Identifying Latent Variables in Dynamic Bayesian Networks with Bootstrapping Applied to Type 2 Diabetes Complication Prediction

Leila Yousefi (Leila.yousefi@brunel.ac.uk)
Brunel University https://orcid.org/0000-0003-1952-0674
Mashael Al-Luhaybi
Brunel University
Lucia Sacchi
Universita degli Studi di Pavia Dipartimento di Studi Umanistici
Luca Chiovato
Fondazione Salvatore Maugeri Istituto Scientifico di Pavia Via Maugeri
Allan Tucker
Brunel University

Research article

Keywords: Latent Variable, Diabetes, Dynamic Bayesian Networks, Time series Bootstrapping, Decision Making, Disease Complications

Posted Date: September 18th, 2020

DOI: https://doi.org/10.21203/rs.2.24145/v2

RESEARCH

Identifying Latent Variables in Dynamic Bayesian Networks with Bootstrapping Applied to Type 2 Diabetes Complication Prediction

Leila Yousefi[1*], Mashael Al-Luhaybi[1], Lucia Sacchi[2], Luca Chiovato[3] and Allan Tucker[1]

*Correspondence:
Leila.Yousefi@brunel.ac.uk
[1]Department of Computer Science, Brunel University London, Kingston Ln, London, Uxbridge, UB8 3PH, UK
Full list of author information is available at the end of the article
†Equal contributor

Abstract

Background: Type 2 Diabetes is a chronic disease with an onset that is commonly associated with multiple life-threatening comorbidities (complications). Early prediction of diabetic complications while discovering the behaviour of associated aggressive risk factors can reduce the patients' suffering time.
Therefore, models of the time series diabetic data (which are often imbalanced, incomplete and involve complex interactions) are needed to better manage diabetic complications.

Aims: The aim of this work is to both deals with imbalanced clinical data using a bootstrapping approach, whilst determining the precise position of latent variables within probabilistic networks generated from the observations. The main motivation behind this paper is to stratify patient groups by means of latent variables to discover how complications in diabetes interact.

Methods: We propose a time series bootstrapping method for building Dynamic Bayesian Networks that includes hidden/latent variables, applied to a case for predicting T2DM complications. A combination of the IC* algorithm on time series bootstrapped data is utilised to identify the latent variables within a Bayesian model. Then, an exploration of inference methods assessed the influences of these latent variables.

Results: Our promising findings show how this targeted use of latent variables improves prediction accuracy, specificity, and sensitivity over standard approaches as well as aiding the understanding of relationships between these latent variables and disease complications/risk factors. The contribution of this paper compared to the previous papers in which time series bootstrapping is used for re-balancing the data and providing confidence in the prediction results.

Conclusion: Our results showed that our re-balancing approach by the use of Time Series bootstrapping method for an unequal number of time series visits demonstrated an improvement in the prediction performance. Additionally, the most highlighted contribution of this paper gained insight by interpreting the latent states (looking at the associated distributions of complications), which led to a better understanding of risk factors and patient-specific interventions: here the fact that the latent variable demonstrated that a patient falls into a
sub-group that is hypertensive but not suffering from retinopathy.

Keywords: Latent Variable; Diabetes; Dynamic Bayesian Networks; Time series Bootstrapping; Decision Making; Disease Complications

Background

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 common form of diabetes, accounting for at least 90 per cent of all instances. The World Health Organisation reported that in the next 11 years there will be about 550 million people suffering from this disease [1]. T2DM 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. For instance, hypertension is a major macrovascular disease risk factor [2].

It has previously been observed that patients with T2DM are also at an increased risk of microvascular comorbidities, including nephropathy, neuropathy, and retinopathy [3]. 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. This non-stationary characteristic of clinical data collected as part of the monitoring of T2DM, creates a difficult context for effective forecasting [4]. Clinical data needs to be considered as time series data to provide a description of the progression of a disease over time, but dealing with time series patient records is known to be a major issue in the prognosis of comorbidities [5], particularly when time series data is imbalanced and contains few examples of patients without comorbidities that are common to all patients. There are various methodologies for T2DM prediction, e.g, risk-prediction equations and Markov models [6]. However, the former suffers from uncertainty as well as only performing one-step-ahead predictions, while the latter is limited to a small number of discrete risk factors. Much of the existing literature on investigating the prognosis of T2DM complications [7] focuses on logistic regression and Naïve Bayes methods. Dagliati et al. [8] presented a Hierarchical Bayesian Logistic Regression model to anticipate patients changes when the individual model parameters are estimated. Research on T2DM prediction has often been restricted to modelling a limited number of visits. For example, in [8], external and internal heterogeneity were explored in T2DM patients for predicting comorbidities in cross-sectional data with just three horizons of time. Time series modelling was not employed in the individual measurements. Similarly, in [9], T2DM data was analysed to understand the influence of H2A1c and other T2DM risk factors in the development of the microvascular complications in 2-year time periods but did not model the data as a time-series. In another work [10], a Bayes Network to predict diabetes was proposed on the Pima Indian Diabetes dataset. However, the study failed to consider time series analysis. Similarly, a study [11] simulated the health state and complications by using Bayesian inferred models, applied to non-time series type 1 diabetes patient data.

Dynamic Bayesian Networks (DBNs) have been suggested as suitable models for handling uncertain, noisy clinical time series data [12]. What is more, DBNs are probabilistic graphical models that can handle missing data and hidden variables. 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 while discovering the behaviour of associated aggressive risk factors can help to improve

a patient's quality of life [13]. Therefore, discovering latent variables can potentially capture unmeasured effects from clinical data, simplifying complex networks of interactions and giving us a better understanding of disease processes. What is more, it can improve classification accuracy and boost user confidence in the classification models [14]. In [15] trees of hidden variables were used to render all observable variables independent and in [16], the authors emphasised the importance of the presence of hidden variables. They determined a hidden variable that interacts with observed variables and located them within the Bayesian Network structure. In addition, they showed that networks without hidden variables are clearly less useful because of the increased number of edges needed to model all interactions, which caused overfitting. In [17], authors provided a factor structure for learning methods that efficiently utilised hidden variables. Nevertheless, this method failed to consider prior belief in the factor structure and therefore, could not rely on the final structure.

As well as modelling unmeasured factors, hidden variables can also be used to model non-stationary processes. Many diseases involve structural changes based upon key stages in the progression, but many models do not appear to take this into account. In [18], clinical features were modelled using a second order time-series model but it was assumed that the temporal dependencies were time-invariant. There has been some work in extending DBNs to model underlying processes that are non-stationary [18]. Previous work on learning DBNs have inferred both network structures and parameters from (sometimes incomplete) clinical data sets [12]. For example, a recent study presented a similar DBN method to this study but to analyse fisheries data [19].

A paper [20] formalised non-stationary DBN models and proposed a Markov Chain Monte Carlo (MCMC) sampling algorithm for learning the structure of the model from time series biological data. Another work [21] 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. Authors in [22] estimated the variance in the data structure parameter with a MCMC approach, but the search space was limited to a fixed number of segments and indirect edges only, which is not suitable for T2DM data. Such studies remained narrow and 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.

Although extensive research has been carried out on the prediction of diabetic progression, no single study exists which has attempted to interpret the impact of latent (hidden) variables in the presence of diabetic disorders. In our previous work, an intuitive stepwise method to learn the latent effects was developed based upon the IC* algorithm, while using a Pair-Sampling re-balancing method [23]. In [24–26], patients were clustered into different sub-groups. In each sub-group, they shared a similar profile of observed risk factors, without taking account of the cluster decision making process. In this article, we expand our previous work [27] on disease progression modelling with latent variables while conducting a bootstrap to both balance the data and calculate confidence bounds. We also carry out an analysis using patient case studies. The first half of the paper is dedicated to explaining

how to balance time series clinical data and learn DBNs with and without latent variables. We compare the proposed method (a latent DBN model) to the standard Bayesian Network approaches using K2 and REVEAL algorithms (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 second half of the paper is dedicated to analysing the results, in terms of classification (predicting two comorbidities associated with T2DM), validation and the potential for adoption in clinical practice.

[Figure 1 about here.]
[Table 1 about here.]
[Table 2 about here.]

Method

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 latent DBN framework. Figure 1 represents the overall strategy to predict the complications. First, we obtained T2DM data and split it into training and test sets. Second, we applied the bootstrap approach to re-balance the data. Third, to learn the structure, the IC* algorithm, Fisher's Z test and the correlation matrix were used to generate a Directed Acyclic Graph (DAG). Moreover, links within the DAG were filtered based upon their strength. Then, a predictive model (IC*LS DBN) was exploited by utilising a latent variable approach. Finally, the findings were assessed by using a number of quantitative and qualitative validation strategies.

Data source

The data for this study consists of pre-diagnosed T2DM patients aged 25 to 65 years (inclusive) that were recruited from clinical follow-ups at the "IRCCS Instituti Clinic Scientifici" (ICS) Maugeri of Pavia, Italy. The MOSAIC project funds the data under the 7th 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 Cholesterol measurements and lipid profile, which is explained in the following section.

Variable selection

For this study, certain complications and risk factors (predictors) were selected based on existing literature on diabetes [28] 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).

Classification and Imbalanced Data

T2DM dataset is highly imbalanced based on the disease common complications. To deal with this firstly, it discretises risk factors into qualitative states: low, medium,

and high. In contrast, comorbidities have (imbalanced) binary states. T2DM dataset is highly imbalanced based on the disease common complications. To predict a target complication, patients are classified into two categories (cases): positive and negative cases. The outcome of the prediction or classification (Y) can be considered as a vector of disease risk factors represents by $Y = (X, C_i)$, where X $C = \{HY\ P, NEU, NEP, LIV, RET\}$. In this study, C_i only takes on binary values is the vector of symptoms, and C_i shows a target complication class selected from ($C_i = \{0 \mid 1\}$) as the main focus is to predict only one complication at time[1].

class value becomes zero ($C_i = 0$) otherwise it sets to one ($C_i = 1$) in which it shows for example, if a patient is diagnosed negatively (not having the complication), the that a patient is diagnosed positively (having a target complication). Considering

a specific (target) complication at each time point, by detecting any 1 in the class over all patient's visits is directed to join to the positive case otherwise the patient becomes a member of the negative case. This explains the "Patient-based" analysis. Once a complication is diagnosed, it is recorded for the rest of the patients

is identified as a positive case ($C_i = 1$), the patient stays in the corresponding time series as people are not recovered once diagnosed. Similarly, once a patient case throughout their time-series. As a result, those patients who are already at a

high risk of developing complications, it is assumed that they do not switch from positive case to the negative case. In particular, if the overall number of patients in the positive case is far less than the negative case, the complication class is labelled as an imbalanced class. In addition, the distribution of data is not equal in imbalance T2DM data. In this research, the minority class represents patients visit during which a complication is present and the patient that are suffering from the target complication. In fact, frequency of samples belonging to one class is severely different from the other ones. Therefore, binary classifiers bias to a class which demonstrates the majority of samples. A common problem with classifying complications in longitudinal data is that there may be many more visits where the complication does not manifest itself compared to those where it does (due to careful management). This explains "Visit-based" analysis in which each time point (as a single visit for a patient) is scored individually as zero or one for each patient[2]. Overall, proportional probability or likelihood ratios are reported (e.g., stating that one complication generated a certain result twice more probable than another complication), it seemed necessary to note the actual incidence at which the result happened. Thus, an unbalanced ratio is calculated as the ratio of negative to positive cases ({number of negative cases}: {number of positive cases}) for a specific complication to ensure a balance. For instance, the unbalance ratio of the

majority to the minority based on the population size of responding binary class proportions in the dataset for RET, NEU, NEP, and LIV are defined as ({3:1({, ({4:1}, ({3:1}, and ({4:1}, respectively. Whilst this ratio for HYP is ({1:5}.

---

[1] It is possible to show patterns of complications for each single visit with respect to any combination of complication co-occurrences, chosen from c as demonstrated in [26].

[2] For more information see Supplementary Material (Pre-processing and Data Structure).

Re-balancing Strategy (Time Series Bootstrapping)

A key aim of this study is to re-balance clinical data to improve our latent variable learning model due to the over representation of patients with specific comorbidities. To address this issue, so far, many methods from weighting, generating new samples to one class classifiers have been proposed. Different learning techniques deal with imbalanced data, such as oversampling, under sampling, boosting, bagging, bootstrapping, and repeated random sub-sampling [29]. In this study, bootstrap approach is adapted to identify the significant statistics from classifiers learnt from such data where the occurrence of the positive class is far less than the negative. This is because, Bootstrap re-balancing methods generally have been found to produce more accurate and reliable statistics [30]. Having considered the temporal and complex nature of T2DM data, the bootstrap approach in the longitudinal dataset is extended by re-sampling consecutive time points, thus enabling the (first-order) to be inferred. The proposed oversampling time-series Bootstrapping methodology is called "TS Bootstrapping" which employs a variant on the re-sampling approaches introduced in [5, 19, 31]. It re-samples the rare complication class with a replacement with respect to the dynamics of progression. The bootstrap pairs of time points

present ($C_i = 1$) than in the original data. Thus, the re-sampling approach of are selected with replacement to ensure more states where the complication is the data involves a bootstrap process to re-sample observed time-series/visits of

pairs of consecutive time points, $t - 1$ and t. It also assumes that patient status a patient with the replacement whereby the original training data is sampled in at time $t - 1$ depends on the corresponding hidden variable at a previous time t (Markov properties). As a result, the bootstrapped data contains an equal number of positive and negative cases for the target complication at time t. TS Bootstrapping approach seems appropriate for T2DM dataset as the prediction in non-stationary models of data was difficult. Moreover, predicting rare cases in clinical data with an unbalanced distribution of a target complication is challenging, where common statistical methods such as standard regression is not appropriate. This is because it only models average score over the different structures throughout the time series. Another method is re-sampling, which can be applied on the learning data and trigger its distribution based on the bias in the data [31]. In the next section, the time series re-balanced data is analysed by DBNs learning models.

[Figure 2 about here.]

[Figure 3 about here.]

Model generation and structure

Data mining and analysis were performed using MATLAB, Bayes Net toolbox [32], and "Graphviz" for visualisation. For learning the structure of the model, we used the K2 [33] and REVerse Engineering ALgorithm (REVEAL) [34] to create and temporal links (Inter) shown in Figure 2 and non-temporal links (Intra) shown in Figure 3. The networks with temporal associations

were inferred from T2DM historical patients time series data whilst represented in two DBNs (t and t-1) under the Markov properties assumption. In the discrete-space/discrete-time DBNs, two-time steps are considered to show the relationship between risk factors. For instance, Figure 2 shows the first complication at time t-1 that affects states of all other comorbidities and risk factors at t.

Dynamic Bayesian Networks

We designed a DBN to model the joint distribution of the domain representing probabilistic relationships between comorbidities and risk factors. DBNs were used to compute the probabilities of the presence of comorbidities over time, given a set of risk factors. DBNs were trained on the balanced T2DM data and tested on their power to predict a complication at the next time point, before the latent variables were explored. In the DBN structure (see Figure 2), 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.

IC*LS methodology

To learn the structure within DBNs and correlation among the latent variable and T2DM risk factors, a combination of the IC* algorithm and the Link Strength methods is used, which is called the IC*LS methodology.

Induction Causation (IC*) and Latent Variable Structure

The IC* algorithm is a constraint-based method which calculates several conditional independence tests. It returns a partially Directed Acyclic Graph (DAG) to characterise the entire Markov equivalence class [35]. The IC* algorithm is similar to the PC algorithm, except that it can detect the presence of latent variables. It learns a latent variable structure associated with a set of observed variables.

Understanding Latent Variables

The causal discovery of BNs is a critical research area, which depends on looking through the space of models for those which can best clarify a pattern of probabilistic conditions in the data [36]. The causal discovery indicates dependencies that are generated by structures with unmeasured factors, i.e., latent variables. The latent variable structure is a projection in which every latent variable is either a root node or a link to observed variables. The probability of a high state of any learned hidden variables is inferred using a standard Bayesian network inference and the Expectation-Maximization (EM) algorithm [37]. The Latent variable discovery in causal structures has been introduced in [38]. Latent variable models have a long tradition in causal discovery. Factor analysis and related methods can be used to position latent variables and measure their hypothetical effects. However, many do not provide clear means of deciding whether or not latent variables are present in the first place. One advantage of a latent variable is that they can better encode the actual dependencies and independencies in the data. For example, Figure 3 demonstrates 13 observed variables in the left-hand side DAG without a latent variable, and from the second DAG in the left side to the first in the right side, IC*LS learned one, two, three and four latent variables from T2DM data, respectively. As Friedman points out, a latent variable as a leaf/child or as root with only one child could be marginalised without affecting the distribution over the remaining variables. Thus, there would be a latent variable that mediates only one parent and one child. The IC* relies on statistical significance tests to decide whether an arc exists between two variables and on its orientation. In addition, a default error

rate (α = 0.05) is used to find the correlation of T2DM risk factors using IC* algorithm.[3]

Link Strength Metric

The Link Strength (LS) measures [39] is a metric to calculate the overall strength of the dependent links. It focuses on the most powerful dependencies between T2DM risk factors and enables us to observe the specific impact of each discovered edge in a DBN. The percentage points of uncertainty reduction in a variable are utilised by knowing the state of another variable if the states of all other parent variables are known. True Average Link Strength (LSTA) calculates LS based on the average over the parent states using their actual joint probability. If there was a link in the IC* adjacent matrix with LSTA greater or equal to some threshold (here 20 percent), a link in the final structure is retained; otherwise, it is 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[4].

Results

This section assessed the effectiveness of the bootstrap re-balancing method and the latent variable discovery approach in T2DM dataset. In Figure 3, in the left-hand side there is a DAG with no latent variable. From the second left to the first DAG in right-hand side of Figure 3 there are four steps to adds hidden variables, which are learned in the enhanced stepwise IC*LS approach. Conditional dependency for the hidden variable observed in the first, second, third and fourth steps, which are scored based on the Link Strength metrics. The selected T2DM nodes (features and predictors) are labelled and ordered from 1 to 13 which are corresponded to complications and risk factors including: HbA1c, retinopathy, neuropathy, nephropathy, Liver disease, hypertension, BMI, creatinine, cholesterol, HDL, DBP, SBP and smoking (as shown previously in Tables 1-2).

The proposed structure has been evaluated by performing the sensitivity analysis on the cohort based on two different perspectives: a “Visit-based” and a “Patient-based” validation test. The results were documented for the following comparative structures:

- UNB-K2-REVEAL: the original data (unbalanced) was trained in the K2 algorithm for Intra links and the REVEAL algorithm for Inter links with the unbalanced data (which is not reliable due to the imbalance issue explained earlier).
- B-K2-REVEAL: a latent variable and a fully learned structure from the K2 algorithm for Intra links and the REVEAL algorithm for Inter links with the balanced data using the TS Bootstrapping approach (shown in the left-hand side of Figure 2).
- NO-latent: the network is fully learned from the re-balanced data by using PC algorithm with no latent variable for Intra links shown as the first DAG in the left-hand side of Figure 3. The dynamic structure for Inter links

[3] More explanation of the Latent Structures is reported in the supplementary Material (Latent Structure).
[4] A detailed description of the LS metric is reported in the Supplementary Material.

is Fully Auto-Regressive; each node is connected to the corresponding node in the next time slice.

- IC*: the structure is obtained by using the IC* algorithm from the balanced data for Intra links seen in Figure 2 in [23] (three hidden variables were learned) and Fully Auto-Regressive structure for the Inter links.
- IC*LS: a combination of the IC* and LS filtering method is used to discover the structure for Intra links shown in the third step of the enhance stepwise IC*LS seen in the second right DAG with three hidden variables in Figure 3 and Fully Auto-Regressive structure for the Inter links.

Visit-based Validation

In Table 3, the patient time series (corresponding to the first four visits of each T2DM patient) were assessed to obtain the classification accuracy in predicting retinopathy, liver disease and hypertension. The time series were analysed considering the Area Under Receiver Operator Characteristic Curves (AUCs). The overall results showed that the proposed TS Bootstrapping method provided more accurate prediction compared to the unbalanced model (UNB-K2-REVEAL versus B-K2-REVEAL in Table 3).

In Table 3, the findings by using the IC* and IC*LS approaches were compared to a NO-latent method. Along with this improvement, the use of the latent methods enhanced AUC values significantly. In contrast, without an LS filter classification results were dropped considerably while the IC*LS approach was compared to the IC* approach. This improvement was potentially achieved because the LS measure filtered out the less robust links, thus avoided overfitting.

[Table 3 about here.]

To eliminate the class imbalance in the predictors, we used re-sampling to bootstrap rare records. Furthermore, we sought a significant improvement of accuracy in prognosis of different complications. The bootstrapped helped us to quantify and deal with bias about the different patient comorbidities. In this way, our approaches to balance the data are different from previous researches.

Patient-based Validation

Table 3 illustrated that whether the complications were predicted correctly or not depending on the characteristics of time-series data (Visit-based analysis). Here, Table 4 represented the performance measures (sensitivity and specificity) obtained for the proportion of T2DM patients who correctly tested positive/negative. Patient-based sensitivity results showed that how many patients that actually having a complication were identified correctly with that complication (comparison between predicted class value for a patient's complication and actual class value of the complication). Patient-based specificity represented that how many patients without a specific complication were tested negative with the complication correctly. Alternatively, Visit-based sensitivity outcomes showed that how many time points across all patients with having a specific complication were also tested positive correctly (comparison between previous and current values for the given visits). Visit-based specificity results revealed that how many visits (time points) which are equal to zero were scored correctly as a negative (zero). By switching the methodology from

B-K2-REVEAL to IC*LS in predicting liver disease showed a massive enhancement, first in sensitivity patient-based assessment, from 71% to 90%; second, in visit-based specificity experiments from 83% to 97% (as can be seen in Table 4-B-K2-REVEAL , IC*LS). Sensitivity was measured for retinopathy prediction by switching method from the standard methods (B-K2-REVEAL) to IC*LS increased sharply from 73% to 90% and 69% to 86% for Visit-based and patient-based, respectively. Despite this, for retinopathy the total number of patients, which was predicted correctly without the disease (specificity or true negative rate) remained almost constant or improved slightly from 87% to 88%. According to these results, it seemed to be evident that better prediction performance had been generally achieved by using the IC*LS method. A detailed explanation of the confusion matrix results is reported in the supplementary material.

[Table 4 about here.]

[Figure 4 about here.]

Confidence Interval Results

In this section the experimental findings and their significance were tested statistically by using the confidence interval. Figure 4 illustrated the influence of the latent variable on the bootstrapped data in predicting liver disease. The 95% confidence interval result demonstrated with high confidence that the IC*LS methodology resulted in a highly significant improvement in the classification accuracy, sensitivity and precision compared to the K2 and REVEAL algorithm as well as no latent variable approaches. The detailed findings are reported in Supplementary Material (Confidence Interval Results).

[Figure 5 about here.]

Latent Variable as Evidence

In this section, we look at the influence of the latent variable as evidence to predict the complications. As mentioned earlier, the aim of this research is to explore the impact of the targeted latent variable on prediction of the T2DM complications. Here, we validate the results to uncover influential factors in the diagnosis with regard to a set of diagnosis targets given the evidence. This target was possible to achieve due to the nature of the Bayesian inference, where any T2DM comorbidity or risk factor could be queried using a joint probability distribution. The bar chart in Figure 5-a shows how the probability of retinopathy being in its high value (the diagnosis point) changes by setting the evidence on the comorbidity-

being diagnosed with retinopathy is 0.05 (no evidence set). According to the DBN related risk factor. For instance, the marginal probability of one or more patient's retinopathy from 0.05 to 0.29. Thus, we guarantee that the latent variable has a model, setting evidence to the latent variable increases the marginal probability of significant impact on the target comorbidity posterior probability. However, setting evidence on the latent variable changes the posterior probability of liver disease from 0.99 to 0.97 and hypertension from 0.95 to 0.88. This showed the latent variable has a negative impact on developing liver disease and hypertension while it has a much stronger positive effect on retinopathy.

Latent Variable Validation Pattern

Figures 6-8 illustrated a case study to investigate how the latent variables have been interacted with other risk factors for predicting a complication in an individual patient. The early time prediction probabilities were represented in X-axis. In contrast, the targeted patient's visits were shown in the Y-axis. The predicted likelihood of liver disease was established in Figure 6-d seemed to be to very similar to its observed probability shown in Figure 6-a, which indicated the complication occurrence slightly earlier than the prediction. The IC*LS latent approach, in Figure 6-c for liver disease, revealed a trigger around the clinician observation time, whereas the latent K2 process in Figure 6-b remained steady. A less significant predicted probability was also captured in Figure 7-d. This illustrated a fluctuation just before retinopathy has been monitored in Figure 7-a. Similarly, a trigger happened in two latent approaches in Figures 7-b-c. It revealed that the latent models had been appeared to be predicting the switches in most patient cases. However, with the small sample size, caution must be applied, as the findings might not be applicable and there have been a few cases where the model could not predict a complication earlier than the clinicians. As a result, the expected findings for predicting hypertension might differ from the conclusions presented here, as it was compared in Figure 8-d comparing to Figure 8-a. It could be argued that the prediction results might be caused because of differences between complications. For example, hypertension has been reported as an easily detected macrovascular disease. In contrast, retinopathy as a chronic microvascular has been known very challenging to be caught at the earlier stage of the disease progression.

[Figure 6 about here.]
[Figure 7 about here.]
[Figure 8 about here.]

Discussion

According to the clinical evidence in Diabetes literature, the experimental results obtained in this study showed how the targeted use of latent variables improves prediction accuracy, specificity, and sensitivity over standard approaches. It also aided the understanding of relationships between these latent variables and disease complications/risk factors. Looking at how the different structures were performed within a DBN for predicting the appearance of complications, Figure 5 revealed that there could be a general trend to improvement in accuracy as more hidden variables have been added. Surprisingly, in Figure 5-b, a slight change was found in liver disease values, whilst the latent variable had the highest value. There was a significant negative correlation between the latent variable and hypertension, which was shown in Figure 5-c. As a result of including this latent variable, there was a steep rise in the prediction accuracy of hypertension from 65% to 99% (in the IC*LS approach in Table 3 compared to NO-latent). Similarly, a positive correlation was found between the latent variable and retinopathy in Figure 5-a. It was apparent from Table 3 that retinopathy prediction was enhanced considerably from 94% (NO-latent) to 99% (IC*LS) by adding the latent variable. Together these findings have provided important insights into the latent variable effects, which helped to reduce the uncertainty in the prediction process by identifying the relationship between

T2DM complications and risk factors. The AUC results obtained in Table 3-UNB-K2-REVEAL predicted hypertension accurately 60% of times comparing to 35% for retinopathy while data was imbalanced. This revealed the degree of improvement in the prediction performance from 35% to 51% for retinopathy and 38% to 51% for liver disease whilst 60% to 51% for hypertension. The reason behind this could be argued that hypertension has been known as a macrovascular complication while retinopathy reported as a typical microvascular complication. Furthermore, hypertension appeared to be the easiest complication to be detected by clinicians due to the routine measurement of blood pressure. Alternatively, retinopathy and liver disease required either ophthalmology consultation or ultrasonography of liver. Finally, if we look at the structure of the links with respect to the hidden variables on their Markov blanket, it can be seen in Figure 3 that there was a strong relationship between *Hidden Variable 2* at the third step of the enhance IC*LS and T2DM key risk factors (e.g., HbA1c, BMI, liver disease, and smoking).

In the first step of the stepwise IC*LS shown in the second left DAG of Figure 3, the initial hidden variable (*Hidden variable 1*) is weakly linked to a small number of clinical factors, notably HbA1c, Liver disease and creatinine. However, as subsequent hidden variables are added at the second step, this structure changes. The second hidden variable (*Hidden variable 2*) is linked stronger to HbA1c by 35.9 as subsequent hidden variables are added at the second step, this structure changes. and also, is connected to more risk factors including *Hidden variable 1* (seen as the third left DAG in Figure 3).

In the third step as seen as the second right DAG in Figure 3, *Hidden variable* 3 is closely connected to HbA1c by 63.2, and there is a 0 scored link between *Hidden variable 3*and *Hidden variable* **2.** This is while *Hidden variable* 2 is strongly connected to *Hidden variable* 1**,** which is scored 44.4**.** Thus, *Hidden variable* 3 (which is closely connected to HbA1c) seems to be irrelevant and independent of *Hidden Variable 2* (that is linked to nephropathy, liver disease and hypertension). At this step, there is a strong relationship between *Hidden variable 2*, retinopathy, liver disease, DBP, SBP and smoking. Having considered these hidden variable results obtained from Figure 3, Diabetes literature (see evidence in [9]), and advice of clinician expert in diabetes, it was suggested that the presence of HbA1c was associated with an increased incidence of nephropathy, while HbA1c emerged as an independent risk factor for developing retinopathy. Additionally, nephropathy and liver disease were independently associated with an increased incidence of hypertension in T2DM patients (see evidence in [40])[5].

The overall approach in this paper is abstracted in Figure 9. In the left-hand side of Figure 9, first the patient's history (including the disease risk factors and complications) was learned and trained in a DBN model (in the middle). The obtained DAG was learned at each step of the stepwise IC*LS approach representing the links from a latent variable to other clinical risk factors. Then the inferred latent variable probabilities were employed to predict a target complication earlier than the actual occurrence time (in the right-hand side). This figure also revealed that

[5]The proposed methodology of this research, in other works [26], are combined with pattern mining approaches to validate the target hidden variables and enhance the understanding of the sub-types of the disease based upon the developing disease complications.

the first latent variable (at visit $t-1$) was closely linked to a small number of clinical factors, while the second latent variable (at visit t) was connected to a larger number of risk factors.

[Figure 9 about here.]

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 a DBN model because of the transparent way of modelling data as well as the flexibility in incorporating latent variables. We incorporated the IC* algorithm and a Mutual Information based scoring metric to identify the strength of relationships between the latent variable and clinical risk factors.

This paper contributed in several ways to our understanding of how the latent variable provides a basis for a better prediction of the T2DM complications. Our results showed that our re-balancing approach by the use of TS bootstrapping method for an unequal number of time series visits demonstrated an improvement in the prediction performance. Additionally, the most highlighted contribution of this paper gained insight by interpreting the latent states while the association among the disease complications are taken into consideration. This led to a better understanding of risk factors and patient-specific interventions. A natural progression of this work involves extending the latent DBN models with more latent variables to capture a greater variety of factors to characterise critical changes. Our proposed approach will be useful for stratifying patients according to their probability of developing complications and clinician advice. For example, there is room for further progress in determining the optimal number of latent variables using Partial Least Squares (PLS). In addition, we will seek more advice from clinicians in interpreting hidden factors and their correlation toward other T2DM risk factors and complications as well as the disease prediction process. We also intend to look at more geographical and clinical factors, such as family history, pollution factors, and glucose levels.

Abbreviations

SD: Standard Deviation
H: Hidden/Latent Variable
BMI: Body mass index
HBA: Glycated Hemoglobin/H2A1c
CRT: Creatinine
COL: Cholesterol
HDL: High-Density Lipoprotein
DBP: Diastolic Blood Pressure
SBP: Systolic Blood Pressure

SMK: Smoking habit
R: Retinopathy
N: Neuropathy
P: Nephropathy
L: Liver disease
T2DM: Type 2 Diabetes Mellitus
DAG: Directed Acyclic Graph
DBN: Dynamic Bayesian Network
PLS: Partial Least Squares
TS: Time Series
IC: Induction Causation for observed variables
IC*: Induction Causation for observed variables and latent variables
LS: Link Strength
LSTA: True Average Link Strength
AUC: Operator Characteristic Curve
EM: Expectation-Maximization
REVEAL: Reverse Engineering Algorithm
TS Bootstrapping: Time Series Bootstrapping approach
MCMC: Markov Chain Monte Carlo

Declaration

The data that support the findings of this study are available from MOSAIC project but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of MOSAIC project for data collected from ICS Maugeri hospital in Pavia - Italy.

Funding

No funding was obtained for conducting this study. However, the data leading to these results had previously received from MOSAIC project funded by the European Commission under the 7 Framework Program, Theme ICT–2011.5.2 Virtual Physiological Human and grant agreement number 600914.

Consent for Publication

Not applicable.

Ethics approval and consent to participate

All participants of focus groups consented to study participation. The studies described in the evaluation activities were approved by the biomedical research ethics committee from the Ethics Committee at Istituti Clinico Scientifici Maugeri.

Authors contributions

LY has designed the study and organised the conduction of all the experiments, as well as contributed to the overall background, methodology and discussion of the manuscript. LY has executed and reported results of all the experiments and contributed to the methodology section. LC and LS have contributed and revised the clinical and scientific information related with T2DM. LY and AT provided contributions related with the latent variables and time series bootstrapping approach. MAL, LS, and AT have contributed to the substantial revision of the manuscript. All authors read and approved the final manuscript.

Competing interests

LS and AT are members of the editorial board for BMC Medical Informatics and Decision Making. All the authors declare that they have no competing interests.

Acknowledgements
Not Applicable.

References

1. Mathers, C.D., Loncar, D.: Projections of global mortality and burden of disease from 2002 to 2030. PLoS medicine 3(11), 442 (2006)
2. Long, A.N., Dagogo-Jack, S.: Comorbidities of diabetes and hypertension: mechanisms and approach to target organ protection. The journal of clinical hypertension 13(4), 244–251 (2011)
3. Raman, R., Gupta, A., Krishna, S., Kulothungan, V., Sharma, T.: 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 26(2), 123–128 (2012)
4. Van Gerven, M.A., Taal, B.G., Lucas, P.J.: Dynamic bayesian networks as prognostic models for clinical patient management. Journal of biomedical informatics 41(4), 515–529 (2008)
5. Van der Heijden, M., Velikova, M., Lucas, P.J.: Learning bayesian networks for clinical time series analysis. Journal of biomedical informatics 48, 94–105 (2014)
6. Mueller, E., Maxion-Bergemann, S., Gultyaev, D., Walzer, S., Freemantle, N., Mathieu, C., Bolinder, B., Gerber, R., Kvasz, M., Bergemann, R.: Development and validation of the economic assessment of glycemic control and long-term effects of diabetes (eagle) model. Diabetes technology and therapeutics 8(2), 219–236 (2006)
7. Dagliati, A., Marinoni, A., Cerra, C., Decata, P., Chiovato, L., Gamba, P., Bellazzi, R.: 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 10(1), 19–26 (2016)
8. Dagliati, Arianna, Alberto Malovini, Pasquale Decata, Giulia Cogni, Marsida Teliti, Lucia Sacchi, Carlo Cerra, Luca Chiovato, and Riccardo Bellazzi. "Hierarchical Bayesian Logistic Regression to forecast metabolic control in type 2 DM patients." In AMIA Annual Symposium Proceedings, vol. 2016, p. 470. American Medical Informatics Association, 2016.
9. Teliti, M., Cogni, G., Sacchi, L., Dagliati, A., Marini, S., Tibollo, V., De Cata, P., Bellazzi, R., Chiovato, L.: Risk factors for the development of micro-vascular complications of type 2 diabetes in a single-centre cohort of patients. Diabetes and Vascular Disease Research 15(5), 424–432 (2018)
10. Guo, Y., Bai, G., Hu, Y.: Using bayes network for prediction of type-2 diabetes. In: 2012 International Conference for Internet Technology and Secured Transactions, pp. 471–472 (2012). IEEE
11. Marini, S., Trifoglio, E., Barbarini, N., Sambo, F., Di Camillo, B., Malovini, A., Manfrini, M., Cobelli, C., Bellazzi, R.: A dynamic bayesian network model for longterm simulation of clinical complications in type 1 diabetes. Journal of biomedical informatics 57, 369–376 (2015)
12. Murphy, K P, Russell, S : Dynamic bayesian networks: representation, inference and learning (2002)
13. Lloyd, A., Sawyer, W., Hopkinson, P.: Impact of long-term complications on quality of life in patients with type 2 diabetes not using insulin. Value in Health 4(5), 392–400 (2001)
14. Friedman, N., Murphy, K., Russell, S.: Learning the structure of dynamic probabilistic networks. In: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, pp. 139–147 (1998). Morgan Kaufmann Publishers Inc.
15. Pearl, J : Probabilistic reasoning in intelligent systems 1988 San Mateo, CA: Kaufmann 23, 33–34
16. Elidan, G., Lotner, N., Friedman, N., Koller, D.: Discovering hidden variables: A structure-based approach. In: Advances in Neural Information Processing Systems, pp. 479–485 (2001)
17. Martin, J., VanLehn, K.: Discrete factor analysis: Learning hidden variables in bayesian networks. Technical report, Technical report, Department of Computer Science, University of Pittsburgh (1995)
18. Tucker, A , Liu, X , Garway-Heath, D : Spatial operators for evolving dynamic bayesian networks from spatio-temporal data. In: Genetic and Evolutionary Computation—GECCO 2003, pp. 205–205 (2003). Springer
19. Trifonova, N., Kenny, A., Maxwell, D., Duplisea, D., Fernandes, J., Tucker, A.: Spatio-temporal bayesian network models with latent variables for revealing trophic dynamics and functional networks in fisheries ecology. Ecological Informatics 30, 142–158 (2015)
20. Robinson, Joshua W., Alexander J. Hartemink, and Zoubin Ghahramani. "Learning Non-Stationary Dynamic Bayesian Networks." Journal of Machine Learning Research 11, no. 12 (2010).
21. Grzegorczyk, M., Husmeier, D.: Non-stationary continuous dynamic bayesian networks. In: Advances in Neural Information Processing Systems, pp. 682–690 (2009)
22. Talih, M , Hengartner, N : Structural learning with time-varying components: tracking the cross-section of financial time series. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67(3), 321–341 (2005)
23. 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. In: 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS), pp. 106–111 (2018). IEEE
24. 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. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1040–1044 (2018). IEEE
25. 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. In: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), pp. 198–203 (2019). IEEE
26. 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)
27. Yousefi, L., Saachi, L., Bellazzi, R., Chiovato, L., Tucker, A.: Predicting comorbidities using resampling and dynamic bayesian networks with latent variables. In: Computer-Based Medical Systems (CBMS), 2017 IEEE 30th International Symposium On, pp. 205–206 (2017). IEEE
28. Turner, R., Millns, H., Neil, H., Stratton, I., Manley, S., Matthews, D., Holman, R.: Risk factors for coronaryartery disease in non-insulin dependent diabetes mellitus: United Kingdom prospective diabetes study (ukpds: 23). Bmj 316(7134), 823–828 (1998)
29. Japkowicz, N., Stephen, S.: The class imbalance problem: A systematic study. Intelligent data analysis 6(5), 429–449 (2002) 30. SIMAR, L.: An invitation to the bootstrap: Panacea for statistical inference? Institut de Statistique, Universite Catholique de Louvain, Louvain (2008)
30. Moniz, N., Branco, P., Torgo, L.: Resampling strategies for imbalanced time series. In: Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference On, pp. 282–291 (2016). IEEE
31. Murphy, K., ET AL.: The bayes net toolbox for matlab. Computing science and statistics 33(2), 1024–1034 (2001)
32. Cooper, G.F., Herskovits, E.: A bayesian method for the induction of probabilistic networks from data. Machine learning 9(4), 309–347 (1992)
33. Liang, S., Fuhrman, S., Somogyi, R.: Reveal, a general reverse engineering algorithm for inference of genetic network architectures (1998)
34. Spirtes P, Glymour CN, Scheines R, Heckerman D. Causation, prediction, and search. MIT press; 2000.

Zhang, X., Korb, K.B., Nicholson, A.E., Mascaro, S.: Latent variable discovery using dependency patterns. arXiv preprint arXiv:1607.06617 (2016)

Figures

Figure 1 IC*LS Diagram: The overall strategy of the proposed predictive model.

Figure 2 Latent DBN Structure: Latent DBN Structure: time series structures using the REVEAL (left-hand side) and Fully Auto-Regressive dynamic links (right hand side). The H, C, and O illustrate Hidden node, Complication, and Observed node, respectively.

Figure 3 IC*LS graph: The latent structure of a Bayesian model is learned from T2DM features. The arrows represent Intra links discovered using IC* and the score on the links shows the LS metric for the corresponding connection.

Figure 4 Bootstrap Confidence Interval in Predicting Liver Disease: Bootstrap Confidence Interval represents accuracy, sensitivity, specificity, and precision to predict liver disease.

Figure 5 Latent Variable as Evidence: Prediction probabilities for retinopathy, liver disease and hypertension using the Latent variable as the evidence.

Figure 6 Latent Variable Behaviour for predicting the onset of Liver disease: A latent prediction pattern of liver disease over time (a patient follow-ups).

Figure 7 Latent Variable Behaviour for predicting the onset of retinopathy: Latent variable prediction pattern of retinopathy over time (a patient follow-ups).

Figure 8 Latent Variable Behaviour for predicting the onset of hypertension: Latent variable prediction pattern of hypertension over time (a patient follow-ups).

Figure 9 A DBN Latent Model: From the left-hand side, in the middle, and the right-hand side demonstrate the patient's history, the inferred latent variable probabilities, the prediction, respectively.

List of Tables

Table 1 The description of T2DM Target Complication, Clinical Node Control Values, and Discretised States.

Node ID	Target Complication	Diagnosis Outcome	Clinical Risk Class
2	Retinopathy (RET)	{Negative,Positive}	{low,high}
3	Neuropathy (NEU)	{Negative,Positive}	{low,high}
4	Nephropathy (NEP)	{Negative,Positive}	{low,high}
5	Liver Disease (LIV)	{Negative,Positive}	{low,high}
6	Hypertension (HYP)	{Negative,Positive}	{low,high}

Table 2 The description of the T2DM Clinical Features, Risk Factors, Control Values, and the Discretised States.

Node ID	T2DM Risk Factors	Control Value (Mean±SD)	Discretised Value
1	HbA1c (HBA)	6.6 ± 1.2 (%)	{low,medium,high}
7	Body Mass Index (BMI)	26.4 ± 2.4 (kg/m2)	{low,medium,high}
8	Creatinine (CRT)	0.9 ± 0.2 (mg/dL)	{low,medium,high}
9	Cholesterol (COL)	0.9 ± 0.2 (mg/dL)	{low,medium,high}
10	High-Density Lipoprotein (HDL)	1.1 ± 0.3 (mmol/l)	{low,medium,high}
11	Diastolic Blood Pressure (DBP)	91 ± 12 (mmHg)	{low,medium,high}
12	Systolic Blood Pressure (SBP)	148 ± 19(mmHg)	{low,medium,high}
13	Smoking Habit (SMK)	{0,1,2}	{low,medium,high}

Table 3 Visit-based performance assessment on the prediction results.

Performance Measure	UNB-K2-REVEAL	B-K2-REVEAL	NO-latent	IC*	IC*LS
AUC of Retinopathy	35	50	94	89	99
AUC of Liver Disease	38	51	71	92	99
AUC of Hypertension	60	51	65	83	99

Table 4 Comparison of Patient-based and Visit-based prediction performance.

Complication Method	Retinopathy B-K2-REVEAL	Retinopathy IC*LS	Liver disease B-K2-REVEAL	Liver disease IC*LS
Sensitivity (Patient-based)	69	86	71	90
Sensitivity (Visit-based)	75	92	94	95
Specificity (Patient-based)	89	90	98	99
Specificity (Visit-based)	89	90	85	99

Figures

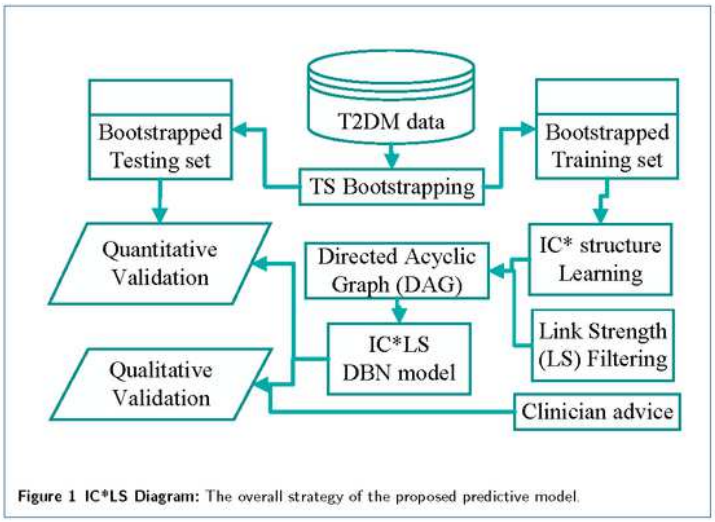

Figure 1 IC*LS Diagram: The overall strategy of the proposed predictive model.

Figure 1

IC*LS Diagram: The overall strategy of the proposed predictive model.

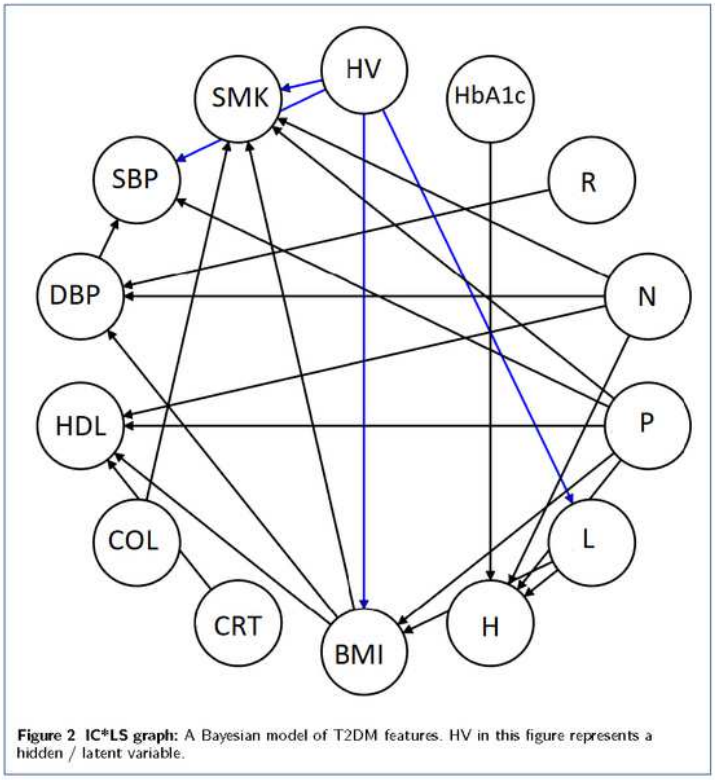

Figure 2 IC*LS graph: A Bayesian model of T2DM features. HV in this figure represents a hidden / latent variable.

Figure 2

Latent DBN Structure: Latent DBN Structure: time series structures using the REVEAL (left-hand side) and Fully Auto-Regressive dynamic links (right hand side). The H, C, and O illustrate Hidden node, Complication, and Observed node, respectively.

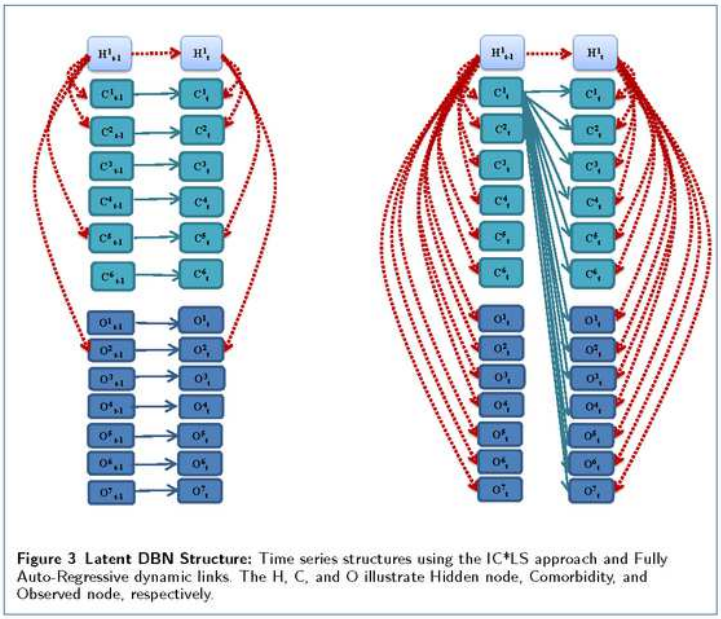

Figure 3 Latent DBN Structure: Time series structures using the IC*LS approach and Fully Auto-Regressive dynamic links. The H, C, and O illustrate Hidden node, Comorbidity, and Observed node, respectively.

Figure 3

IC*LS graph: The latent structure of a Bayesian model is learned from T2DM features. The arrows represent Intra links discovered using IC* and the score on the links shows the LS metric for the corresponding connection.

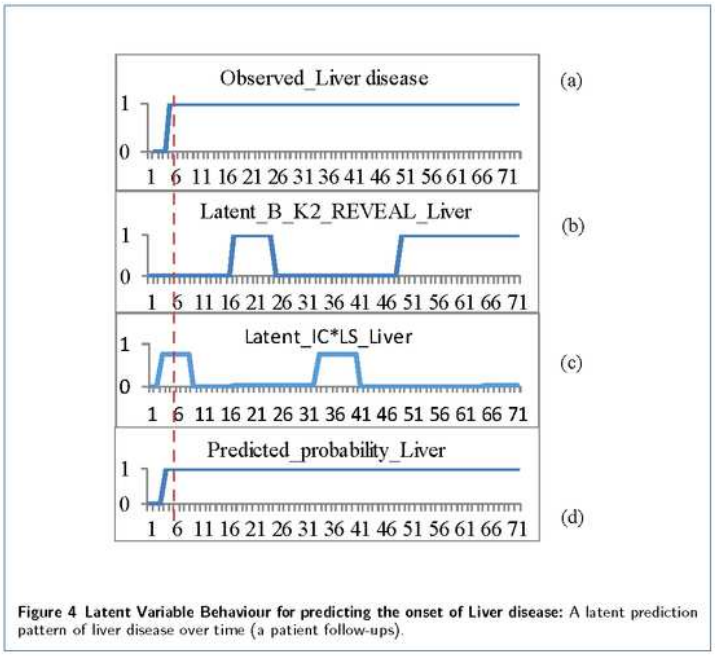

Figure 4 Latent Variable Behaviour for predicting the onset of Liver disease: A latent prediction pattern of liver disease over time (a patient follow-ups).

Figure 4

Bootstrap Confidence Interval in Predicting Liver Disease: Bootstrap Confidence Interval represents accuracy, sensitivity, specificity, and precision to predict liver disease.

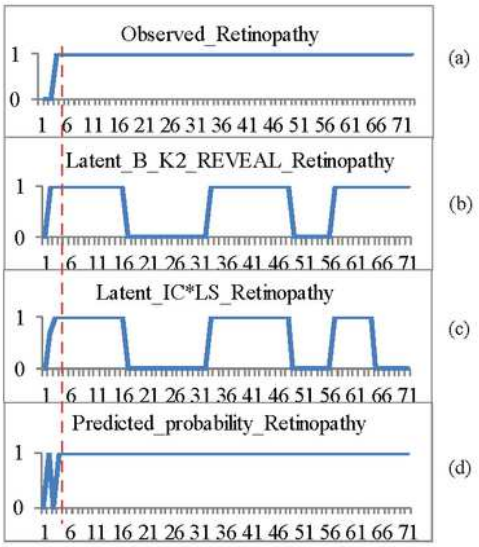

Figure 5 Latent Variable Behaviour for predicting the onset of retinopathy: Latent variable prediction pattern of retinopathy over time (a patient follow-ups).

Figure 5

Latent Variable as Evidence: Prediction probabilities for retinopathy, liver disease and hypertension using the Latent variable as the evidence.

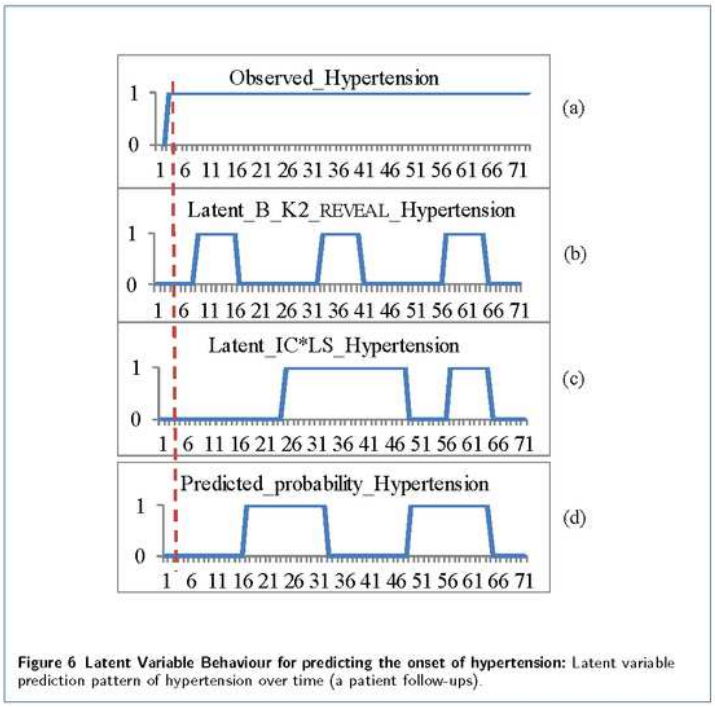

Figure 6 Latent Variable Behaviour for predicting the onset of hypertension: Latent variable prediction pattern of hypertension over time (a patient follow-ups).

Figure 6

Latent Variable Behaviour for predicting the onset of Liver disease: A latent prediction pattern of liver disease over time (a patient follow-ups).

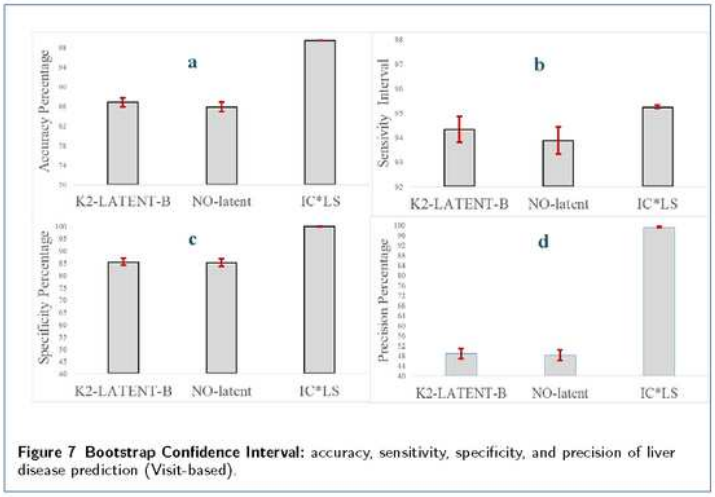

Figure 7 Bootstrap Confidence Interval: accuracy, sensitivity, specificity, and precision of liver disease prediction (Visit-based).

Figure 7

Latent Variable Behaviour for predicting the onset of retinopathy: Latent variable prediction pattern of retinopathy over time (a patient follow-ups).

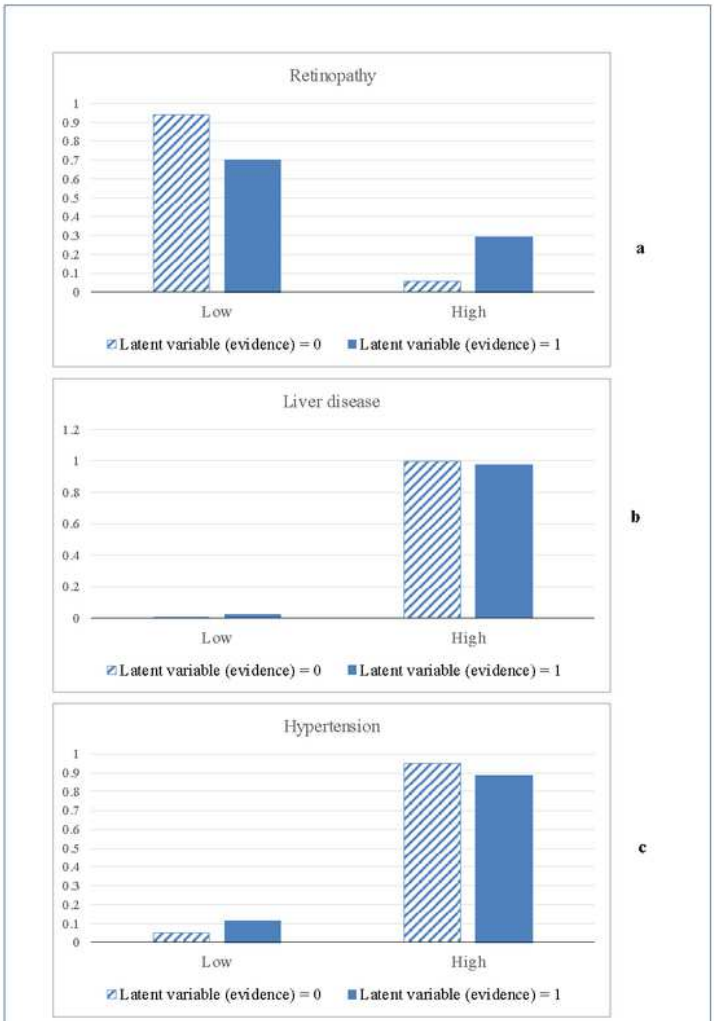

Figure 8 Prediction probabilities: The obtained posteriors for retinopathy, liver disease, and hypertension using Latent variable as the evidence.

Figure 8

Latent Variable Behaviour for predicting the onset of hypertension: Latent variable prediction pattern of hypertension over time (a patient follow-ups).

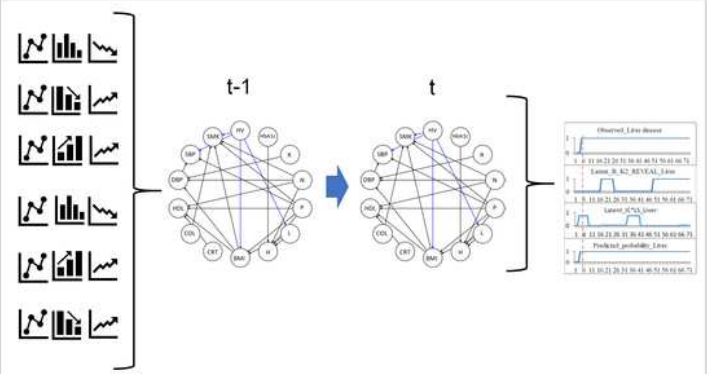

Figure 9 A DBN Latent Model: From the left hand side, in the middle, and the right hand side demonstrate the patients history, the inferred latent variable probabilities, the prediction, respectively.

Figure 9

A DBN Latent Model: From the left-hand side, in the middle, and the right-hand side demonstrate the patient's history, the inferred latent variable probabilities, the prediction, respectively.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryBMCjournal.pdf