TYPE Original Research
PUBLISHED 12 January 2023
DOI 10.3389/fpubh.2022.1035025

Identifying risk factors of developing type 2 diabetes from an adult population with initial prediabetes using a Bayesian network

Pilar Fuster-Parra[1,2], Aina M. Yañez[2,3,4]*, Arturo López-González[5,6], A. Aguiló[2,3] and Miquel Bennasar-Veny[2,3,7]

[1]Department of Mathematics and Computer Sciences, Balearic Islands University, Palma, Spain, [2]Institut d'Investigació Sanitària Illes Balears (IdISBa), Hospital Universitari Son Espases, Palma, Spain, [3]Department of Nursing and Physiotherapy, Balearic Islands University, Palma, Spain, [4]Research Group on Global Health and Human Development, Balearic Islands University, Palma, Spain, [5]Escuela Universitaria ADEMA, Palma, Spain, [6]Prevention of Occupational Risk in Health Services, Balearic Islands Health Service, Palma, Spain, [7]CIBER de Epidemiología y Salud Pública (CIBERESP), Instituto de Salud Carlos III (ISCIII), Madrid, Spain

OPEN ACCESS

EDITED BY
MinJae Lee,
University of Texas Southwestern Medical Center, United States

REVIEWED BY
Marija Jevtic,
University of Novi Sad, Serbia
Pedro Larrañaga,
Polytechnic University of Madrid, Spain
Baoping Cai,
China University of Petroleum, China

*CORRESPONDENCE
Aina M. Yañez
✉ aina.yanez@uib.es

SPECIALTY SECTION
This article was submitted to Life-Course Epidemiology and Social Inequalities in Health, a section of the journal Frontiers in Public Health

RECEIVED 02 September 2022
ACCEPTED 15 December 2022
PUBLISHED 12 January 2023

CITATION
Fuster-Parra P, Yañez AM, López-González A, Aguiló A and Bennasar-Veny M (2023) Identifying risk factors of developing type 2 diabetes from an adult population with initial prediabetes using a Bayesian network. *Front. Public Health* 10:1035025. doi: 10.3389/fpubh.2022.1035025

Background: It is known that people with prediabetes increase their risk of developing type 2 diabetes (T2D), which constitutes a global public health concern, and it is associated with other diseases such as cardiovascular disease.

Methods: This study aimed to determine those factors with high influence in the development of T2D once prediabetes has been diagnosed, through a Bayesian network (BN), which can help to prevent T2D. Furthermore, the set of features with the strongest influences on T2D can be determined through the *Markov blanket*. A BN model for T2D was built from a dataset composed of 12 relevant features of the T2D domain, determining the dependencies and conditional independencies from empirical data in a multivariate context. The structure and parameters were learned with the bnlearn package in R language introducing *prior* knowledge. The *Markov blanket* was considered to find those features (variables) which increase the risk of T2D.

Results: The BN model established the different relationships among features (variables). Through inference, a high estimated probability value of T2D was obtained when the body mass index (BMI) was instantiated to *obesity* value, the glycosylated hemoglobin (HbA1c) to more than 6 value, the fatty liver index (FLI) to more than 60 value, physical activity (PA) to *no* state, and age to 48–62 state. The features increasing T2D in specific states (warning factors) were ranked.

Conclusion: The feasibility of BNs in epidemiological studies is shown, in particular, when data from T2D risk factors are considered. BNs allow us to order the features which influence the most the development of T2D. The proposed BN model might be used as a general tool for prevention, that is, to improve the prognosis.

KEYWORDS

Bayesian networks, Markov blanket, prediabetes, type 2 diabetes, prevention, risk factors

Introduction

Type 2 diabetes (T2D) is recognized as a global serious health concern with a considerable impact on human life and health expenditures (1) whose prevalence has steadily increased and is now one of the main causes of morbidity and mortality in adults (2, 3). T2D is a preventable condition and is very likely to develop in people whose blood glucose levels are higher than normal but do not fulfill the criteria for a diagnosis of T2D (prediabetes) (4). Prediabetes also increases the risk of cardiovascular disease and mortality (5), similar to people with diabetes (6). The rate of progression to T2D for this population ranges between 5 and 10% annually (7, 8). According to the IDF Diabetes Atlas, a total of 6.0% of the worldwide population (463 million individuals aged 20–79 years) are estimated to have prediabetes (9). Acting on prediabetes could be a window of opportunity to prevent or delay T2D. Different risk factors for the progression from prediabetes to T2D have been described such as obesity (general or abdominal), a family history of diabetes, ethnicity, gestational diabetes, high systolic blood pressure (SBP), low level of high-density lipoprotein cholesterol (HDL-C), and tobacco smoking (4, 10, 11). More studies and data analysis frameworks are needed to evaluate the complex relationship among different risk factors on the progression to T2D. In this sense, a model for T2D based on Bayesian networks (BNs) is considered.

Models, such as BNs (12–14), capture the potential relationships among features (factors) like an expert understands them (15) by a directed acyclic graph (DAG) where the nodes have conditional probability tables. They constitute an established framework and an efficient reasoning tool for uncertainty management in artificial intelligence (AI), which has been used to discover the relationships between variables determining the direct and indirect dependencies (14, 16, 17). Models include graph theory together with probability theory in order to represent the relationships between variables (18), and they are especially selected because they provide probability estimates rather than predictions. Moreover, BNs offer a paradigm for interpretable AI, where high-stakes applications have increased, and therefore, the use of interpretable models is important (19). In this sense, they can be applied to help health practitioners by providing T2D characterization estimates as a probability network that can be continuously updated according to patient information given by practitioners.

A BN (12, 13) $\mathcal{B}$ encodes a joint probability distribution P over a vector of random variables (features) $\mathbf{X} = (X_1, \dots, X_n)$, and it consists of (20) (i) a DAG $\mathcal{G}$, which is composed of a set of variables (features), each variable has a finite set of mutually exclusive states, where these variables are the vertices (i.e., nodes) of $\mathcal{G}$, and a set of directed edges (i.e., arcs) between these variables (features) of vector $\mathbf{X}$, and (ii) a set of parameters θ such that $\mathcal{B} = (\mathcal{G}, \theta)$, where $\mathcal{G}$ denotes a structure represented by a DAG, θ is a set of local parameters according to the structure $\mathcal{G}$, the parameters are the conditional probability distributions for the values of each variable given different value combinations of their parent nodes. The joint probability distribution P encoded by the BN $\mathcal{B}$ factorizes as a product of several local conditional distributions which denotes the dependency/independency structure by a DAG:

$$P(X_1, \dots, X_n) = \prod_{i=1}^{n} P(X_i \mid pa(X_i^{\mathcal{G}})). \tag{1}$$

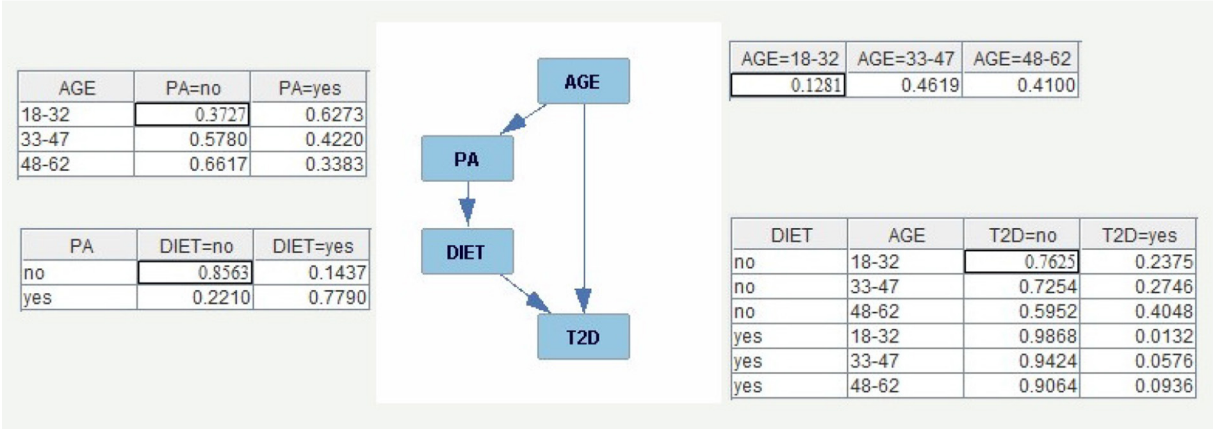
Equation (1) is also referred to as the *chain rule for BNs*, where $pa(X_i^{\mathcal{G}})$ denotes the parent nodes of X_i in DAG $\mathcal{G}$, which is the main reason for the formulation of a multivariate distribution by BNs.

An example of BN, including the structure and conditional probabilities tables, is presented in Figure 1. In this case, the chain rule is given by

$$\begin{aligned} P(AGE, PA, DIET, T2D) &= P(AGE) \cdot P(PA \mid AGE) \\ &\cdot P(DIET \mid PA) \cdot P(T2D \mid DIET, AGE). \end{aligned}$$

Bayesian networks are used to make an inference (21); therefore, the basic concepts for inference flow when new information is introduced in a BN are presented here.

Two variables X and Y in a BN are *d-separated* if, for every possible path between X and Y, there is an intermediate variable Z such that either (i) the connection is serial ($X \to Z \to Y$ or $X \leftarrow Z \leftarrow Y$) or diverging ($X \leftarrow Z \to Y$) and Z is instantiated, or (ii) the connection is converging ($X \to Z \leftarrow Y$) and neither Z nor any of Z's descendants have received evidence. It seems necessary to know when influence flows from a node X to another node Y *via* a node Z, when this occurs, it is said that the trail $X \rightleftharpoons Z \rightleftharpoons Y$ is active. A causal trail $X \to Z \to Y$ (serial



AGE	PA=no	PA=yes
18-32	0.3727	0.6273
33-47	0.5780	0.4220
48-62	0.6617	0.3383

PA	DIET=no	DIET=yes
no	0.8563	0.1437
yes	0.2210	0.7790

AGE=18-32	AGE=33-47	AGE=48-62
0.1281	0.4619	0.4100

DIET	AGE	T2D=no	T2D=yes
no	18-32	0.7625	0.2375
no	33-47	0.7254	0.2746
no	48-62	0.5952	0.4048
yes	18-32	0.9868	0.0132
yes	33-47	0.9424	0.0576
yes	48-62	0.9064	0.0936

FIGURE 1
Example of BN with the conditional probability tables of each node. In the Figure P(AGE = 18−32) = 0.1281, P(PA = no|AGE = 18−32) = 0.3727, P(DIET = no|PA=no) = 0.8563, and P(T2D = no|AGE = 18−32, DIET=no) = 0.7625.

connection), an evidential trail $X \leftarrow Z \leftarrow Y$ (serial connection) or, a common cause trail $X \leftarrow Z \rightarrow Y$ (diverging connection) is active if and only if Z is not observed. A common effect trail $X \rightarrow Z \leftarrow Y$ (converging connection) is active if and only if either Z or one of Z's descendants is observed. For instance, in Figure 2, there are two possible paths between *PA* and *T2D*, i.e., $PA \rightarrow DIET \rightarrow T2D$ and $PA \leftarrow AGE \rightarrow T2D$, when *DIET* and *AGE* are instantiated then *PA* and *T2D* would be *d-separated*.

The Bayesian network $\mathcal{B} = (\mathcal{G}, \theta)$ satisfies the local Markov condition if, for each variable (feature) X_i, X_i is conditionally independent of the set of all its non-descendants given the set of all its parents. The global Markov property states that any node X_i is conditionally independent of any other node given its *Markov blanket*, i.e., $I(X_i, non\text{-}markov\text{-}blanket(X_i) \mid markov\text{-}blanket(X_i))$; the *Markov blanket* of a node includes its parents, its children, and the children's other parents (spouses). Any node in the BN $\mathcal{B}$ would be *d-separated* from the nodes belonging to the non-Markov blanket given its Markov blanket.

The Bayesian network models have been widely used successfully in different fields, such as diagnostic diseases (14, 22–28), neuroscience (15, 29), analysis of complex disease systems (30–32), clinical decision support (33, 34), human immunodeficiency virus (HIV) and influenza research (35, 36), and even in interactions between multiple diseases (37).

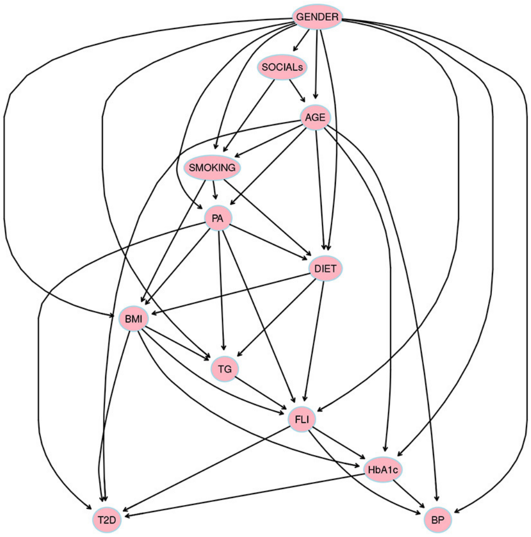


FIGURE 2
The structure obtained through hill climbing learning algorithm *hc* from *bnlearn* package in R language using a threshold = 0.85 by *model averaging* over 500 networks. Prior knowledge was included in model selection, thus variables were divided into four blocks: 1) *background variables* = {*GENDER*, *AGE*, and *SOCIALs*}, 2) *conditional variables* = {*DIET*, *SMOKING*, *PA*, and *BMI*}, 3) *intermediate variables* = {*HbA1c*, *FLI*, *BP*, and *TG*}, and, 4) *diagnostic variable* = {*T2D*}.

Methods

Participants

The present study examined a cohort of Spanish working adults ($n = 16,648$, Men: 12,080, Women: 4,568), who had prediabetes at baseline. Participants were selected from the working population of 234,995 potentially eligible individuals, belonging to different productive sectors (public administration, construction, healthcare, service industries, or postal services) who received occupational health examinations between 2012 and 2013. Criteria for selecting the subjects were as follows: age between 20 and 65 years and FPG between 100 and 125 mg/dl, according to ADA criteria (38). Exclusion criteria were FPG $\geq$ 126 mg/dl or HbA1c $\geq$ 6.5% at baseline, current treatment with oral antidiabetic or diagnosed with T2D, anemia, current treatment with systematic glucocorticoid, cancer treatment, and pregnancy. All subjects underwent a standard health examination, sociodemographic characteristics, anthropometric measurements, lifestyle, and clinical data at baseline. Follow-up examinations were performed at 5 years (2017–2018). The study protocol was conducted in accordance with the Declaration of Helsinki with human participants and was approved by the Institutional Review Board of the Balearic Islands Health Research Ethics Committee (CEI-IB Ref. No: 1887). All participants were informed of the purpose of the study before they gave consent to participate.

Data collection and definition of variables

Most methods have been described in greater detail previously (39, 40). The following data were collected: sociodemographic variables, including age, gender, education level, and social status. Social status was ascertained using the Spanish adaptation of the Goldthorpe classification suggested by the Spanish Epidemiology Society classification (41).

Anthropometric measurements: All anthropometric measurements were performed by well-trained technicians and were made according to the recommendations of the International Standards for Anthropometric Assessment (ISAK) (42). Body weight was measured to the nearest 0.1 kg using an electronic scale (Seca 700 scale, Seca gmbh, Hamburg). Height was measured to the nearest 0.5 cm using a stadiometer (Seca 220 Telescopic Height Rod for Column Scales, Seca gmbh, Hamburg). BMI was calculated as weight (kg) divided by height (m) squared (kg/m^2). Participants were categorized depending on their BMI following WHO criteria: normal weight (BMI = 18.5–24.9 kg/m^2), overweight (BMI = 25.0–29.9 kg/m^2), and obese (BMI $\geq$ 30.0 kg/m^2) (43). Waist circumference (WC) was measured half-way between the lower costal border and the iliac crest, using a flexible steel tape (Lufkin Executive Thinline W 606). Systolic (SPB) and diastolic blood pressure (DBP) were measured in triplicate, with a 1 min gap between measurements, using an electric and calibrated sphygmomanometer (OMRON M3, Healthcare Europe, Spain). Hypertension was defined as SBP $\geq$ 140 mmHg, or DBP $\geq$ 90 mmHg, or taking antihypertensive medication.

Blood samples were taken at baseline and 5 years. Venous blood samples were collected from the antecubital vein after a 12 h overnight fast in suitable vacutainers without anticoagulants to obtain serum. Serum concentrations of glucose, TG, GGT, and cholesterol were measured by standard procedures using an autoanalyzer (SYNCHRON CX®9 PRO, Beckman Coulter clinical system, Brea, CA, USA). High triglycerides (TG) was defined as $\geq$ 150 mg/dl, and high cholesterol was defined as $\geq$ 200 mg/dl.

Fatty liver index (FLI) was used as a surrogate measure of fatty liver. FLI is a validated risk score system based on routine measurements of TG and gamma-glutamyl transferase (GGT) concentrations, WC, and BMI (44), and accurately identifies non-alcoholic fatty liver disease (NAFLD) and hepatic steatosis in the general population (40, 45).

Participants were asked if they engaged in moderate and/or vigorous physical activity (PA) at least 150 min/week according to WHO recommendations and if they consumed vegetables and fruits daily. Participants, also, were classified as non-smokers, current smokers, or former smokers, according to WHO criteria.

Prediabetes was defined as FPG between 100 and 125 mg/dl according to the ADA criteria (38). Incident T2D was defined as the clinical diagnosis of T2D, FPG $\geq$ 126 mg/dl, or the initiation of anti-hyperglycemic treatment at follow-up.

Bayesian networks

In the process of learning a BN, the following steps have to be carried out: i) *structural learning* that will determine the topology of the BN or DAG and ii) *parametric learning* or estimation of conditional probabilities among the nodes once the network topology is established.

Structural learning

The problem of discovering the causal structure increases with the number of variables (46–48). Table 1 shows a description of the 12 variables considered.

Basically, two approaches to structural learning can be considered (49): (i) *search-and-score* structure learning and (ii) *constraint-based* structure learning. Search algorithms based on search and score assign a number (score) to each BN, and then the BN structure with the highest score is selected. Constraint-based search algorithms determine a set of conditional independence analyses on the data (50), which is used to generate an undirected graph, then additional independence tests have to be considered in order to obtain the BN structure.

The package *bnlearn* (51, 52) of R language (53) was used to learn the structure, where prior knowledge of the variables under study was taken into account in order to reduce the number of structures that are consistent with the same set of independencies and to choose a structure which reflects the causal order and dependencies. The following blocks of variables were considered: i) *background variables* = {*Gender*, *Age*, and *Social status*}, ii) *conditional variables* = {*Diet*, *Smoking*, *PA*, and *BMI*}, iii) *intermediate variables* = {*HbA1c*, *FLI*, *BP*, and *TG*}, and iv) *diagnostic variable* = {*T2D*}.

By blacklisting arrows that point from a later block to an earlier block, the process of model selection was restricted

TABLE 1 Description of 12 data set features used to learn the structure.

Variable name	Description	Values
GENDER	Male and Female	Men, Women
T2D	Type 2 diabetes	Yes, No
SOCIALs	Social status	I, II, III
SMOKING	Never, Former or current smoker	No, Former smoker, Yes
PA	Physical activity (At least 150 min/week)	Yes, No
DIET	Daily consumption of fruits and vegetables	Yes, No
HbA1c	Glycosylated haemoglobin	Less 6.0, More 6.0
FLI	Fatty liver index	Less 30, 30–60, More 60
BMI	Body mass index (kg/m^2)	Underweight, Normal weight,
		Overweight, Obesity
BP	Blood pressure (mmHg)	Normal, High, Grade 1, Grade 2
TG	Triglycerides (mg/dl)	Normal, Limit, Hyper
AGE	Age interval in years	18-32, 33-47, 48-62

(54). There are two possible options to obtain the structure: either select a *single best* model or else obtain some average model, which is known as *model averaging* (55). The model (see Figure 2) was learned with *hill-climbing (hc)* algorithm, a score-based algorithm that explores the search space starting from a network structure (usually the empty graph) and adding, deleting, or reversing one arc at a time until the score can no longer be improved (this process is also known as greedy search). The score used by the structure learning algorithm was the Akaike Information Criterion (AIC).

Parametric learning

Given the topology of the network, parameters were also obtained with the *bnlearn* package in R language by performing a Bayesian parameter estimation using the Dirichlet distribution (56).

A conditional probability distribution is obtained for each node. An example of conditional probability distribution is shown in Table 2.

T2D model

This T2D model allow us to obtain conditional independencies among the variables. In a BN, any node is conditionally independent of its non-descendants given its parents' nodes, i.e., $I(X, non - descendants(X) \mid Pa(Xi))$, as the local Markov property states. For instance, in the BN obtained for T2D model which structure is given in Figure 2, we show some independencies:

$$\begin{aligned} &I(FLI, \{SMOKING, AGE, SOCIALs\} \mid Pa(FLI) \\ &\quad = GENDER, DIET, PA, TG, BMI), \\ &I(HbA1c, \{TG, DIET, PA, SMOKING, SOCIALs\} \mid Pa(HbA1c) \\ &\quad = GENDER, AGE, FLI, BMI), \end{aligned}$$

$$\begin{aligned} &I(TG, \{SMOKING, AGE, SOCIALs\} \mid Pa(TG) \\ &\quad = DIET, PA, GENDER, BMI), \\ &I(BP, \{T2D, TG, BMI, DIET, PA, SMOKING, SOCIALs\}|Pa(BP) \\ &\quad = GENDER, AGE, HbA1c,\ FLI), \\ &I(T2D, \{BP, TG, DIET, SMOKING, SOCIALs, GENDER\} \\ &\quad \mid Pa(T2D) = PA, AGE, BMI, FLI, HbA1c). \end{aligned}$$

The *Markov blanket* of the diagnostic feature T2D is composed of five features *AGE*, *BMI*, *FLI*, *HbA1c*, and *PA*. According to the Global Markov property, T2D will remain independent of any other node (in the DAG) given the features that compose its Markov blanked, i.e., once the features in the Markov blanket are instantiated. The Markov blanket of a specific node (in a BN) determines the only features that influence such a node. However, the influence of features such as habit smoking (SMOKING), physical activity (PA), or dietary habits (DIET) may be of interest. In the flow of influence between SMOKING and T2D, the following causal trails in the DAG may be of interest: i) $SMOKING \to BMI \to T2D$; ii) $SMOKING \to BMI \to FLI \to T2D$; iii) $SMOKING \to PA \to BMI \to T2D$; iv) $SMOKING \to PA \to DIET \to TG \to FLI \to T2D$; v) $SMOKING \to BMI \to HbA1c \to T2D$; or vi) $SMOKING \to PA \to DIET \to FLI \to HbA1c \to T2D$. The obtained model allow us to rank the features from the Markov blanket of T2D led to maximization of the probability of the T2D variable in the Yes state (see Table 3).

Validation of the BN

The Bayesian network was validated using a 10-fold cross-validation for BN, using a log-likelihood loss function, obtaining an expected loss of 8.0470. In Table 4, the area under the ROC curve (AUC) and the percentage correctly classified for the different features is shown.

Performance comparison

Other classification performances (see Table 5) have been included in order to have reference benchmarks with respect

TABLE 2 Expected values of probabilities for *age* feature conditional on combinations of its parent values, in this case, conditional on *gender*, and *social status* features.

Gender	Social status	Age = *18–32*	Age = *33–47*	Age = *48–62*
Men	I	0.0681	0.4301	0.5018
Men	II	0.0726	0.4716	0.4558
Men	III	0.1430	0.4682	0.3887
Women	I	0.1147	0.5301	0.3552
Women	II	0.1163	0.5143	0.3694
Women	III	0.1306	0.4279	0.4416

TABLE 3 Ranking the features from the Markov blanket of *T2D* led to maximization of the probability of the T2D variable in the *Yes* state, where in the initial BN without introduction of evidence *T2D* in *Yes* state reached a probability (expressed in percentage) of 22:3%.

Rank	Instantiated variable		Value	T2D = Yes
1	*BMI*	=	Obesity	67.3%
2	*HbA1c*	=	More than 6.0	63.5%
3	*FLI*	=	More than 60	56.4%
4	*PA*	=	No	37.2%
5	*AGE*	=	48–62	30.7%

to our BN, in particular, we include naïve bayes (NB), random forest, multilayer perceptron (MLP), and the ID3 algorithms WEKA (57). The performance of each classification model is evaluated using four statistical measures: accuracy, sensitivity, specificity, and ROC area.

Results

Once the BN is built (see Figures 2, 3), it is used to make inferences, i.e., probabilities are updated when new information is introduced (21). In order to make inferences, different reasoning patterns can be adopted (12, 14): *causal reasoning* (from top to bottom), *evidential reasoning* (from bottom to up), and *intercausal reasoning* (very close to human reasoning, it happens when different causes of the same effect can interact). The concept of the *Markov blanket* of a node [composed of its parents, its children, and the children's other parents (spouses)] is frequently used in order to reduce the features (variables) that may influence one another.

Analysis with the Markov blanket

In order to maximize the *T2D* variable in the *Yes* state, we considered the *Markov blanket* of *T2D* which is composed of *age*, *BMI*, *FLI*, *HbA1c*, and *PA* features, and we introduced evidence, selecting for each feature from the Markov blanket, the state which maximizes the most *T2D* in the *Yes* state (as *T2D* is on the bottom, a causal reasoning pattern is performed).

In Table 3, the features from the Markov blanket of *T2D* are ranked, showing the variable that increases the most *T2D* in the *Yes* state, being *BMI* in state *Obesity* which increases the most *T2D* in the *Yes* state going from an initial conditional probability value of 0.2230 to a 0.6730 value, followed by *HbA1c* in *more than 6.0* value increasing the conditional probability to 0.6350, and by *FLI* in *more than 60* value increasing the conditional probability to 0.5640, which is also shown in Figures 4–8 at step 1 in each of them.

In the following, a study of the likelihood variability for the *T2D* variable is considered taking into account the different states for each of the features (variables) that belong to the Markov blanket.

Influence of age on T2D

Figure 4 shows the likelihood variability for *T2D* in the different age groups, with 48–62 years being the age group that increases the most *T2D* in the *yes* state. Furthermore, as we observe in the groups of 18–32 and 33–47 years, it is the *HbA1c* in *more than 6* state which increases the most *T2D* in the *yes* state followed by *BMI* in the *Obesity* state. In group 48–62 is the feature *BMI* in *Obesity* state the one which increases the most *T2D* in *yes* state followed by *HbA1c* in *more than 6* state and *Physical Activity (PA)* in *No* state. In the age group 33–47, there is also a high influence of *BMI* in the *Obesity* state, it is observed that as the age decreases the influence of *BMI* in the *Obesity* state on *T2D* in the *yes* state also decreases. As we can observe, the group of 48–62 years is the one that increases the most *T2D* in the *yes* state, and therefore the one with the highest risk of developing *T2D*. Once *PA*, *BMI*, and *FLI* have been instantiated to *No*, *Obesity*, and *more than 60* states, respectively, the group of 18–32 goes from a conditional probability of *T2D* in the *Yes* state of 0.572 (57.2% expressed in percentage) to 0.907 (90.7% expressed in percentage) when the *HbA1c* is instantiated to *more than 6* state. In the case of the 33–47 group, this likelihood variates from 0.669 (66.9% expressed in percentage) to 0.932 (93.2% expressed in percentage).

TABLE 4 AUCs and percentage correctly classified for the different features.

Variable name	State	AUC	Accuracy
GENDER	Men	0.8100	77.62
GENDER	Women	0.8099	77.62
T2D	Yes	0.9829	94.56
T2D	No	0.9825	94.56
SOCIALs	I	0.5973	78.86
SOCIALs	II	0.5928	78.86
SOCIALs	III	0.5952	78.86
SMOKING	Former smoker	0.6552	48.82
SMOKING	Yes	0.6229	48.82
SMOKING	No	0.6165	48.82
PA	No practice	0.9685	92.24
PA	Practice	0.9685	90.24
DIET	Yes	0.9037	83.52
DIET	No	0.9039	83.52
HbA1c	Lower 6.0	0.8415	86.49
HbA1c	Upper 6.0	0.8419	86.49
FLI	Lower 30	0.9596	80.45
FLI	30-60	0.8655	80.45
FLI	Upper 60	0.9683	80.45
BMI	Underweight	0.9042	78.23
BMI	Normal weight	0.9381	78.23
BMI	Overweight	0.8476	78.23
BMI	Obesity	0.9588	78.23
BP	Normal	0.7137	49.60
BP	High	0.5742	49.60
BP	Grade 1	0.6723	49.60
BP	Grade 2	0.7367	49.60
AGE	18–32	0.7443	55.72
AGE	33–47	0.6281	55.72
AGE	48–62	0.7076	55.72
TG	Normal	0.8949	76.47
TG	Limit	0.7806	76.47
TG	Hyper	0.9031	76.47

Influence of BMI in T2D

In Figure 5, the likelihood variability for *T2D* in the different labels of BMI is shown, with *Obesity* being the one that increases the most *T2D* in the *yes* state, follow by *Overweight*. On other hand, no influence is shown in the groups for *BMI* in *Normal* and *Underweight* states. The highest influence in the groups of BMI for *Obesity* and *Overweight* states is obtained again when *HbA1c* is instantiated to *more than 6* state. The *Obesity* group is less influenced by the remaining instantiations while the *Overweight* group has still a high influence when *FLI* to *more than 60* state is instantiated, being also influenced (to a lesser extent) by *PA* in the *No* state and the *48–62* group of *age*. With both groups *BMI* in the *Obesity* state and in the *Overweight* state groups of risk, achieving a likelihood of having *T2D* in the *yes* state at the end of the instantiations was 0.994 (99.40% expressed in percentage) and 0.717 (71.70% expressed in percentage), respectively.

Influence of PA in T2D

Figure 6 shows the likelihood variability for *T2D* in the different labels of PA, the *No* state being the one that increases the most *T2D* in the *Yes* state. For those that practice PA, the highest influence in *T2D* (increasing its likelihood in yes state) is given by the *HbA1c* in *more than 6* state, the *BMI* in the *Obesity* state, and *FLI* in *more than 60* state, while for those that do not practice PA, the highest influence on T2D includes the *FLI* in *more than 60* state and the *HbA1c* in state *more than 6*.

In the worst-case scenario, i.e., when (*BMI*) is instantiated to the *Obesity* state, *HbA1c* is instantiated to *more than 6* state, *FLI* is instantiated to *more than 60*, and *age* is instantiated to *48-62*, the estimated conditional probability for those who practice PA achieves a value of 0.962 (96.2% expressed in percentage), while for those that do no practice PA achieves a value of 0.994 (99.4% expressed in percentage).

Influence of Fatty Liver Index in T2D

Figure 7 shows the likelihood variability for *T2D* in the different labels of *FLI*, with *more than 60* being the one that increases the most *T2D* in the *Yes* state, and therefore the one with highest risk of developing *T2D*. The *HbA1c* in *more than 6* state has the highest influence in all the groups determined by *FLI*, in the group *30–60*, *BMI* in the *Obesity* state also has a high influence.

Once the variables have been instantiated, the estimated conditional probability for *T2D* in the *yes* state is 0.0031 (0.31% expressed in percentage) in the group *<30*, 0.08 (8.00% expressed in percentage) in the group *30–60*, and 0.56 (56% expressed in percentage) in the group *more than 60* at step 1 in Figure 7, achieving at step 5 an estimated conditional probability of 0.0064 (0.64% expressed in percentage), 0.9420 (94.20% expressed in percentage), and 0.9940 (99.40% expressed in percentage), respectively.

Influence of HbA1c in T2D

Figure 8 shows the likelihood variability for *T2D* at the different labels of *HbA1c*, with *more than 6* being the one

TABLE 5 Performance for T2D feature comparing our BN and using a 10-fold cross validation experiments with the corresponding algorithms.

Algorithms	Accuracy	Sensitivity	Specificity	ROC area
Bayesian network	94.5639	0.9455	0.8268	0.9826
Logistic regression	94.3589	0.9440	0.83429	0.9740
Naïve Bayes	90.7196	0.9070	0.9160	0.9740
Random forest	94.2636	0.9430	0.8770	0.9670
Multilayer perceptron	94.4918	0.9450	0.8870	0.9440
ID3	93.6389	0.9440	0.8720	0.9170

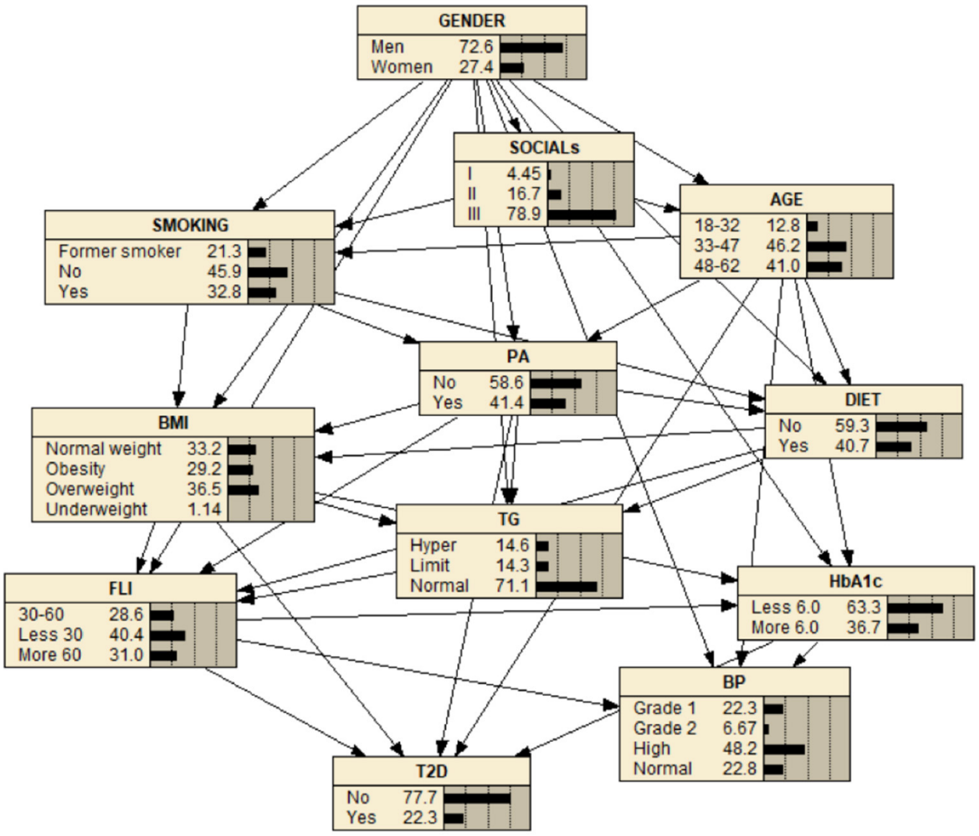


FIGURE 3
BN for the study of T2D. The BN shows high (48.2%) blood pressure (BP), normal (71.1%) triglycerides (TG), less 30 level (40.4%) FLI, overweight (36.5%) (BMI), less than 6.0 level HbA1c (63.3%), and practice physical activity (PA) (40.4%) and no practice physical activity (PA) (58.6%). It also shows a level of T2D equal to 22.3% in the yes state.

that increases the most *T2D* in the *Yes* state. The *BMI* in the *Obesity* state has a high influence on developing *T2D* in <6 group followed by *FLI* in *more than 60* state and *PA* in the *No* state; while in the *more than 6* group, *PA* in the *No* state, *FLI* in *more than 60* state, and *BMI* in the *Obesity* state have a similar influence. Once the variables have been instantiated in step 1 the estimated conditional probability for *T2D* in the *yes* state is 0.0269 (2.69% expressed in percentage) in the group $HbA1c < 6$, and 0.6350 (63.50% expressed in percentage), in the group $HbA1c > 6$, achieving at step 5 an estimated conditional probability of 0.3070 (30.70% expressed in percentage) and 0.994 (99.4% expressed in percentage), respectively.

Intercausal reasoning

The influence of some variables to reduce the risk of developing *T2D* is considered, taking into account that

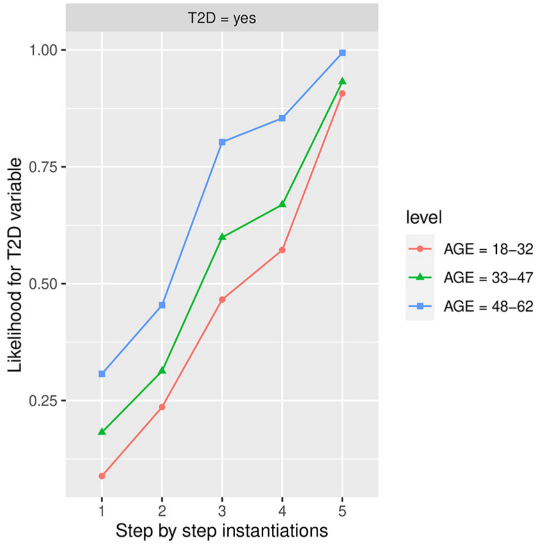


FIGURE 4
The different steps: step 1 = *Age*, step 2 = *PA* in the *No* state, step 3 = *BMI* in the *Obesity* state, step 4 = *FLI* in state *More than 60*, and step 5 = *HbA1c* in state *More than 6.0*, to evaluate *T2D* feature. The different steps are represented in the horizontal axis, while the estimated probability for the *T2D* variable at the value *Yes* is shown in the vertical axis.

influence flows, in this sense, the Markov blanket cannot be completely considered, only some variables to allow the flow among variables.

Influence of diet and PA

Figure 9 shows the influence of *Diet*, once *Diet* has been instantiated to the *No* state or the *Yes* state, the estimated likelihood of developing *T2D* is 0.3310 (33.10% expressed in percentage) and 0.0618 (6.18% expressed in percentage) respectively. The highest influence in the *No* diet group is given by *Body Mass Index* (*BMI*) in the *Obesity* state, while in the *Yes* diet group, it is given by *PA*, in the *No* state and *BMI* in the *Obesity* state. The risk of developing *T2D* is increased in both groups when *BMI* is instantiated to the Obesity state, reaching an estimated likelihood of developing *T2D* is 0.7130 (71.30% expressed in percentage) and 0.4660 (46.60% expressed in percentage) respectively, showing the strongest influence. Other factors such as *smoking* in the *former* state, *age* in *48-62* state, and *gender* in the *men* state increase the risk of developing *T2D*, reaching an estimated likelihood of developing *T2D* at step 5 of 0.8430 (84.30% expressed in percentage) and 0.5910 (59.10% expressed in percentage), respectively.

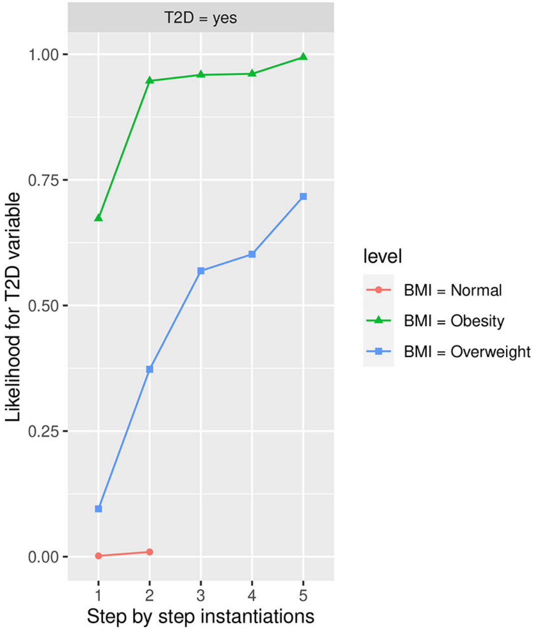


FIGURE 5
The different steps: step 1 = *BMI*, step 2 = *HbA1c* in state *More than 6*, step 3 = *FLI* in state *More than 60*, step 4 = *PA* in state *No*, and step 5 = *Age* in state *48-62* to evaluate *T2D* feature. The different steps are represented in the horizontal axis, while the estimated probability for the *T2D* variable at the value *Yes* is shown in the vertical axis.

Discussion

In this study, the feasibility of BNs in epidemiological studies is demonstrated, in particular when data from T2D risk factors are considered. Clinical questions based on unobserved evidence can be answered through specific BN models due to automatically updated probability distributions when new patient information is introduced.

The development and analysis of models to examine the relationships between different factors acting on T2D could be not only of theoretical interest but can serve as a generic tool for application oriented activities: explanation, prediction, monitoring, and prevention. BN models allow the theoretical analysis of the relationships between numerous variables in an appealing way, taking into account the probabilistic nature of the causal dependencies, in this sense, BNs constitute an adequate tool in the study of T2D. The ability of BN models of creating different scenarios based on hypothetical cases when new observations are considered to make BN models a powerful knowledge representation and an efficient reasoning tool under conditions of uncertainty. Furthermore, using the T2D model, a characterization of the whole set of variables could be given.

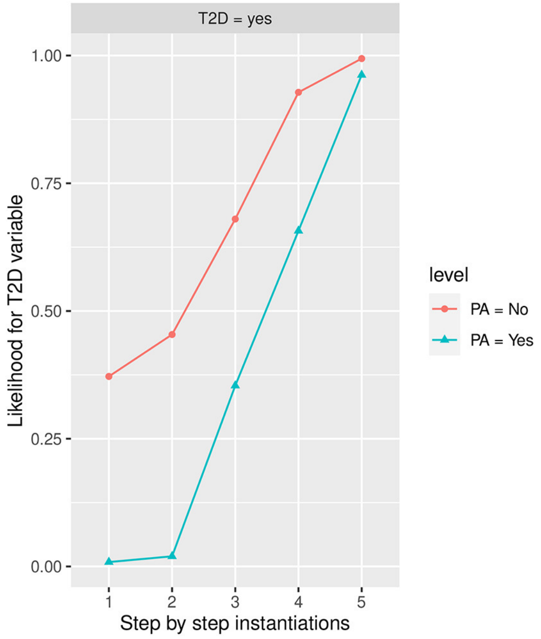


FIGURE 6
The different steps: step 1 = *PA*, step 2 = *AGE* in state *48-62*, step 3 = *FLI* in state *More than 60*, step 4 = *HbA1c* in state *More than 6.0*, and step 5 = *BMI* in state *Obesity* to evaluate *T2D* feature. The different steps are represented in the horizontal axis, while the estimated probability for the *T2D* variable at the value *Yes* is shown in the vertical axis.

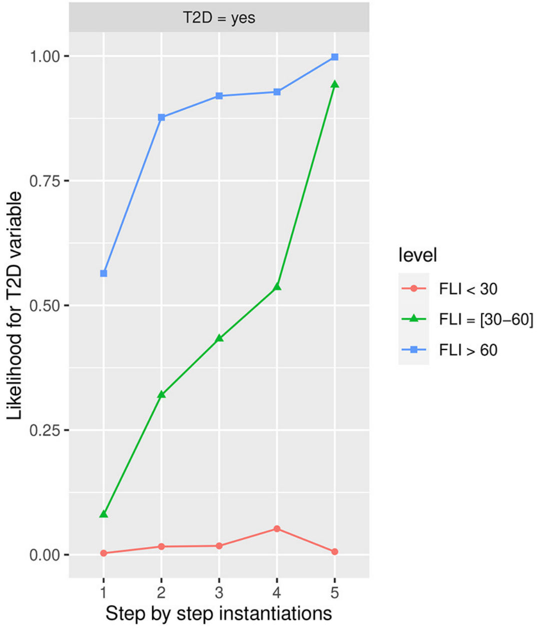


FIGURE 7
The different steps: step 1 = *FLI*, step 2 = *HbA1c* in state *More than 6.0*, step 3 = *AGE* in state *48-62*, step 4 = *PA* in the *No* state, and step 5 = *BMI* in the *Obesity* state, to evaluate *T2D* feature. The different steps are represented in the horizontal axis, while the estimated probability for the *T2D* variable at the value *Yes* is shown in the vertical axis.

The main difference with respect to other T2D studies from prediabetes in the literature is that intercausal reasoning together with the concept of a Markov blanket were considered in order to optimize the T2D feature. The BN model is selected because they produce probability estimates rather than predictions. The process of learning the structure of a BN is a form of unsupervised learning, the learner does not distinguish the dependent variable from the independent ones, which is an advantage when compared with regression.

This longitudinal 5-year follow-up study evaluates risk factors for the progression from prediabetes to T2D among workers using a BN model and the *Markov blanket* concept. Our results showed that obesity and high levels of HbA1c are determinants for the progression to T2D. Furthermore, PA is an important protective factor even in the presence of other risk factors. The results of the present analysis are in accordance with previous evidence reporting that obesity is the main risk factor for T2D (8, 58, 59). Specifically, the risk for progression to T2D is very high (more than 50% risk) in prediabetic obese patients with and without high levels of HbA1c. However, high levels of HbA1c may help distinguish overweighted patients who will convert from those who will not. Similarly, when the instantiation begins with HbA1c, we also observe that obesity is a strong risk factor for conversion in subjects with high levels of HbA1c. But obesity and high levels of HbA1c are strongly associated; more than 60% of the patients who were obese also presented high levels of HbA1c. Notably, most of the normal weighted patients will not convert independently of the presence of other risk factors. High levels of FLI could also contribute to determining the risk of T2D, as it has also been previously described (60), especially in the case of overweight and obese patients. In addition, in overweight patients with high HbA1c levels, FLI (> 60) strongly increases the risk of conversion to T2D. Interestingly, age is not a good predictor of developing T2D compared with other factors, and the main risk factors are important at all ages. In this way, our results highlighted that the main risk factors for conversion to T2D apply at different age groups with similar behavior at different steps and reaching a very high risk for progression to T2D (near 90% risk) independently of age. Although, some studies suggest that age could be a modest independent risk factor (61, 62), the practice of PA is an important lifestyle that could delay or avoid the progression to T2D in people with prediabetes, as previously described (8). We observed that people who practice and do not practice PA, obesity, HbA1c, and FLI are the factors that strongly increases T2D risk (according

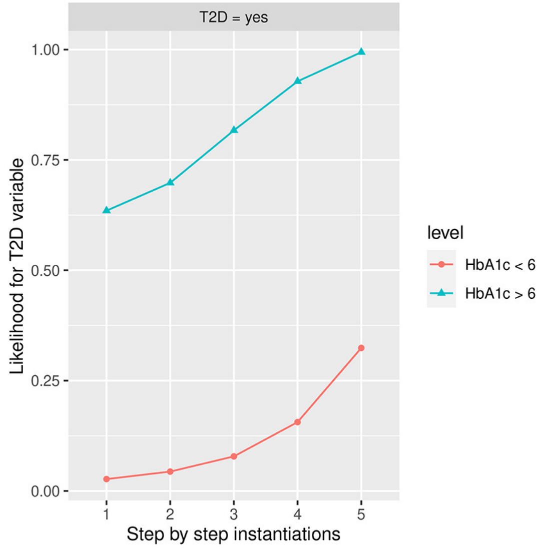


FIGURE 8
The different steps: step 1 = *HbA1c*, step 2 = *AGE* in state *48–62*, step 3 = *PA* in state *No*, step 4 = *FLI* in state *More than 60*, and step 5 = *BMI* in the *Obesity* state to evaluate *T2D* feature. The different steps are represented in the horizontal axis, while the estimated probability for *T2D* variable at the value *Yes* is shown in the vertical axis.

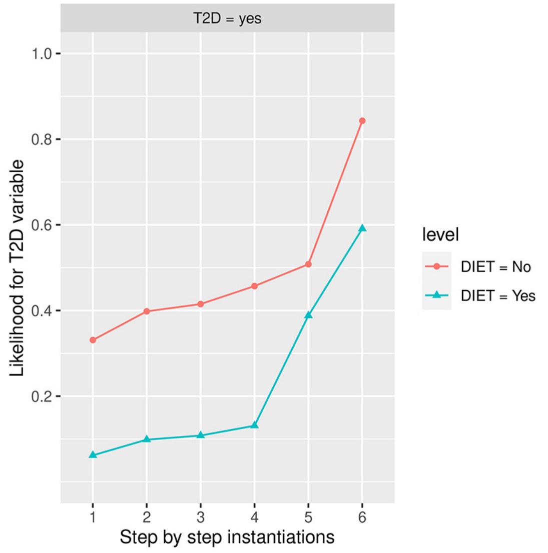


FIGURE 9
The different steps: step 1 = *DIET*, step 2 = *age* in state *48–62*, step 3 = *gender* in state *men*, step 4 = *SMOKING* in state *former smoker*, step 5 = *PA* in the *No* state, and step 6 = *BMI* in the *Obesity* state. The different steps are represented in the horizontal axis, while the estimated probability for *T2D* variable at the value *Yes* is shown in the vertical axis.

to the previously mentioned instantiation). Notably, even the presence of high levels of HbA1c is only an important risk factor for patients that do not practice PA. We used BN analysis to evaluate the influence of different variables in the progression to T2D from prediabetes. Compared with other types of analysis, such as logistic regression that also use the outcome as a binary, BN analysis evaluates the risk of different conditions (FLI, HbA1c, PA, and Diet), adding the presence of different variables instead of adjusting for the effect of other factors as logistic regression (63) that evaluate the adjusted effect of each condition. BN modeling is a more practical approach for clinical purposes since it allows more clinical use because it provides probability estimates for different scenarios that clinicians and patients could easily interpret. BNs could serve as a tool for helping clinicians in the management of risk factors assessment and clinical decision-making (14, 64). Early intervention is essential for T2D prevention, and BNs may allow clinicians to identify patients at high risk of developing T2D. For example, early intervention should be done without additional tests in patients with obesity, at a very high risk of conversion. But in overweight patients, clinicians should request additional tests to determine the real risk. It would not be necessary for patients with healthy weight because they are at low risk of conversion to T2D.

This study presents some limitations that should be acknowledged. First, a possible misclassification bias was when subjects were categorized as having prediabetes based on a single blood sample. Second, diet and PA were not evaluated with a validated questionnaire. On the other side, the main strength of the study was the large sample size with a 5-year follow-up. Furthermore, the study population was representative of the Spanish workforce.

Conclusion

Our results confirm that obesity and high levels of HbA1c are the main risk factors for the progression to T2D, while PA is an important lifestyle protective factor. The BN analysis is an advanced model for dynamic description and prediction of the development of T2D. Furthermore, the BNs tool could be a feasible strategy to help clinicians with T2D prevention and motivate patients to adopt a healthier lifestyle that reduces their T2D risk.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by Institutional Review Board of the

Balearic Islands Health 77 Research Ethics Committee (CEI-IB Ref. No: 1887). The patients/participants provided their written informed consent to participate in this study.

Author contributions

PF-P, MB-V, and AY: conceptualization and writing original draft preparation. PF-P and MB-V: methodology. PF-P and AY: formal analysis. AL-G: investigation. AL-G and AA: data curation, writing review, and editing. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Khan MAB, Hashim MJ, King JK, Govender RD, Mustafa H, Al Kaabi J. Epidemiology of Type 2 diabetes - global burden of disease and forecasted trends. *J Epidemiol Glob Health*. (2020) 10:107–11. doi: 10.2991/jegh.k.191028.001

2. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 44 million participants. *Lancet*. (2016) 387:1513–30. doi: 10.1016/s0140-6736(16)00618-8

3. Chamnan P, Simmons RK, Forouhi NG, Luben RN, Khaw KT, Wareham NJ, et al. Incidence of type 2 diabetes using proposed HbA1c diagnostic criteria in the european prospective investigation of cancer-norfolk cohort: implications for preventive strategies. *Diabetes Care*. (2011) 34:950–6. doi: 10.2337/dc09-2326

4. Tabak AG, Herder C, Rathmann W, Brunner EJ, Kivimaki M. Prediabetes: a high-risk state for diabetes development. *Lancet*. (2012) 379:2279–290. doi: 10.1016/S0140-6736(12)60283-9

5. Huang Y, Cai X, Mai W, Li M, Hu Y. Association between prediabetes and risk of cardiovascular disease and all cause mortality: systematic review and meta-analysis. *BMJ*. (2016) 355:i5953. doi: 10.1136/bmj.i5953

6. Vistisen D, Kivimaki M, Perreault L, Hulman A, Witte DR, Brunner EJ, et al. Reversion from prediabetes to normoglycaemia and risk of cardiovascular disease and mortality: The Whitehall II cohort study. *Diabetologia*. (2019) 62:1385–90. doi: 10.1007/s00125-019-4895-0

7. Ezquerra-Lazaro I, Cea-Soriano L, Giraldez-Garcia C, Ruiz A, Franch-Nadal J, Diez-Espino J, et al. Lifestyle factors do not explain the difference on diabetes progression according to type of prediabetes: results from a Spanish prospective cohort of prediabetic patients. *Diabetes Res Clin Pract*. (2019) 153:66–75. doi: 10.1016/j.diabres.2019.05.033

8. Bennasar-Veny M, Fresneda S, López-González A, Busquets-Cortés C, Aguiló A, Yañez AM. Lifestyle and progression to Type 2 diabetes in a cohort of workers with prediabetes. *Nutrients*. (2020) 12:1538. doi: 10.3390/nu12051538

9. Saeedi P, Petersohn I, Salpea P, Malanda B, Karuranga S, Unwin N, et al. Global and regional diabetes prevalance estimates for 2019 and projections for 2030 and 2045: results from the International Diabetes Federation Diabetes Atlas, 9(th) edition. *Diabetes Res Clin Pract*. (2019) 157:107–843. doi: 10.1016/j.diabres.2019.107843

10. Richter B, Hemmingsen B, Metzendorf MI, Takwoingi Y. Development of type 2 diabetes mellitus in people with intermediate hyperglycaemia. *Cochrane Database Syst Rev*. (2018) 10: CD012661. doi: 10.1002/14651858.CD012661.pub2

11. Stevens, JW, Khunti K, Harvey R, Johnson M, Preston L, et al. Preventing the progression to type 2 diabetes mellitus in adults at high risk: a systematic review and network meta-analysis of lifestyle, pharmacological and surgical interventions. *Diabetes Res Clin Pract*. (2015) 107:320–31. doi: 10.1016/j.diabres.2015.01.027

12. Koller D, Friedman N. *Probabilistic Graphical Models. Principles and Techniques*. Cambridge, MA, London: The MIT Press (2010).

13. Pearl J. *Causality Models, Reasoning and Inference*. Cambridge: Cambridge University Press (2000).

14. Fuster-Parra P, Bennasar-Veny M, Ligeza A, López-González AA, Aguiló A. Bayesian network modeling: a case study of an epidemiologic system analisys of cardiovascular risk. *Comput Methods Programs Biomed*. (2016) 126:128–42. doi: 10.1016/j.cmpb.2015.12.010

15. DeFelipe J, López-Cruz PL, Benavides-Piccione R, Bielza C, Larra naga P, et al. New insights into the classification and nomenclature of cortical GABAergic interneurons. *Nat Rev Neurosci*. (2013) 14:202–16. doi: 10.1038/nrn3444

16. Fuster-Parra P, García-Mas A, Ponseti FJ, Palou P, Cruz J. A Bayesian network to discover relationships between negative features in sport: a case study of teen players. *Qual Quant*. (2014) 48:1473–91. doi: 10.1007/s11135-013-9848-y

17. Fuster-Parra P, García-Mas A, Ponseti FJ, Leo FM. Team performance and collective efficacy in the dynamic psychology of competitive team: a Bayesian network analysis. *Hum Mov Sci*. (2015) 40:98–118. doi: 10.1016/j.humov.2014.12.005

18. Larrañaga P, Moral S. Probabilistic graphical models in artificial intelligence. *Appl Soft Comput*. (2011) 11:1511–28. doi: 10.1016/j.asoc.2008.01.003

19. Mihaljević B, Bielza C, Larrañaga P. Bayesian networks for interpretable machine learning and optimization. *Neurocomputing*. (2021) 456:648–65. doi: 10.1016/j.neucom.2021.01.138

20. Jensen FV, Nielsen TD. Bayesian networks and decision graphs. In: Jordan M, Lauritzen SL, Lawless JF, Nair V editors. *Information Science & Statistics*. New York, NY: Springer Science+Business Media New York (2007).

21. Butz CJ, Hua S, Chen J, Yao H. A simple graphical approach for understanding probabilistic inference in Bayesian networks. *Inform Sci*. (2009) 179:699–716. doi: 10.1016/j.ins.2008.10.036

22. Antal P, Fannes G, Timmerman D, Moreau Y, Moor BD. Bayesian applications of belief networks and multilayer perceptrons for ovarian tumor classification with rejection. *Artif Intell Med*. (2003) 29:29–60. doi: 10.1016/S0933-3657(03)00053-8

23. Antal P, Fannes G, Timmerman D, Moreau Y, Moor BD. Using literature and data to learn Bayesian networks as clinical models of ovarian tumors. *Artif Intell Med*. (2004) 30:257–81. doi: 10.1016/j.artmed.2003.11.007

24. Charitos T, Gaag LC, Visscher S, Schurink KAM, Lucas PJF. A dynamic Bayesian network for diagnosing ventilator-associated pneumonia in ICU patients. *Expert Syst Appl*. (2009) 36:1249–58. doi: 10.1016/j.eswa.2007.11.065

25. Liu S, Zhang R, Shang X, Li W. Analysis for warning factors of type 2 diabetes mellitus complications with Markov blanket based on a Bayesian network model. *Comput Methods Programs Biomed*. (2019) 188:105302. doi: 10.1016/j.cmpb.2019.105302

26. Maskery SM, Hu H, Hooke J, Shriver CD, Liebman MN. A Bayesian derived network of breast pathology co-occurrence. *J Biomed Inform*. (2008) 41:242–50. doi: 10.1016/j.jbi.2007.12.005

27. Wang XH, Zheng B, Good WF, King JL, Chang YH. Computer assisted diagnosis of breast cancer using a data-driven Bayesian belief network. *Int J Med Inform*. (1999) 54:115–26. doi: 10.1016/S1386-5056(98)00174-9

28. Kanda E, Kanno Y, Katsukawa F. Identifying progressive CKD from healthy population using Bayesian network and artificial intelligence: a worksite-based cohort study. *Sci Rep*. (2019) 9:5082. doi: 10.1038/s41598-019-41663-7

29. Mihaljević B, Larrañaga P, Benavides-Piccione R, DeFelipe J, Bielza C. Comparing basal dendrite branches in human and mouse hippocampal CA1 pyramidal neurons with Bayesian networks. *Sci Rep*. (2020) 10:18592. doi: 10.1038/s41598-020-73617-9

30. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*. (2003) 302:449–453. doi: 10.1126/science.1087361

31. Lewis FI, Brälisauer F, Gunn GJ. Structure discovery in Bayesian networks: an analytical tool for analysing complex animal health data. *Prev Veterin Med*. (2011) 100:109–15. doi: 10.1016/j.prevetmed.2011.02.003

32. Lewis FI, McCormick BJ. Revealing the complexity of health determinants in resource-poor settings. *Am J Epidemiol*. (2012) 176:1051–9. doi: 10.1093/aje/kws183

33. Sesen MB, Nicholson AE, Banares-Alcantara R, Kadir T, Brady M. Bayesian networks for clinical decision support in lung cancer care. *PLoS ONE*. (2013) 8:e82349. doi: 10.1371/journal.pone.0082349

34. Zhang Z, Zhang J, Wei Z, Ren H, Song W, Pan J, et al. Application of tabu search-based Bayesian networks in exploring related factors of liver cirrhosis complicated with hepatic encephalopathy and disease identification. *Sci Rep*. (2019) 9:6251. doi: 10.1038/s41598-019-42791-w

35. Lycett SJ, Ward MJ, Lewis FI, Poon AFY, Kosakovsky Pond SL, Leigh Brown AJ. Detection of mammalian virulence determinants in highly pathogenic avian influenza H5N1 viruses: multivariate analysis of published data. *J Virol*. (2009) 83:9901–10. doi: 10.1128/JVI.00608-09

36. Poon AF, Lewis FI, Pond SL, Frostt SDW. Evolutionary interactions between N-linked glycosylation sites in the HIV-1 envelope. *PLoS Comput Biol*. (2007) 3:e30011. doi: 10.1371/journal.pcbi.0030011

37. Lappenschaar M, Hommerson A, Lucas PJF, Lagro J, Visscher S. Multilevel Bayesian networks for the analysis of hierarchical health care data. *Artif Intell Med*. (2013) 57:171–83. doi: 10.1016/j.artmed.2012.12.007

38. American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes Care*. (2012) 35: S64-S71. doi: 10.2337/dc12-s064

39. Bennasar-Veny M, Fresneda S, López-González A, Busquets-Cortés C, Yáñez AM. Lifestyle and progression to Type 2 diabetes in a cohort of workers with prediabetes. *Nutrients*. (2020) 12:1–13.

40. Busquets-Cortés C, Bennasar-Veny M, López-González AA, Fresneda S, Abbate M, Yáñez AM. Utility of Fatty Liver Index to predict reversion to normoglycemia in people with prediabetes. *PLoS ONE*. (2021) 16:e0249221. doi: 10.1371/journal.pone.0249221

41. Domingo-Salvany A, Bacigalupe A, Carrasco JM, Espelt A, Ferrando J, Borrell C. del Grupo de Determinantes Sociales de Sociedad Espanola de Epidemiología. [Proposals for social class classification based on the Spanish National Classification of Occupations 2011 using neo-Weberian and neo-Marxist approaches]. *Gac Sanit*. (2013) 27:263–72. doi: 10.1016/j.gaceta.2012.12.009

42. Stewart A, Marfell-Jones M, Olds T, De Ridder J. *International Standards for Anthropometric Assessment*. Wellington: International Standards for Anthropometric Assessment (ISAK) (2011).

43. World Health Organization. *Obesity: Preventing and managing the global epidemic. Report of a WHO consultation*. World Health Organization technical report series (2000).

44. Jäger S, Jacobs S, Kröger J, Stefan N, Fritsche A, Weikert C, et al. Association between the fatty liver index and risk of type 2 diabetes in the EPIC-Potsdam study. *PLoS One*. (2015) 10:1–14. doi: 10.1371/journal.pone.0124749

45. Calori G, Lattuada G, Ragogna F, Garancini MP, Crosignani P, Villa M, et al. Fatty liver index and mortality: the cremona study in the 15th year of follow-up. *Hepatology*. (2011) 54:145–52. doi: 10.1002/hep.24356

46. Buntine W. A guide to the literature on learning probabilistic networks from data. *IEEE T Knowl Data Eng*. (1996) 8:195–210. doi: 10.1109/69.494161

47. Cheng J, Greiner R, Kelly J, Bell D, Liu W. Learning Bayesian networks from data: an information-theory based approach. *Artif Intell*. (2002) 137:43–90. doi: 10.1016/S0004-3702(02)00191-1

48. Sucar LE, Martínez-Arroyo M. Interactive structural learning of Bayesian networks. *Expert Syst Appl*. (1998) 15:325–32. doi: 10.1016/S0957-4174(98)00050-5

49. Daly R, Shen Q, Aitken S. Learning Bayesian networks: approaches and issues. *Knowl Eng Rev*. (2011) 26:99–157. doi: 10.1017/S0269888910000251

50. Margaritis D. *Learning Bayesian network model structure from data* (Ph.D. thesis) of CMU-CS-03-153 (2003).

51. Nagarajan R, Scutari M, Lèbre S. *Bayesian Networks in R: with Applications in Systems Biology*. New York, NY: Springer (2013).

52. Scurati M. Learning Bayesian networks with the bnlearn R package. *J Stat Soft*. (2010) 35:1–22. doi: 10.18637/jss.v035.i03

53. R Core Team. *R: A language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing (2021). Available online at: https://www.R-project.org/

54. Hojsgaard S, Edwards D, Lauritzen S. *Graphical Models with R*. New York, NY: Springer (2012).

55. Claeskens G, Hjort NL. *Model Selection and Model Averaging*. Cambridge: Cambridge University Press (2008).

56. Neapolitan RE. *Learning Bayesian networks*. Upper Saddle River, NJ: Prentice Hall, Inc. (2003).

57. Frank E, Hall MA, Witten IH. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, 4th Edn*. Burlington, MA: Elsevier (2016).

58. Schnurr TM, Jakupović H, Carrasquilla GD, Ängquist L, Grarup N, Sϕrensen TIA, et al. Obesity, unfavourable lifestyle and genetic risk of type 2 diabetes: a case-cohort study. *Diabetologia*. (2020) 63:1324–32. doi: 10.1007/s00125-020-05140-5

59. Eckel RH, Kahn SE, Ferrannini E, Goldfine AB, Nathan DM, Schwartz MW, et al. Obesity and type 2 diabetes: what can be unified and what needs to be individualized? *J Clin Endocrinol Metab*. (2011) 96:1654–63. doi: 10.1210/jc.2011-0585

60. Busquets-Cortés C, Bennasar-Veny M, López-González AA, Fresneda S, Aguiló A, Yañez A. Fatty liver index and progression to type 2 diabetes: a 5-year longitudinal study in Spanish workers with pre-diabetes. *BMJ Open*. (2021) 11:e045498. doi: 10.1136/bmjopen-2020-045498

61. Alva ML, Hoerger TJ, Zhang P, Gregg EW. Identifying risk for type 2 diabetes in different age cohorts: does one size fit all? *BMJ Open Diabetes Res Care*. (2017) 5:e000447. doi: 10.1136/bmjdrc-2017-000447

62. Caspersen CJ, Thomas GD, Boseman LA, Beckles GLA, Albright AL. Aging, diabetes, and the public health system in the United States. *Am J Public Health*. (2012) 102:1482–97. doi: 10.2105/AJPH.2011.300616

63. Lemeshow S, Hosmer DW Jr. A review of goodness of fit statistics for use in the development of logistic regression models. *Am J Epidemiol*. (1982) 115:92–106. doi: 10.1093/oxfordjournals.aje.a113284

64. Sesen MB, Nicholson AE, Banares-Alcantara R, Kadir T, Brady M. Bayesian networks for clinical decision support in Lung Cancer Care. *PLoS ONE*. (2013) 8:e82349.