no. 1, pp. 135–139, 1997.

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