2018 IEEE 31st International Symposium on Computer-Based Medical Systems

Predicting Disease Complications Using a Stepwise Hidden Variable Approach for Learning Dynamic Bayesian Networks

Leila Yousefi
Dept. of Computer Science
Brunel University London
London, United Kingdom
Leila.Yousefi@brunel.ac.uk

Allan Tucker
Dept. of Computer Science
Brunel University London
London, United Kingdom
Allan.Tucker@brunel.ac.uk

Mashael Al-luhaybi
Dept. of Computer Science
Brunel University London
London, United Kingdom
Mashael.Al-luhaybi@brunel.ac.uk

Lucia Saachi
Dept. of Computer Science and Systems
University of Pavia
Pavia, Italy
Lucia.Sacchi@unipv.it

Riccardo Bellazzi
Dept. of Computer Science and Systems
University of Pavia
Pavia, Italy
Riccardo.Bellazzi@unipv.it

Luca Chiovato
Unit of Endocrinology
Fondazione Maugeri
Pavia, Italy
Luca.Chiovato@unipv.it

***Abstract*—Predicting Diabetes Type 2 Mellitus (T2DM) complications such as retinopathy and liver disease is still a challenge despite being a growing public health concern worldwide. This is due to the complex interactions between complications and other features, as well as between the different complications, themselves. What is more, there are likely to be many unmeasured effects that impact the disease progression of different patients. Probabilistic graphical models such as Dynamic Bayesian Networks (DBNs) have demonstrated much promise in the modeling of disease progression and they can naturally incorporate hidden (latent) variables using the EM algorithm. Unlike deep learning approaches that attempt to model complex interactions in data by using a large number of hidden variables, we adopt a different approach. We are interested in models that not only capture unmeasured effects but are also transparent in how they model data so that knowledge about disease processes can be extracted and trust in the model can be maintained by clinicians. As a result, we have developed a step-wise hidden variable structure learning process that incrementally adds hidden variables based on the IC* algorithm. To the best of our knowledge, this is the first study for classifying disease complication using a step-wise learning methodology for identifying hidden and T2DM features with a DBN structure from clinical data. Our extensive set of experiments show that the proposed method improves classification accuracy, identifying the correct number of hidden variables, and targeting their precise location within the network structure.**

***Keywords*-Disease Prediction; Dynamic Bayesian Networks; Hidden Variables; Longitudinal data**

I. Introduction

Diabetes affects about 10% of the population. As such diabetes and its associated complications are in fact the eight leading cause of death worldwide. Diabetes Mellitus Type 2 (T2DM) [1] or insulin resistant type is the most common form of diabetes, which includes 90% of all cases. The World Health Organization reported that in the next 11 years,

there will be about 550 million people suffering from this disease [1]. T2DM is associated with severe long-term complications and large health maintenance costs to providers. Previous research has revealed that 86% of patients with T2DM have other associated complications such as kidney damage (nephropathy), nerve damage (neuropathy) and eye disease (retinopathy). Up to 75% of adults with diabetes also have hypertension, and patients with hypertension often show evidence of insulin resistance [2]. Similar to Diabetic type 1 patients, although genetic factors impact on developing T2DM, it is believed ignorance of developing complications harms patients' life. What is more, T2DM patients develop a different profile of complications and features, which changes over time per follow-up visit. One of the most important factors in the high number of dependencies among T2DM features and complications is the appearance of unmeasured risk factors. Surprisingly, the effect of understanding unmeasured variables, which play an important role in disease prediction, does not seems that closely examined. In this work, a predictive framework of T2DM comorbidities was implemented using latent variables to predict complication onset. Furthermore, the relationship between the discovered hidden (latent) variables and observed T2DM complications was explored using Dynamic Bayesian Networks (DBNs) [3]. Unlike deep learning approaches that attempt to model complex interactions in data by using a huge number of hidden variables, we are interested in models that incrementally add individual latent variables that not only capture unmeasured effects but are also transparent in how they model data so that knowledge about disease processes can be extracted and trust in the model can be maintained by clinicians.

[1]The T2DM dataset in this study was provided at the IRCCS Istituti Clinici Scientifici (ICS) Maugeri of Pavia, Italy.

2372-9198/18/$31.00 ©2018 IEEE
DOI 10.1109/CBMS.2018.00026

A. Related research

1) Dynamic Bayesian Networks and T2DM prediction: Dynamic Bayesian Networks (DBNs) are popular for modeling uncertain noisy time series clinical data [3]. DBNs can represent probabilistic relationships between comorbidities and symptoms. Previous work on learning DBNs has inferred network structures and parameters from clinical data sets in the presence of missing data and hidden (latent) variables. There is a growing body of literature that recognizes the importance of analyzing the structure and parameters of DBNs in the prognosis or diagnosis of clinical tasks, for example, for finding the relationship among different brain regions in several disorders[4]. In [5], Khalilia attempted to predict disease in the National Inpatient Sample (NIS) data by training random forest classifiers using a repeated subsampling method. There is also some research into using these methods for investigating the diagnosis of T2DM disorders [6] [7] [8].

2) Dealing with bias in data: There are different learning techniques that deal with imbalanced data, such as oversampling, undersampling, boosting, bagging, bootstrapping and repeated random sub-sampling [9]. In this study, since T2DM dataset is highly imbalanced based on the common complications, we employ an oversampling approach in time series patient modeling for increasing the size of the rare class. [10] presented a Dynamic Bayesian Network method with a latent variable after balancing data using a bootstrap approach for modeling fisheries data.

3) causal structure learning and Latent variable discovery: Factor Learning (FL) in [11] is one method for learning a probabilistic model from data. FL contrasts with most other BN learning methods in that it learns a factor structure. Factor structure indicates the joint probability distribution among discrete observed variables. Factor structures contribute an explanation across a small amount of variables. Although they are suitable for polynomial time inference, caused reducing accuracy and precision. In [11] author provided a factor structure for learning methods that efficiently understood hidden variables. However, its method failed to use prior belief in factor structure and therefore, could not rely on the final structure. [12] explored the possibility of using trees of hidden variables that render all observable variables independent. However, trees of hidden variables are non-optimal when there are independencies between observable variables. The casual discovery of BNs is a critical research territory, which depends on looking through the space of casual models for those which can best clarify a pattern of probabilistic conditions appeared in the data [13]. The casual discovery indicates dependencies that are generated by casual structures with unmeasured factors, i.e., latent variables. Hidden variable discovery in casual structure has been introduced in [14]. [15] used Bayesian techniques to find the most probable structure and can use this technique to add hidden variables. In principle, exact Bayesian methods for hidden variables could identify the most probable structures of factors given the data and suitable priors. However, with a large number of variables, exact methods are prohibitively expensive. Our previous work [16] exploited DBNs on rebalanced T2DM data while incorporating the Inductive Causation (IC*) algorithm and a Mutual Information based scoring metric to identify the strength of relationships between the latent variable and clinical factors.

II. METHOD

In this section, we discuss how the explicit consideration of latent processes contributes to an improved modeling of T2DM features. Hence, we develop a novel algorithm that iteratively adds hidden variables to a DBN structure so that unmeasured effects can be captured.

A. Data

A total of 1000 patients newly diagnosed as having type II diabetes, aged 25 to 65 years inclusive, were recruited between 2009 and 2013 from clinical follow-ups of diabetes patients at the IRCCS Istituti Clinici Scientifici (ICS) Maugeri of Pavia, Italy. The data is part of the MOSAIC project funded by the European Commission under the 7th Framework Program, Theme ICT–2011.5.2 Virtual Physiological Human (600914). It should be noted that after balancing the data set, a number of pre-diagnosed patients by the specific complication were removed, so the number of patients reduced to less than 400. In addition, in different balanced dataset based on the specific complication, a different population of patients were used in testing and training sets. Risk factors found to be influencing T2DM [17] [18] included physical examination and laboratory data (Body Mass Index (BMI), Systolic Blood Pressure (SBP), High-Density Lipoprotein (HDL), Triglycerides (TRIG), Glycated Hemoglobin (HbA1c), Diastolic Blood Pressure (DBP), total Cholesterol (Cholestrol), Smoking habit and Creatinine). Data mining and analysis were performed using MATLAB and Bayes Net toolbox [19] and for visualisation we used Graphviz.

B. Learning from Imbalanced Data using Pair-sampling

In this paper, Pair-sampling was exploited to effectively address unbalanced time series medical data. This method divided the dataset into Positive and negative instances, from which the train and test data sets are generated. Firstly, we removed pre-diagnosed patient samples (patients were diagnosed at the first follow-up visit) from the original T2DM dataset. The remaining samples were divided into two separate cases: P positive data samples and N negative data samples. Since the data was highly imbalanced $(N >> P)$, we randomly oversampled $(N - P)$ patients from positive cases to have the same amount of P and N.

In addition, positive cases was increased to the amount of their subtraction from the negative cases. Half of P (Positive patient cases) was fitted into test set while the remaining is allocated to a train set. Similarly, we split out N instances for training and testing sets. Now data was partitioned into two patient samples with each partition containing an equal number of patient samples from the target complication, while ensuring that one sub-sample contains only positive cases and another one negative cases of patients. We defined a pair-visit of $[P_1, P_2]$ for each patient, in which P_1 indicated that target comorbidity was not diagnosed (denoted by a '0') and P_2 represented the following visit of P_1, which could be either not diagnosed ('0') or diagnosed (denoted by '1'). For all undiagnosed patients for a target comorbidity we selected two random consecutive time points of data (pair-visit = [0 0]), which indicates there was no diagnosed comorbidity in the two consecutive visits of a patient. For all patients diagnosed with a specific comorbidity, we select two consecutive time points so that they represent the switch from no diagnosed comorbidity to diagnosed comorbidity (pair-visit = [0 1]). At the end, we shuffled the positive and negative cases to randomly spread out around within the data set.

C. Step-Wise IC*

1) Bayesian Networks and The Inductive Causation (IC) Algorithm:* Bayesian Networks (BNs) are graphical models that represent the joint probability distribution over a set of variables (e.g. risk factors). These variables are represented as nodes within the Bayesian graphical structure and directed connections between these nodes capture independencies between them. We assume a positive diagnosis when greater than a threshold of 0.1. The IC* algorithm provides a procedure to determine which causal connections among nodes in a network can be inferred from empirical observations even in the presence of latent variables. IC* is a constraint-based method which applies conditional independence analyses to infer causal structures and learns a partially-oriented directed acyclic graph with latent variables [20]. This algorithm can be used to analyze effective connectivity among T2DM features. We propose a new methodology which combines the basic principles of the IC* algorithm to obtain a directed acyclic graph in addition to a dynamic process that is inferred using the REVEAL algorithm [21]. A step-wise approach used to incrementally add latent variables which are now described.

2) Stepwise discovery strategy: We employed an incremental strategy using IC* algorithm, which we called Stepwise IC*. A diagram describing the process of Pair-sampling and the stepwise procedure is presented in Figure 1. In this method, IC* is applied to a dataset. The probability of a high state of any learned latent variables at this current step is then inferred using the expectation maximization (EM) algorithm. The inferred probabilities of the hidden variable are treated as observations which means that we can then treat the hidden variable as an observed variable in the subsequent step. In this next step, IC* is applied again to see if the new observed variable uncovers any new hidden variables. This is repeated until no other hidden variables are discovered. For example, a balanced dataset based on retinopathy is provided using our Pair-sampling. The model was trained using the structure obtained from the balanced dataset at the first step and a hidden variable was found. At the next step, there are 14 observed variables including 13 different T2DM features plus one additional hidden variable probability that was inferred in the previous step. Then we learn the structure from this new 14 variable dataset and discover a new hidden variable. Moreover, we retrieved the prediction probability of the hidden variable and to be used to create another observed feature in the third version of the dataset. Later we used this obtained dataset to train and test next step of adding more hidden variables in the third step. Therefore, data was trained with a new hidden variables pointing to neuropathy, HbA1c, liver disease, smoking, and BMI (see Figure 2). If IC* could find a new hidden variable and there was an improvement then we would continue adding one more hidden variable in the next step.

III. Experimental Results

To test our approach we selected three T2DM comorbidities with high prevalence in elderly patients. We performed a set of experiments to compare the step-wise approach in each step of latent variable discovery and without using latent variables. Our results show that we can predict complications with an improved accuracy using the Pair-sampling whilst dealing with highly imbalanced T2DM data. In addition, with limiting the number of latent variables, we ease process of understanding them.

In Figure 2 we can see the discovered relations among DBN nodes for the retinopathy structure at the three steps of the stepwise approach. It can be seen how the addition of the first hidden node influences the hidden node discovered at the second step (by the explaining away effect via neuropathy). The third hidden node is then added based upon being linked in part to hidden variable 2. Each of the components of T2DM exposure (HbA1c, hidden at the first step, and hidden at the second step) were significantly associated with risk of retinopathy progression. Similar results involving varying interacting hidden variables have been observed when applied to other complications. The influence of a new learned latent variable in each step of the Stepwise algorithm has been demonstrated by a bar chart in Figure 3. In figure 3, we monitor the effects of the targeted latent variable by changing the evidence of the observed latent variable in different states (0 and 1). Bayesian network inference was used to query target complication to capture the probability distribution. As can be seen in Figure 3, the prediction probability of retinopathy dropped while evidence

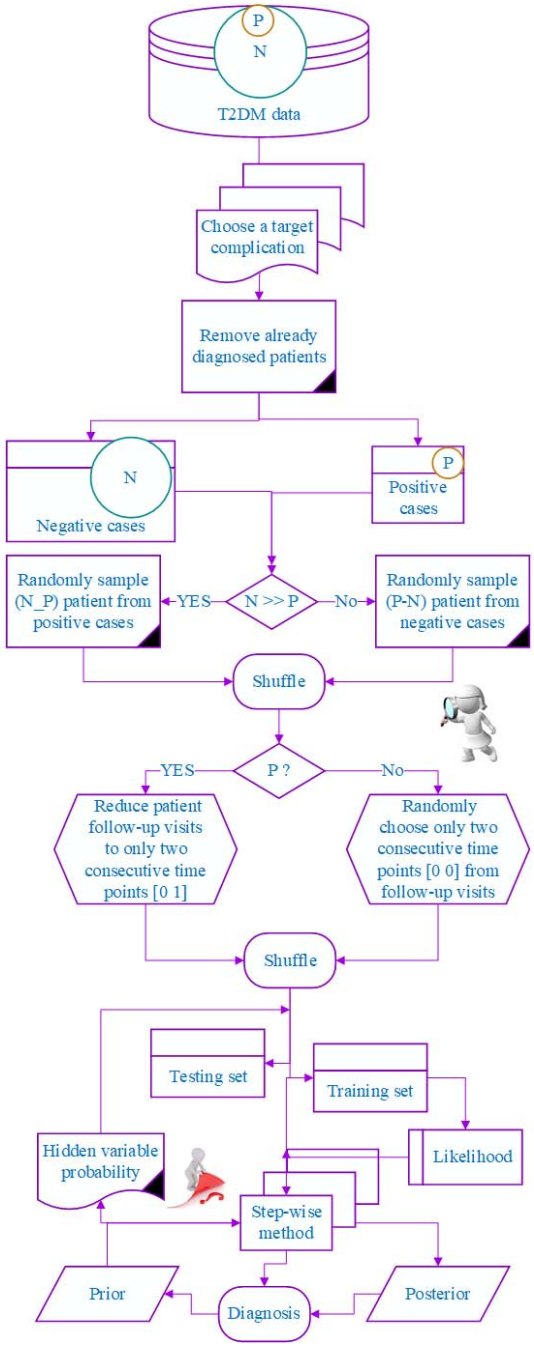

Figure 1. Diagram of Pair-sampling and the Stepwise approach.

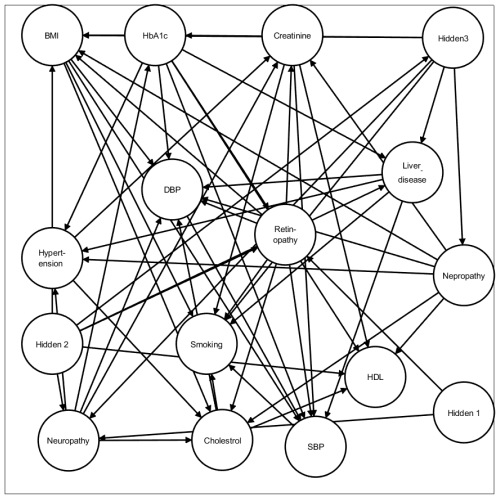

Figure 2. Graph of static relationships among T2DM risk factors by applying the third step of the Stepwise approach.

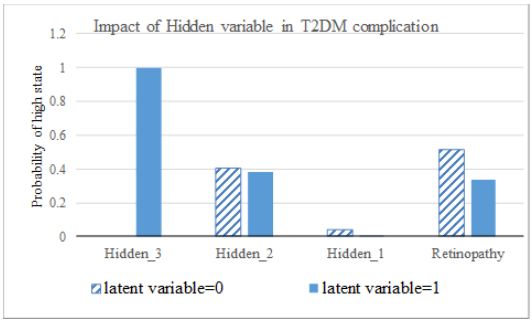

Figure 3. Changes in target complication (retinopathy) in respond to different values of evidence (latent variable at the third step of the Stepwise approach.

(latent variable in the first time slot) was switched to one. The hidden variable at the third step of the prediction (hidden 3) is generally seen as a factor strongly related to retinopathy while the probability of retinopathy being diagnosed is dropped from 0.5 to 0.3 by setting evidence from 0 to 1. Thus, we can interpret this as a discovered hidden variable in the higher step of the Stepwise approach is a significant contributory factor to the development of retinopathy. Moreover, it is also understood that hidden variable at the second step of learning plays an important role in the diagnosis of retinopathy. We now explore the use of our step-wise approach to learning latent variables with respect to disease prediction. Figure 4 emphasizes the power of our step-wise approach with a generally improving accuracy as a number of hidden variables are added. The accuracy plots indicate that the performance of the predictor for retinopathy from the no-hidden step to the first step of our approach is less significant than precision (comparing Figure 4a and 4b). The sensitivity of retinopathy prediction has improved

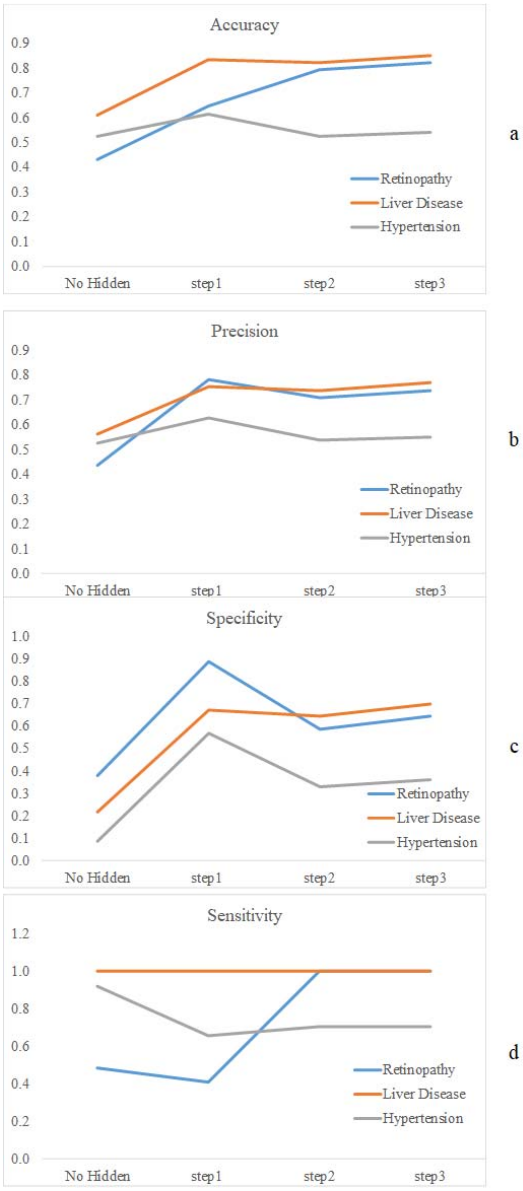

Figure 4. Assessing performance based on the confusion matrix results for three complications (retinopathy, liver disease and hypertension).

sharply from 0.4 to 1. However, sensitivity for prediction of liver disease and hypertension remained constant after the first step of approach. Despite the sharp rise in performance measures by adding a hidden variable at the first step of the approach, for the rest of trend (learning more hidden variable) has a slight increase from the first step (step1) to higher steps. Additionally, Figures 4a-4c illustrate that there is a sharp rise in prediction results (accuracy, precision and specificity) by exploiting a latent variable in the first step of the Stepwise method. Thereafter, from Figure 4a it is obvious that prediction accuracy has been improved from the first step (step1) to the third step (step3) especially in

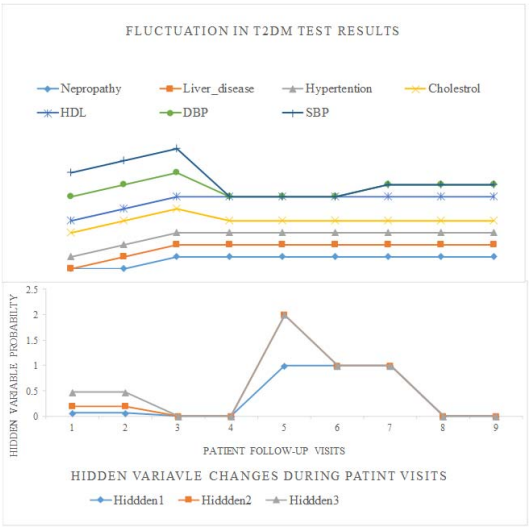

Figure 5. Impacts of understanding Hidden variable patterns at each step of Stepwise method for a patient with nine follow-up visits based upon the DBN inferred from the balanced data for retinopathy.

retinopathy and liver disease. Precision is easier to interpret, e.g. in Figure 4b, a precision of 0.77 in the third step (step3) of learning hidden variable can immediately be understood as 77% correct predictions among the positive predictions of liver disease. Overall, Figures 5 explain how the changes in hidden variable in three steps of our learning method reflect fluctuation in the observed variables at different points per patient visits. As can be seen in Figure 5, latent variable learned at the third step of Stepwise approach is on its peak and higher than the other steps and at the same time earlier than observed variables (SBP and DBP) rise points, which emphasized the importance of hidden variable discovery in early time disease prediction.

IV. Conclusion

Predicting disease complications at the early stage of a longitudinal study has been known as a critical issue which has high practical benefits in clinical applications. For many clinical problems in patients, the underlying structure of risk factors (hidden factors) plays an important role in medical interventions. The Relationship of T2DM risk factors affects the risk of Development and Progression of complications in follow-up visits. A systematic understanding of how latent variables contribute to T2DM complications is still lacking and in this paper, we have made a start by developing an intuitive step-wise method to learn these latent effects based upon the IC* algorithm. More specifically, our approach effectively integrates Bayesian methods with latent variables by adapting the prior probability of the event occurrence for future time points. We show how the hidden variables influence the T2DM risk factors. Our results reveal that the proposed method is more accurate than using one of hidden

variable step or no hidden variables at all.

A. Future Works

One limitation of our approach could be the stopping rule to the step-wise approach and in some cases, it seems that accuracy starts to drop after the final hidden variable is added. This may represent overfitting. Classification accuracy could be monitored and used as a stopping condition (i.e. if it drops significantly). Although the IC* algorithm only learns static structure (we use another process to learn temporal links) there is potential to update the IC* algorithm to learn temporal links. A future approach will use mutual information metrics [22] to filter some of the hidden variable relationships where IC* results in uncertainty in choosing either a latent variable or a direct link between two nodes. Finally, we will explore deep learning and Bayesian Neural Networks for disease prediction where huge numbers of hidden factors are used to extract features from clinical data prior to prediction.

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] 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.

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

[4] D. Chicharro and S. Panzeri, "Algorithms of causal inference for the analysis of effective connectivity among brain regions," *Frontiers in neuroinformatics*, vol. 8, p. 64, 2014.

[5] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," *BMC medical informatics and decision making*, vol. 11, no. 1, p. 51, 2011.

[6] 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.

[7] 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.

[8] 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.

[9] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent data analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[10] 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.

[11] J. Martin and K. VanLehn, "Discrete factor analysis: Learning hidden variables in bayesian networks," Technical report, Department of Computer Science, University of Pittsburgh, Tech. Rep., 1995.

[12] J. Pearl, "Probabilistic reasoning in intelligent systems. 1988," *San Mateo, CA: Kaufmann*, vol. 23, pp. 33–34.

[13] X. Zhang, K. B. Korb, A. E. Nicholson, and S. Mascaro, "Latent variable discovery using dependency patterns," *arXiv preprint arXiv:1607.06617*, 2016.

[14] C. SPEARMAN, "" general intelligence," objectively determined and measured," *American Journal of Psychology*, vol. 15, pp. 201–293, 1904.

[15] 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.

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

[17] 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.

[18] 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.

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

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

[21] K. Murphy, S. Mian *et al.*, "Modelling gene expression data using dynamic bayesian networks," Technical report, Computer Science Division, University of California, Berkeley, CA, Tech. Rep., 1999.

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