doi.org/10.26398/asaproc.0135

The combination of expert knowledge and Bayesian networks to identify the relevant factors in diabetic kidney disease (DKD)

Debora Slanzi[ab], Irene Poli[b], Roger D. Jones[b], Gert Mayer[c]

[a] Dept of Management, Venice School of Management, Ca' Foscari University of Venice, Italy
[b] European Centre for Living Technology (ECLT), Ca' Foscari University of Venice, Italy
[c] Dept of Internal Medicine IV (Nephrology and Hypertension), Medical University Innsbruck, Austria

1. Introduction

Type 2 diabetes mellitus (T2DM) is a chronic condition characterized by elevated blood glucose levels due to insulin resistance or reduced insulin secretion. The number of adults with T2DM in Europe is expected to increase from 52.8 million in 2011 to 69 million by 2045. About 30-40% of these individuals develop diabetic kidney disease (DKD), a severe complication that decreases both the quality and duration of life and imposes a significant burden on healthcare systems. DKD is the leading cause of end-stage renal disease in developed countries (Perco et al., 2019). Initially, kidney disease in T2DM was thought to mimic that in type 1 diabetes, driven primarily by genetic predisposition and metabolic control. However, it is now understood to be more complex and multifactorial, influenced by comorbidities such as hypertension and disruptions in various biological pathways. This complexity results in significant variability in disease progression and therapy response among individuals (Galicia-Garcia et al. 2020). Understanding these mechanisms is essential for improving clinical care and developing targeted treatments (Slanzi et al., 2024, Abebe et al., 2024).

Identifying the most relevant variables that influence DKD is crucial for a more targeted approach to treatment and management (Jones et al., 2023). By focusing on these key factors, we can reduce uncertainty and improve the precision of interventions, ultimately leading to better patient outcomes. This targeted approach minimizes the need for broad-spectrum treatments, which can be less effective and carry a higher risk of side effects. Additionally, understanding the specific variables at play can help to develop personalized therapies that address each patient's unique needs, thereby enhancing treatment efficacy and slowing DKD progression.

Integrating data-derived insights with expert knowledge can significantly enhance the understanding and management of DKD. Data provides empirical evidence that uncovers patterns and relationships among variables, while expert knowledge offers contextual understanding and clinical experience. Combining these sources allows for a more comprehensive analysis, adding valuable context to raw data and improving the selection of the most impactful variables. This integration not only refines the focus of research and treatment but also ensures that the chosen variables are clinically relevant and actionable, ultimately leading to more effective and personalized healthcare strategies for patients with DKD.

The aim of this work is to develop an approach based on Bayesian networks for selecting the relevant variables that influence diabetic kidney disease (DKD). A Bayesian network is a probabilistic graphical model that represents a set of variables and their conditional dependencies via a directed acyclic graph. By utilizing Bayesian networks, we aim to uncover the complex relationships among various factors involved in DKD. This method allows for the integration of data and expert knowledge, providing a systematic way to identify and analyse the most impactful variables. Ultimately, this approach can lead to more precise and effective interventions for managing DKD, improving patient outcomes by focusing on the most critical aspects of the disease. The paper is structured as follows: in Sect. 2 we introduce the study conducted to achieve the data used in the analyses and the statistical approach developed to address the proposed objectives, in particular the steps of the procedure used to variable selection and how to include prior knowledge available from the literature and experts to produce more informative models;

then in Sect. 3 we present the main results achieved in analysing DKD patients highlighting the main findings of the research.

2. Materials and methods

The Data. The data utilized in this study are sourced from the PROVALID study ("PROspective cohort study in patients with type 2 diabetes mellitus for VALIDation of biomarkers"), a prospective observational study that enrolled over 4,000 T2DM patients across five European countries with normal, mild, or moderately reduced kidney function. These patients were monitored for at least four years, with annual collection of clinical data, laboratory values, and medication information. For a more detailed description of the study and the available data, refer to (Eder et al., 2018 and 2019). The outcome of the study is determined by the estimated glomerular filtration rate (eGFR) value, a marker of renal excretory capacity, recorded at each visit. The disease trajectories measured by changes in eGFR vary significantly among PROVALID participants even under stable therapy (Kerschbaum et al., 2020). This variability is likely due to differences in drug adherence, environmental factors, and heterogeneity in pathophysiology.

The number of initial variables considered in the study exceeds 120. The aim of the work is to identify which of these variables are most relevant for predicting changes in eGFR, specifically the difference in eGFR between two consecutive visits, measured as $\Delta eGFR = eGFR_{ti} - eGFR_{ti-1}$. Additionally, the study seeks to elucidate how these variables are interconnected to better understand the underlying mechanisms of the disease. By pinpointing the key variables and mapping their interrelations, the study would like to provide a comprehensive explanation of the factors driving changes in eGFR and to shed light on the complex interactions contributing to the progression of diabetic kidney disease (DKD).

After preprocessing the data to remove incomplete cases and to adjust skewed distribution through log transformation when appropriate, we include n=537 observations in our analysis. We point out that we consider datapoints without considering the longitudinal aspect because, in this initial phase, we are interested in identifying the variables responsible for a change in eGFR, independently of the timing of the visit.

Bayesian networks. To identify the network of relationships among the many factors that may be responsible for diabetic kidney disease (DKD), we use a class of probabilistic graphical models known as Bayesian networks (BNs) (Koller and Friedman, 2009; Scutari and Denise, 2014; Arora et al., 2019). BNs are directed acyclic graphs where nodes represent variables and edges denote probabilistic dependencies between these variables. Each node in the network is associated with a probability distribution that quantifies the likelihood of different values of the variable, given the values of its parent nodes.

Specifically, a BN for a set of random variables $\mathbf{X}=\{X_1,...,X_p\}$ is identified by:

– a *network structure* G, a directed acyclic graph (DAG) where nodes represent the variables **X** of the system and the directed arcs between nodes represent the probability dependences between them,

– a *set of parameters*, representing conditional probability distributions $P(X_i|Pa(X_i))$ associated to each variable X_i, i=1,...,p, where $Pa(X_i)$ are the variables that correspond to the parents of X_i in the DAG (i.e. the nodes with an arc pointing towards X_i that indicates a directed dependency with it).

The global distribution of the variables **X** is decomposed into the local distributions of the individual variables X_i as

$$P(\boldsymbol{X}) = \prod_{i=1}^{p} P(X_i|Pa(X_i))$$

The process of estimating a BN is called learning and typically involves two main steps: (1) the *structure learning* to identify the topological structure, i.e. which arcs are present in the graph and therefore which probabilistic relationships are supported by the data, and (2) the *parameter learning* to learn the conditional probability distributions that regulate the strength of the relationships.

There are many approaches in literature to estimate BNs from the data (Kitson et al., 2023): in this work we will focus on a Search & Score strategy which uses a score function in order to compare the structures of the network and then selects the structure which better fits the data. Specifically, we develop

structure learning by means of a hill-climbing search procedure and a BDe score (Koller and Friedman, 2009). BNs can model complex, multivariate relationships and are able to handle uncertainty by combining prior knowledge with observed data. They also facilitate inference and reasoning, allowing for the prediction of outcomes and the identification of causal relationships. Through these properties, BNs provide a robust framework for understanding the intricate interplay of factors contributing to DKD.

Variable selection. The Markov blanket (MB) is a crucial concept in BNs for variable selection. It represents the set of nodes that directly influence a specific node, consisting of its parents, children, and the parents of its children. This set encapsulates all the information needed to predict the target variable, effectively isolating it from the rest of the network. By focusing on the MB, one can identify the minimal and most relevant subset of variables necessary for prediction, thereby eliminating redundant or irrelevant data. This not only simplifies the model and enhances computational efficiency but also improves interpretability by highlighting the key variables that influence the outcome. Overall, using the Markov Blanket for variable selection ensures a theoretically optimal and practical approach to building efficient and comprehensible BNs (Wang et al., 2020).

Since experts in DKD are interested also in variables that indirectly impact the target of interest, it has been proposed to expand the concept of MB in a BN. This expansion includes not only the directly related variables but also those indirectly connected to the target. By doing so, the analysis can capture a broader range of influential factors, providing a more comprehensive understanding of the various elements that may affect the target outcome. This approach allows for a deeper exploration of the relationships within the network, potentially uncovering hidden or subtle influences that might otherwise be overlooked. We thus develop a variable selection procedure using the properties of BNs in several steps. Specifically, we proceed as follow:

1. BN estimation: A model averaging procedure is developed where arcs that exceed a certain threshold t are identified. This process involves performing bootstrap resampling, which means re-sampling the data k times using the bootstrap and performing structure learning separately on each of the resulting samples to collect k DAGs. The arc strength is then calculated by determining the frequency with which each arc appears in those k graphs. An "average" consensus DAG is derived by selecting those arcs that have a frequency above t. Following the literature, in this work we set the number of bootstrap replications to $k = 100$ and threshold to $t = 0.5$ (selecting only arcs with strength > 0.5) (Scutari and Nagarajan, 2013).
2. Markov blanket identification: the MB of ΔeGFR (target of the system) is selected. MB includes the target variable's parents, children, and the parents of its children, effectively encapsulating all variables that directly influence or are directly influenced by the target.
3. MB expansion: for all variables identified in the initial MB, their MB is also selected. This process iteratively expands the network to include all relevant variables and their direct and undirected relationships, ensuring a comprehensive selection of variables most pertinent to the target variable.

Additionally, expert knowledge can be integrated into the process by imposing certain known relationships in the pathophysiology of diabetic kidney disease (DKD). This ensures that domain-specific insights enhance the accuracy and relevance of the variable selection. The prior information was provided by study physicians in the form of 32 prior relationships (whitelisted arcs) derived from the pathophysiological theoretical framework as described in Slanzi et al. (2024).

3. Results

Starting with the whole set of over 120 variables, the selection process described above identify 18 variables as the most important for the study's target, ΔeGFR. The selected variables are listed in Table 1. Only two of these, CST3 and SCR, are part of the Markov blanket MB(ΔeGFR). The remaining variables constitute the MB expansion, indicating variables that appear to have undirected relationships with the target. These variables are important for eGFR and its variation because they are associated with various physiological and clinical aspects that can influence kidney function and health.

Table 1. The selected variables (in alphabetical order)

Covariate	*Description*	*Type*	*Descriptive statistics (Mean±sd / Percent%)*
ADIPOQ	Adiponectin, C1Q, And Collagen Domain Containing	Numerical	8490781± 3947052
ADMD	Age at DM2 diagnosis	Numerical	52.36 ±9.93
ADMET	Biguanides (metformin)	Categorical	True: 74% False: 26%
CST3	Cystatin C concentration in serum	Numerical	1.29±0.43
DLOOP	Loop diuretics	Categorical	True: 29% False: 71%
EGF	Epidermal growth factor concentration in urine normalized by UCREA	Numerical	8.42±0.77
EGFR	eGFR creatinine based	Numerical	64.36±18.13
FGF21	Fibroblast growth factor 21 concentration in urine normalized by UCREA	Numerical	104.62±74.39
GE	Gender	Categorical	Female: 46% Male: 54%
HB	Haemoglobin	Numerical	13.48±1.57
HDLCHOL	Serum cholesterol (HDL)	Numerical	48.04±13.60
LGALS3	Galectin 3 concentration in serum	Numerical	8824±2581.98
PHRDB	Personal history of renal disease at baseline	Categorical	Diabetic renal disease:13% Other renal disease: 12% False: 75%
SCR	Serum creatinine	Numerical	1.10±0.37
SPOT	Serum potassium	Numerical	4.54±0.49
TNFRSF1A	TNF Receptor Superfamily Member 1A concentration in serum	Numerical	2174±1020.38
UACR	Mean UACR	Numerical	2.49±1.81
UCREA	Urinary creatinine	Numerical	78.80±42.39
Target			
DEGFR	ΔeGFR= $eGFR_{ti}$- $eGFR_{ti-1}$	Numerical	-1.45±18.76

For instance, biomarkers like Adiponectin (ADIPOQ), Cystatin C (CST3), and TNF receptor superfamily member 1A (TNFRSF1A) are related to metabolic and inflammatory processes that affect renal function. Other variables like age at DM2 diagnosis (ADMD), personal history of renal disease (PHRDB), and medication usage (e.g., ADMET for metformin and DLOOP for loop diuretics) provide context on patient history and treatment, which can significantly impact kidney health. Additionally, factors such as serum creatinine (SCR), urinary creatinine (UCREA), and mean UACR provide direct measures of kidney function, while variables like haemoglobin (HB) and serum potassium (SPOT) can indicate related systemic health conditions. These variables collectively help in understanding and predicting changes in eGFR, providing a comprehensive overview of factors influencing kidney function.

To understand the relevance of the selected variables for the target, we estimate a Bayesian network using p_1=18+1=19 variables (the selected variables plus the target, ΔeGFR) to evaluate the strength of their relationships. Additionally, since it is possible to incorporate expert prior information, we also estimate a Bayesian network by introducing prior knowledge from the literature and practical experience in the field. This is accomplished using whitelisted arcs, which represent well-known dependencies that should be enforced in the graph, based on 32 prior relationships derived from the pathophysiological theoretical framework constructed on 17 specific variables (Slanzi et al. 2024). As many of these prior relationships involved variables not included in the selected set, it was necessary to expand the list of variables to encompass all relevant ones, resulting in a new set of p_2=30 variables (only 7 of these are in the set of 18 selected). Subsequently, we estimate the BNs as described in step 1 of the proposed procedure outlined in Section 2. In Table 2, we present several measures of graphical differences: the number of nodes (Num. nodes), the number of arcs (Num. arcs), the average Markov blanket size (Av. MB size), the average neighbourhood size (Av. neighb. size), and the Bayesian information criterion (BIC) to evaluate the model fit to the data.

Table 2. Measures of graphical differences. *Original* refers to the BN estimated using the initial complete set of variables, *Selected* refers to the BN learned using the set of selected variables only, *Selected + Prior* refers to the BN learned by using the set of selected variables and imposing expert knowledge.

	Num. nodes	Num.arcs	Av. MB size	Av. neighb. size	BIC	Num. MB(ΔeGFR)
Original	123	199	5.14	3.24	-137443.4	2
Selected	19	43	6.74	4.53	-35847.18	6
Selected+Prior	30	105	13.00	7.00	-55796.59	6

Additionally, we highlight the cardinality of the MB of the target variable (Num. MB(ΔeGFR)). From the results, we can see that the "selected" BN has fewer arcs, indicating that the data suggest relationships considered to be more robust. Adding prior relationships enriches the set of arcs, making the network denser. Notably, the number of variables forming the MB of the target is larger compared to the original network, highlighting that reducing the number of variables to those that are truly relevant can lead to more accurate results.

To evaluate the effectiveness of the proposed procedure in identifying an efficient set of variables that are useful for reducing the noise from redundant variables and deriving the pattern of DKD pathophysiology with clinically relevant and actionable variables, we calculate the predictive accuracy of the BNs in terms of the correlation between the observed and predicted values for the target variable. This predictive accuracy is assessed using 10-fold cross-validation. We compute the correlation between the observed and predicted values for ΔeGFR and this quantity is called predictive correlation. The procedure is repeated M=30 times to achieve the mean posterior predictive correlation for the three BNs as previously described. The results are presented in Figure 1. We find that changes in eGFR are more accurately predicted when using the reduced set of variables. Furthermore, using prior information improves the prediction accuracy, although not substantially, demonstrating how the expertise of clinicians can enhance the description of the system.

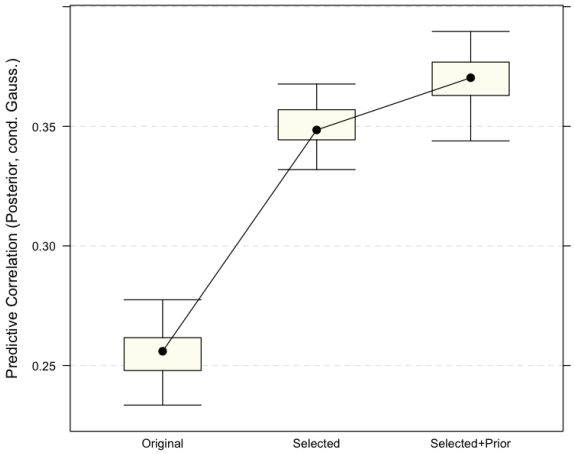


Figure 1. Predictive correlation of ΔeGFR. *Original* refers to the BN estimated using the initial complete set of variables, *Selected* refers to the BN learned using the set of selected variables only, *Selected + Prior* refers to the BN learned by using the set of selected variables and imposing expert knowledge.

By incorporating expert knowledge in the modelling phase, we demonstrate that the synergy between data and expert prior information is a great source of valuable assistance in gaining new insights into studying complex disorders and understanding DKD with respect to a predefined target of interest. In fact, our results indicate that more than half of the variables identified as relevant to the target of interest are not part of the a priori pathophysiology theoretical framework identified by clinical experts. This has provided new evidence regarding the pathophysiology of the disease, highlighting how statistical methods and data-driven technologies can support expert decisions with new knowledge. Future research will consider the longitudinal dimension of the data to assess the model's ability to capture temporal dynamics,

and comparisons with other methods for variable selection.

Acknowledgments

This work is supported by the project DC-ren, that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 848011.

References

Abebe, S., Poli, I., Jones, R.D., Slanzi, D. (2024) Learning optimal dynamic treatment regime from observational clinical data through reinforcement learning. *Machine Learning & Knowledge. Extraction*, **6**, pp. 1798–1817.

Arora, P., Boyne, D., Slater, J.J., Gupta, A., Brenner, D.R., Druzdzel, M.J. (2019) Bayesian networks for risk prediction using real-world data: A tool for precision medicine. *Value in Health,* **22**(4), pp. 439–445.

Eder, S., Leierer, J., Kerschbaum, J., Rosivall, L., Wiecek, A., de Zeeuw, D., Mark, P.B., Heinze, G., Rossing, P., Heerspink, H.L., Mayer, G. (2018) A prospective cohort study in patients with type 2 diabetes mellitus for validation of biomarkers (PROVALID) - Study design and baseline characteristics, *Kidney Blood Press Research*, **43**(1), pp. 181–190.

Galicia-Garcia, U., Benito-Vicente, A., Jebari, S., Larrea-Sebal, A., Siddiqi, H., Uribe, K.B., Ostolaza, H., Martin, C. (2020) Pathophysiology of type 2 diabetes mellitus, *International Journal of Molecular Sciences*, **21**(17), 6275.

Eder, S., Leierer, J., Kerschbaum, J., Rosivall, L., Wiecek, A., de Zeeuw, D., Mark, P.B., Heinze, G., Rossing, P., Heerspink, H.L., Mayer, G. (2019) Guidelines and clinical practice at the primary level of healthcare in patients with type 2 diabetes mellitus with and without kidney disease in five European countries, *Diabetes & Vascular Disease Research*, **16**(1), pp. 47–56.

Jones, R.D, Abebe, S., Distefano, V., Mayer, G., Poli, I., Silvestri, C., Slanzi, D. (2023) Candidate composite biomarker to inform drug treatments for diabetic kidney disease. *Frontieres in Medicine*, **10**, 1271407

Kerschbaum, J., Rudnicki, M., Dzien, A., Dzien-Bischinger, C., Winner, H., Heerspink, H.L., Rosivall, L., Wiecek, A., Mark, P.B., Eder, S., Denicolò, S., Mayer, G. (2020) Intra-individual variability of eGFR trajectories in early diabetic kidney disease and lack of performance of prognostic biomarkers, *Scientific Reports*, **10**, 19743.

Kitson, N.K., Constantinou, A.C., Guo, Z., Liu, Y., Chobtham, K.(2023) A survey of Bayesian network structure learning, *Artificial Intelligent Review*, **56**, pp. 8721-8814.

Koller, D., Friedman, N. (2009) *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Perco, P., Pena, M., Heerspink, H.J.L., Mayer, G. (2019) Multimarker panels in diabetic kidney disease: The way to improved clinical trial design and clinical practice?, *Kidney International Reports*, **4**(2), pp. 212–221.

Scutari, M., Denis, J.B. (2021) *Bayesian Networks with Examples in R*. Chapman & Hall.

Scutari, M., Auconi, P., Caldarelli, G., Franchi, L. (2017) Bayesian networks analysis of malocclusion data, *Scientific Report*, **7**(1), 15236.

Scutari, M., Nagarajan, R. (2013) On Identifying Significant Edges in Graphical Models of Molecular Networks. *Artificial Intelligence in Medicine*, **57**(3), pp. 207–217.

Slanzi, D., Silvestri, C., Poli, I., Mayer, G. (2024). Exploiting the potential of Bayesian networks in deriving new insight into diabetic kidney disease (DKD), in *Artificial Life and Evolutionary Computation. WIVACE 2023*, eds M. Villani, S. Cagnoni, R. Serra, R. *Communications in Computer and Information Science*, vol 1977. Springer, Cham.

Wang, H., Ling, Z., Yu, k., Wu, X. (2020) Towards efficient and effective discovery of Markov blankets for feature selection, *Information Sciences*, **509**, pp. 227-242.