Computers in Biology and Medicine 190 (2025) 109967

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

A Bayesian Belief Network model for the estimation of risk of cardiovascular events in subjects with type 1 diabetes

Ornella Moro [a,*], Inger Torhild Gram [b,c], Maja-Lisa Løchen [d], Marit B. Veierød [e], Ana Maria Wägner [f], Giovanni Sebastiani [a,g,h]

[a] Istituto per le Applicazioni del Calcolo "Mauro Picone", Consiglio Nazionale delle Ricerche, Rome, Italy
[b] Norwegian Centre for E-health Research, University Hospital of North Norway, Tromsø, Norway
[c] Department of Community Medicine, Faculty of Health Sciences, UiT The Arctic University of Norway, Tromsø, Norway
[d] Department of Clinical Medicine, Faculty of Health Sciences, UiT The Arctic University of Norway, Tromsø, Norway
[e] Oslo Centre for Biostatistics and Epidemiology, Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, Oslo, Norway
[f] Endocrinology and Nutrition Department. Complejo Hospitalario Universitario Insular Materno-Infantil. Instituto de Investigaciones Biomédicas y Sanitarias (IUIBS). Universidad de Las Palmas de Gran Canaria (ULPGC). Las Palmas de Gran Canaria, Spain
[g] Dipartimento di Matematica Guido Castelnuovo, Sapienza Università di Roma, Rome, Italy
[h] Department of Mathematics and Statistics, UiT The Arctic University of Norway, Tromsø, Norway

ARTICLE INFO

Keywords:
Risk assessment
Bayesian Belief Network
Cox proportional hazard model
Statistical inference
Simulation study
Cardiovascular diseases
Type 1 diabetes

ABSTRACT

Objectives: Cardiovascular diseases (CVDs) represent a major risk for people with type 1 diabetes (T1D). Our aim here is to develop a new methodology that overcomes some of the problems and limitations of existing risk calculators. First, they are rarely tailored to people with T1D and, in general, they do not deal with missing values for any risk factor. Moreover, they do not take into account information on risk factors dependencies, which is often available from medical experts.

Method: This study introduces a Bayesian Belief Network (BBN) model to quantify CVD risk in individuals with T1D. The developed methodology is applied to a large T1D dataset and its performances are assessed. A simulation study is also carried out to quantify the parameter estimation properties.

Results: The performances of individual risk estimation, as measured by the area under the ROC curve and by the C-index, are about 0.75 for both real and simulated data with comparable sample sizes.

Conclusions: We observe a good predictive ability of the proposed methodology with accurate parameter estimation. The BBN approach takes into account causal relationships between variables, providing a comprehensive description of the system. This makes it possible to derive useful tools for optimising intervention.

1. Introduction

Type 1 diabetes (T1D), also known as insulin-dependent diabetes, is an autoimmune metabolic disease characterised by chronic high blood glucose concentration. In this disease, insulin, which controls blood glucose concentration, is not produced by the pancreas, requiring intensive injections [1]. Cardiovascular disease (CVD) is a leading cause of morbidity and mortality worldwide, with an estimated 19.8 million of deaths from CVD in 2022 [2,3]. People with T1D are at increased risk of developing CVD due to the effects of chronic hyperglycaemia [4–9].

Risk of CVD for individuals with T1D can be reduced by improving the control of hypoglycaemia and adopting a healthy lifestyle, such as quitting smoking or increasing physical activity [10–12]. Personalised risk prediction methods, specifically developed for people with T1D, can be an important tool to help these individuals prevent the occurrence of CVD events. There are several validated risk calculators (RCs) that estimate the risk of CVD in a certain time span. Two of the most commonly used are the NORRISK2 [13], for the general population, and the STENO1 [4], for people with T1D.

Most of these RCs are based on the Cox proportional hazard (CPH) model [14], which describes the evolution of risk over time given the values of some risk factors. However, among the available RCs, only a couple of them are specifically designed for people with T1D [4]. Moreover, in the standard application of the CPH, there is no need to consider the dependence between the factors involved, while this is relevant when focusing on the dependence of the output on a subset of factors. Another example is related to the estimation of the risk for

* Corresponding author.
E-mail address: o.moro@iac.cnr.it (O. Moro).

https://doi.org/10.1016/j.compbiomed.2025.109967
Received 2 October 2024; Received in revised form 8 January 2025; Accepted 1 March 2025
Available online 20 March 2025

a new patient when there are missing values for some factors, which are usually replaced by their population means.

A possible alternative to the CPH model is to adopt a Bayesian Belief Network (BBN), as done here. We remark that BBN is the most suited type of statistical model to take into account the causal relationships between the factors. This also allows to cope with missing values and to estimate the dependence of the output on any subset of factors for all possible combinations of their values. Information on the causal dependencies between factors can often be provided by medical experts. However, it cannot be used by the CPH, which is the basic model of standard RCs. The situation is the same if one wants to use a model belonging to Artificial Intelligence (AI), e.g. Deep Learning (DL), which is now being successfully applied in many different fields. In addition, such models require much larger data samples for training than those used here. In fact, DL has recently been successfully applied to predict CVD events in subjects with T1D [15]. However, there the model is trained with a very large sample.

The use of BBN in medical science has become progressively more frequent in various fields [16–21]. In recent years, some studies have also proposed BBN models to predict the risk of occurrence of non-communicable diseases, such as cancer or CVD. Stojadinovic and colleagues [22] developed a BBN-based model to perform prognostic assessment for colon carcinomatosis and decision-making for consequent treatment. They applied the model proposed to a sample of 53 patients and performed a 10-fold cross-validation procedure [23,24]. The overall area under the receiver operating characteristic (ROC) curve (AUC) [25] obtained was equal to 0.71, while the sensitivity and specificity values were 68.3% and 63.3%, respectively. There are also a couple of works using BBN model to describe probabilistically the relationships between CVD events and some related risk factors [26, 27].

The first of the two methods above used a Dynamical Bayesian Network to perform prognosis for coronary heart disease (CHD) and applied it to a longitudinal dataset of about 850 middle-aged men. Dynamical Bayesian Networks are an extension of BBNs structured to include time, whose use in medical application is increasing [28–30]. In this case, the risk of a given event at time t was estimated based on the patient's medical history before that time. The authors proposed two variations of the model, one using time point data and one including temporal abstractions (TAs) which are high-level temporal concepts obtained from point time data [31]. Given the low probability of the event considered, the sample population resulted highly imbalanced towards the subjects who did not experience any event. Thus, four different types of resampling techniques were applied to the real dataset obtaining a total of five different datasets. The two variations of the model were then tested on all the five samples using a 10-fold cross-validation procedure. The maximum sensitivity value obtained was 75%, which corresponded to the application of the model with TAs to one of the resampled dataset. For this same dataset, they also built the ROC curve and calculate the AUC: 0.60 and 0.78 for the model without and with TAs, respectively. The overall performances of the model showed a good prediction ability for CHD events and the used approach resulted very promising from the medical point of view. However, the performances dropped substantially when considering the model without TAs where the maximum sensitivity obtained was 63%. This could represent a strong limitation, given that the effective use of the TAs version requires a high number of records per subject over time and a low proportion of missing values. Furthermore, the performance values depended on the resampling technique used, for both model types. Indeed, the sensitivity value for the model including TAs ranged between 52% and 75%, while the model without TAs fluctuated from 38% to 63%.

In the second study, Ordovas et al. [27] developed a BBN to describe the relationships between a disease, i.e. diabetes, and two medical abnormalities, i.e. hypertension and hypercholesterolemia, and ten related factors, both modifiable and non-modifiable. A first structure of the BBN was retrieved using the Greedy Thick Thinning algorithm [32]. The resulting version was then critically modified by medical experts. To estimate the parameters of the model (i.e. the local conditional probabilities), the authors applied a Bayesian approach using a multinomial-Dirichlet distribution with uniform prior. However, the setup and the main objectives of this work are quite different from ours. First, we only consider subjects with T1D. Instead, in the cited work, subjects with type 1, type 2 and no diabetes are included. Furthermore, we predict probabilistically future outcomes (within 7 years) of CVD events, e.g. stroke, myocardial infarction, based on the values of some risk factors in recent past (5 years). In contrast, in the work of Ordovas et al. no prediction is performed. Furthermore, the whole set of (13) variables considered for each subject, including also those for the disease and the medical abnormalities considered, are relative to the same time. More precisely, they cumulate data acquired within a 5-year interval, after checking homogeneity with respect to time.

In this study, we propose a new methodology for the quantification of the risk of CVD events in people with T1D:

- A specific BBN model for CVD events in people with T1D, developed in close collaboration with medical experts, is adopted here for the first time.
- In contrast to existing RCs for the above scope, the BBN model takes into account dependencies between risk factors.
- Instead of providing risk score, as done by standard RCs, we quantify the risk in terms of a real number, i.e. probability value.
- One component of the BBN model incorporates the basic mathematical expression of the CPH model.
- The method also copes with the case when the values of some factors are missing.
- Unlike known RCs, after statistical inference is drawn, the probability of the event under study conditioned on any subset of factors can be computed, which can be useful for intervention.

The training and testing of the model is performed on a large dataset of about 1300 subjects with T1D [33]. This is also done in the case where the values of some factors are missing. To quantify the performances of the estimation process, a simulation study is performed.

2. Models and methods

Below, we start giving some details of the CPH model and of the relative parameter inference. In fact, we will adopt it in the BBN model for the conditional probability of the output (CVD), given the factors. Then, we focus on the BBN model, also describing the parameter estimation procedure. Finally, we illustrate the methods used here to quantify the performances of the proposed methodology.

2.1. CPH model

The CPH model assumes that, given the time T_E of first occurrence of an event E, the hazard function $h(t)$, defined as

$$h(t) = \lim_{\Delta \to 0^+} \frac{\mathcal{P}(t \le T_E < t + \Delta \mid t \le T_E)}{\Delta}, \tag{1}$$

can be expressed as a product of two separate functions $h_0(t)$ and $\rho(\boldsymbol{x})$ [14]. The first one is the baseline hazard function and depends only on the time, whereas the second is a function of some covariates (risk factors) $\boldsymbol{x}$

$$h(t) := h_{\boldsymbol{x}}(t) = h_0(t) \cdot \rho(\boldsymbol{x}). \tag{2}$$

As a consequence, the survival function $S(t) = \mathcal{P}(T_E \ge t)$ can be expressed as

$$S(t) = e^{-\int_0^t h_{\boldsymbol{x}}(u)du} = S_0(t)^{\rho(\boldsymbol{x})}, \tag{3}$$

where $S_0(t) = e^{-\int_0^t h_0(u)du}$ is the baseline survival function. A common choice, also made here, corresponds to a Weibull distribution for the baseline survival $S_0(t) := e^{-\frac{t^\kappa}{\gamma}}$, which leads to the following survival function

$$S(t) = e^{-\frac{t^\kappa}{\gamma}\cdot\rho(\boldsymbol{x})}. \tag{4}$$

The most frequently used parametric model for the covariate function $\rho := \rho(\boldsymbol{x};\boldsymbol{\beta})$, which we also adopt here, is the exponential one:

$$\rho(\boldsymbol{x};\boldsymbol{\beta}) = e^{\boldsymbol{\beta}^T\boldsymbol{x}}, \tag{5}$$

where $\boldsymbol{\beta}$ is a parameter vector, and T indicates the transposition.

To estimate the model parameters, longitudinal data are usually used. Following Cox's approach, we first estimate the $\boldsymbol{\beta}$ parameters by maximising the *partial likelihood* of the data [14]. To build it, suppose we have n individuals, of which d experienced the event. For the ith individual, among the d who experienced the event E, let t_i denote the first time the event occurred. For all the remaining $n-d$ individuals, the time t_i represents their censoring time. In fact, some of the individuals not experiencing the event during the study may have left it before its end. Note that uncensored individuals not experiencing the event until the final time are here considered 'censored' at that time. The partial likelihood is then expressed as follows [14]

$$\mathcal{L}_p(\boldsymbol{\beta}) = \prod_{j=1}^{d} \frac{\rho(\boldsymbol{x}_j;\boldsymbol{\beta})}{\sum_{k:t_k\geq t_j}\rho(\boldsymbol{x}_k;\boldsymbol{\beta})}. \tag{6}$$

In this case, the product is indexed only over the d individuals experiencing the event, while the contributions of the others appear in the denominator. The estimated values $\hat{\boldsymbol{\beta}}$ of the parameters are obtained by maximising the partial likelihood:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \mathcal{L}_p(\boldsymbol{\beta}). \tag{7}$$

Now, we can plug these estimates into the full likelihood to estimate the remaining parameter vector $\boldsymbol{\tau} = (\gamma,\kappa)$ of the baseline survival $S_0(\cdot)$. In that likelihood, the contributions of censored and uncensored individuals multiply as follows

$$\mathcal{L}(\hat{\boldsymbol{\beta}},\boldsymbol{\tau}) = \prod_{uncensored} f(t_i;\hat{\boldsymbol{\beta}}) \prod_{censored} S(t_i;\hat{\boldsymbol{\beta}}), \tag{8}$$

where $f(t_i;\boldsymbol{\beta}) = -S'(t_i)$ and the times t_i are defined as above. The values $\hat{\boldsymbol{\tau}}$ are obtained maximising the likelihood function in Eq. (8)

$$\hat{\boldsymbol{\tau}} = \arg\max_{\boldsymbol{\tau}} \mathcal{L}(\hat{\boldsymbol{\beta}},\boldsymbol{\tau}). \tag{9}$$

In the followings, for different purposes, we will use an empirical version of the survival curve. Due to the presence of censoring, the Kaplan–Meier (KM) estimator [34] is almost always the one adopted to estimate the survival curve, as done here. This is due to its versatility, as it is a non-parametric estimator. In addition, the way it is obtained, i.e. by a product over measured and censoring times, allows to take into account data censored up to each of those times. Furthermore, it has been proven [35,36] that the asymptotic properties of this estimator are close to those of other parametric and non-parametric ones.

2.2. BBN model

One of the two main elements of a BBN model is a directed acyclic graph (DAG), whose nodes represent the variables describing the system under study and whose directed edges represent the causal relationships between them [37]. Here, a generic node represents either the variable corresponding to the output event (CVD) or one of the risk factors. The set of all variables is denoted here by $\boldsymbol{x} = (CVD, \boldsymbol{x}_f)$, where CVD is the output node and $\boldsymbol{x}_f$ the set of all risk factors. Note that the symbol $\boldsymbol{x}_f$ used here corresponds to the one $\boldsymbol{x}$ used in the previous section. Each variable is here considered random. The second element of a BBN is represented by the so-called *local conditional probabilities*. These are the conditional probabilities of each factor, given its *parents*, according to the DAG structure. As usual, a node A is *parent* of a node B if there exists a directed arrow from A to B, meaning that B is causally dependent on A. Assigned the two elements of a BBN, we are able to express the joint probability of the whole set of factors as a product of the local conditional probabilities. We can in general express the joint probability of a BBN model as follows:

$$\mathcal{P}(\boldsymbol{x}) = \mathcal{P}(x_1,\dots,x_h) = \prod_j \mathcal{P}(x_j \mid pa_j), \tag{10}$$

where, for simplicity, we denote by $\mathcal{P}(\boldsymbol{x})$ the probability that $\boldsymbol{X}$ assumes the value $\boldsymbol{x}$, and by $\mathcal{P}(x_j \mid pa_j)$ the probability that the variable X_j assumes the value x_j given that the set of its parent variables $\mathcal{P}a_j$ assumes the configuration pa_j.

To retrieve the structure of the DAG, there are different approaches, which can be data-driven, expert-based, or a combination of both (see for example [38]). In our case, we take the second choice and the result is shown in Fig. 1. Finding the structure of the DAG consists of drawing oriented edges between pairs of variables. The comprehensive list of dependencies for the DAG in Fig. 1 is presented in Figure S1 of the Supplementary material. The presence of each edge here is justified by the existence of a physiological process. All the processes involved have been intensively discussed with clinicians and epidemiologists of the WARIFA project [39].

The estimation procedure for such a complex graph requires a huge computational cost. Furthermore, one main scope of this work consists of being beneficial like the risk calculators STENO1 [4] (for people with T1D) and NORRISK2 [13] (for general population), through an app developed within the WARIFA project. Therefore, the inclusion of a high number of covariates may be not at all ideal both for the parameters estimation and/or to retrieve information for a new user when computing the risk probability. Hence, we proceed as follows. Starting from the original DAG in Fig. 1, we also consider the two RCs for CVD cited above (STENO1 and NORRISK2). We then select the intersection of the sets of variables used in both these RCs and in our DAG. These are age (AGE), sex at birth (SEX), smoking status (SM) and systolic blood pressure (SBP). All the three considered models include also some information on cholesterol levels. In our DAG we have both low-density lipoprotein (LDL) and high-density lipoprotein (HDL) cholesterol, the NORRISK2 uses the HDL cholesterol and the total cholesterol, while the STENO1 includes only the LDL cholesterol. Although the intersection in this case is empty, we prefer to keep both LDL and HDL cholesterol in the model. Indeed, they are quantifying the phenomenon of dyslipidaemia which plays a major role in CVD events.

Since the large database available to us is regarding individuals with T1D, we additionally include some factors specific for this population. Of the two RCs considered, only the STENO1 is designed for people with T1D. The variables of interest for diabetes considered there are albuminuria, estimated glomerular filtration rate, diabetes duration and glycated haemoglobin level (HBA1C). Among those, only HBA1C and diabetes duration are also included in our original DAG in Fig. 1. We decide to exclude the diabetes duration from the analysis since the sample population that we are using is composed of very young people, whose age is highly correlated with the duration of diabetes.

The resulting DAG of the BBN model used in the following, which corresponds to the above set of variables, derived from the original one in Fig. 1, is shown in Fig. 2. We notice that the resulting DAG is not exactly the induced sub-graph relative to the included factors. Indeed, in a preliminary analysis of the real dataset, we do not find some of the dependencies drawn in the chosen DAG. Therefore, in all the following results, we omit them. In particular, HDL depends uniquely on SM and SEX, and neither SBP nor LDL depend on SM. All the other dependencies are kept the same. On the other hand, based on the same analysis, some additional arrows are included. More in details, both HBA1C and SEX are assumed directly influencing the output node CVD. As for the original DAG, in Figure S2 of the Supplementary material we report the list of dependencies for the reduced subDAG.

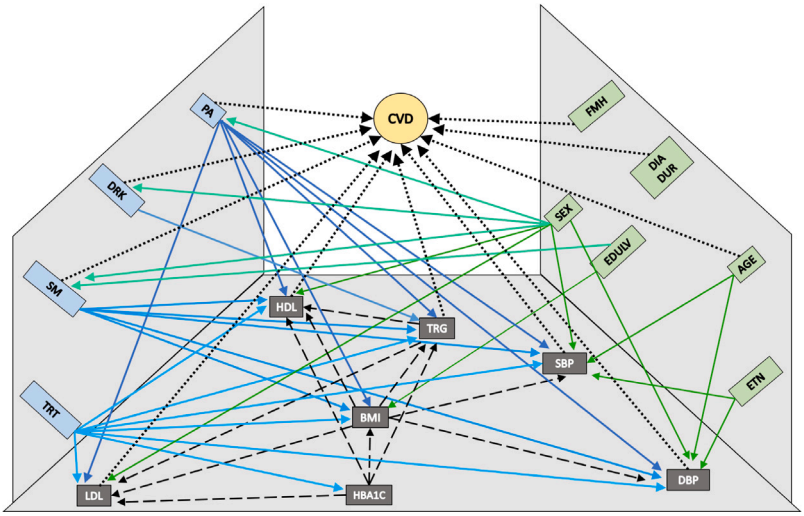

Fig. 1. The original DAG representing causal relationships between pairs of variables for CVD in individuals with T1D. On the three different planes of each figure are reported the variables of different classes. The left plane contains all the modifiable variables, the horizontal plane contains the variables describing the clinical measurements of the subjects and the right plane contains the non-modifiable variables. The dotted arrows are those relative to the direct effects of the factors on the output node (CVD). All the other colours and styles of the arrows are used to improve readability of the DAG. For the meaning of the acronyms, see the description in the text or the list of acronyms at the end of the paper.

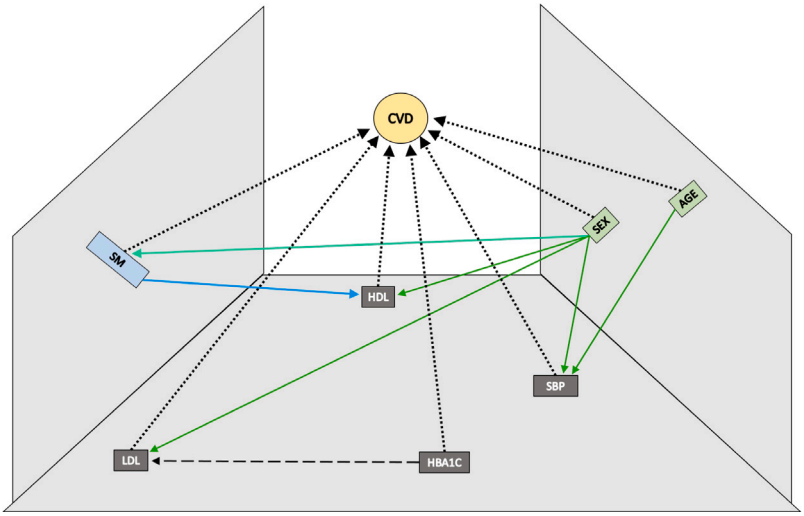

Fig. 2. The subDAG relative to a subset of factors from the complete DAG in Fig. 1. The 3-dimensional organisation of the DAG and the meaning of different arrows are the same as in the caption of Fig. 1.

The seven risk factors considered are of different types. The one of the output node, i.e. the first time of adverse event occurrence, SBP, LDL, HDL and HBA1C are continuous. The SM and SEX variables are binary. The only discrete variable is AGE. The cost required for parameter estimation and some computation with the BBN model, e.g. the population risk, is high. Therefore, we compute the local conditional probability of the continuous variable discretising their supports in subintervals. Specifically, we divide the measured ranges of each factor into equally sized intervals of a length δ specific to each factor. In particular, for HBA1C we choose $\delta = 0.5\%$, for LDL and HDL δ is equal to 10 and 5 mg/dL, respectively, and for SBP it is equal to 10 mmHg. Furthermore, there is evidence in literature [40] showing that while lower values of HDL may increase the risk of CVD, increasing these values does not necessarily protect against such adverse events. Therefore, following a similar approach to the one used in the NORRISK2 [13], we set HDL values greater than 50 mg/dL exactly equal to this value.

We now focus on the chosen local probability models for the BBN in Fig. 2. We consider the DAG of this BBN as a union between two different subDAGs. The first one, DAG_o, models the dependency of the output (CVD) on the covariates ($\boldsymbol{X}_f$). The second one, DAG_f, gives us the dependencies among the set of risk factors. We first focus on the local probability models related to DAG_f, denoted in the followings by $\mathcal{P}(\boldsymbol{X}_f)$. We recall that, as usual for this type of models, the joint distribution is obtained as a product of marginal and conditional distributions.

First, we have to consider the marginal distributions of the three root variables (variables without parents), i.e. AGE, SEX and HBA1C. We model SEX as a Bernoulli random variable and AGE by a probability mass function. For the last variable (HBA1C) we use a discretisation of a continuous Gamma distribution. Then, we cope with the conditional probability models. Looking at the reduced DAG_f in Fig. 2, we notice three different patterns of the relationship children–parents. For one of them, we have a binary variable whose parent is also binary: SM

whose parent is SEX. In this case, for the two possible values of SEX we model SM as a Bernoulli random variable. The second pattern involves HDL, which depends on the two binary variables. Here, for the four possible combinations of SEX and SM, we model HDL as a discretised Gaussian distribution. The last pattern concerns LDL and SBP. Each of these two variables depends on one binary and one continuous or discrete variable (AGE or HBA1C). For these cases, we proceed as follows. For example, for the two possible values of SEX, we model LDL as a discretised Gaussian distribution whose expected value is a linear function of HBA1C with parameters depending on the combination. We proceed similarly for SBP.

Finally, with regard to DAG_o, we adopt a CPH model for the conditional probability of the output given its parents $\boldsymbol{X}_f$

$$\mathcal{P}(CVD \mid \boldsymbol{X}_f = \boldsymbol{x}) = 1 - \exp\left(-\frac{t^\kappa}{\gamma} \cdot e^{\boldsymbol{\beta}^T \boldsymbol{x}}\right). \tag{11}$$

The variables $\boldsymbol{X}_f$ are used here after their standardisation. As a result, the values estimated for the $\boldsymbol{\beta}$ parameters of the linear combination of the covariates quantify the relative influence of the factors on the risk.

To estimate the BBN model parameters, we maximise the likelihood of the measured longitudinal data. Due to the factorisation of the complete joint distribution, the maximisation can be performed separately for each conditional/marginal model appearing in the factorisation. Specifically, we proceed as follows. For the variable X corresponding to each of the two root variables SEX and AGE, we estimate the probability $p_i := \mathcal{P}(X = i)$ by its empirical frequency

$$\hat{p}_i = \frac{N_i}{m}, \tag{12}$$

where m is the total number of observations, N_i is the number of observations for which the variable X assumes the value i equal to 0,1 for SEX and for AGE a number between the minimum and the maximum of the observed values. For the case X = HBA1C, we first estimate the shape κ and the scale θ parameters of the Gamma distribution by maximising the likelihood of the original data, without discretisation. We then compute the cumulative Gamma distribution function $F(\cdot\,; \hat{\kappa}, \hat{\theta})$ corresponding to the estimated shape and scale parameters. Given the discrete division of the HBA1C range in M subintervals $I_1, \ldots, I_M$ as described above, we estimate the probability $\mathcal{P}(X \in I_k)$ of X to belong to $I_k = [z_{k+1}, z_k]$ as

$$\hat{p}_k = F(z_{k+1}) - F(z_k). \tag{13}$$

Finally, since we consider a finite range, we normalise the discrete distribution so obtained.

For the conditional model of SM given SEX, we first select the two subsamples corresponding to the possible values of SEX. The estimated value of the conditional probability $p_{ij} := \mathcal{P}(SM = i \mid SEX = j)$, for $i, j = 0, 1$, is then obtained using the empirical frequency of SM in each of the two subsamples. Thus, we have

$$\hat{p}_{ij} = \frac{N_{ij}}{m_j}, \tag{14}$$

where N_{ij} is the number of observations in the subsample j where $SM = i$ and m_j is the size of the subsample j.

For the conditional model of HDL, we select the four subsamples relative to all possible combinations of SEX and SM. For each combination, we then estimate the parameters of the Gaussian distribution maximising the likelihood of the original data without discretisation relative to the corresponding subsample. For each of the remaining two continuous variables (i.e. LDL and SBP), we first extract the two subsamples corresponding to the values of the involved binary parent variable. For each of the two cases, we need to estimate the two parameters of the linear combination providing variable expected value, conditioned on the value of the remaining continuous parent variable, and the unknown variance. The first two are estimated by maximising the likelihood of the assumed conditional model, using each subsample of data without discretisation. This corresponds to a least square linear fitting. The constant variance σ^2 is estimated as the mean squared errors between measured and theoretical values from the estimated conditional model. Then, as already done for the marginal distribution of HBA1C, we estimate the probability for the child variable to belong in a certain subinterval of the discretisation by numerical integration. Also in this case, given the finite range of values considered, we normalise the discrete distribution obtained.

We estimate the parameters of the CPH model assumed for the output, conditioned on its parents, using first the partial and then the full likelihood of the system, as explained in Section 2.1. Once all the parameters are estimated, we have a full description of the model in terms of the joint probability distribution of the system. In particular, we can compute the population risk $R(t)$ as

$$R(t) = \sum_{\boldsymbol{x}} \mathcal{P}(CVD \mid \boldsymbol{X}_f = \boldsymbol{x}) \cdot \mathcal{P}(\boldsymbol{X}_f = \boldsymbol{x}). \tag{15}$$

Moreover, from the factor values $\boldsymbol{x}$ of a new subject, we can compute the individual risk through the conditional probability $\mathcal{P}(CVD|\boldsymbol{X}_f = \boldsymbol{x})$, using Eq. (11).

2.3. Performance evaluation criteria

To assess the performances of the proposed methodology we consider two indicators: the AUC of the ROC curves [25] and the concordance index or C-index [41] which is obtained as follows. Let us consider all ordered pairs of individuals (i, j), such that the first individual i experiences the event of interest at a time when the individual j is still at risk. This can happen if both the individuals experience the event and $t_i < t_j$, or if the individual j is censored at a time $t_j > t_i$. A pair is considered *concordant* if the risk estimate p_i for the individual i is higher than p_j. The C-index value is the proportion of concordant pairs out of all evaluable ones.

We first perform an internal validation where the training and the test sets correspond to the whole sample available. We also consider a 5-fold cross-validation [23,42]. To minimise the differences due to the particular 5-fold partition of the sample, we repeat the procedure many times and average the results.

In addition to the risk quantification at population or individual levels, one might be interested in proposing intervention. This can be done based on the event probability conditioned on the values of a chosen subset of factors $\bar{\boldsymbol{X}}$. To compute the conditional probability of the output (CVD) given $\bar{\boldsymbol{X}}$, we sum the joint distribution of the BBN model with respect to all values of the remaining variables $\boldsymbol{X}_f \setminus \bar{\boldsymbol{X}}$:

$$\mathcal{P}(CVD \mid \bar{\boldsymbol{X}}) = \frac{\sum_{\boldsymbol{X}_f \setminus \bar{\boldsymbol{X}}} \mathcal{P}(CVD, \boldsymbol{X}_f)}{\mathcal{P}(\bar{\boldsymbol{X}})} = \frac{\sum_{\boldsymbol{X}_f \setminus \bar{\boldsymbol{X}}} \mathcal{P}(CVD \mid \boldsymbol{X}_f) \cdot \mathcal{P}(\boldsymbol{X}_f)}{\mathcal{P}(\bar{\boldsymbol{X}})}. \tag{16}$$

We can also cope with the case of a new user with missing values for a set of factors $\boldsymbol{X}_{miss}$. Then, we compute the probability of CVD, given the values for the observed factors $\boldsymbol{X}_{obs} = \boldsymbol{X}_f \setminus \boldsymbol{X}_{miss}$. A first straightforward possibility is to replace the missing covariates with their averages over the training population. Now, as a better alternative, we can impute the missing values by means of their expected value according to the conditional density, as follows

$$\mathbb{E}\left(\boldsymbol{X}_{miss} | \boldsymbol{X}_{obs}\right) = \int \mathcal{P}(\boldsymbol{x}_{miss} | \boldsymbol{x}_{obs})\, \boldsymbol{x}_{miss}\, d\boldsymbol{x}_{miss}. \tag{17}$$

This quantity is estimated numerically using the average of a sample drawn from the conditional distribution. In the followings, when the method is applied in presence of missing values, we will refer to it as the 'missing model'. Conversely, when no values are missing, we will refer to the method as the 'complete model'.

Table 1
Measured values of the risk factors included in this analysis for the sample population considered and corresponding estimated β parameters. The units used for the different measurement types, the corresponding medians, modes, means, standard deviations and measured ranges are reported from the second to the seventh column, where applicable. In the last column are shown the estimated values of each covariate parameters β appearing in the conditional CPH model of the output node (see Eq. (5)). The variables are ordered according to their estimated (absolute) value of β.

Acronym	Units	Median	Mode	Mean (sd)	Measured range	Estimated β
AGE	years	42	42	41.6 (7)	[28, 54]	0.53
HBA1C	%	8	7.7	8.1 (1.2)	[5, 13]	0.40
HDL	mg/dL	55	50	55 (13)	[28, 112]	−0.37
SBP	mmHg	119	115	118 (10)	[89, 172]	0.30
SEX	–	–	–	–	–	0.16
LDL	mg/dL	113	107	114 (28)	[28, 255]	0.14
SM	–	–	–	–	–	0.04

3. Real database and factors

In the following, when applying the proposed methodology to real data, we use the DCCT/EDIC database which is a large 30-year study led by the National Institute of Diabetes and Digestive and Kidney Diseases of the US National Institutes of Health [33]. The study consists of two phases: the Diabetes Control and Complications Trial (DCCT) which ends at year 10, when the Epidemiology of Diabetes Interventions and Complications (EDIC) observational study starts. This last one continues for the following 20 years [43,44]. In the first DCCT phase, 1441 people with T1D were enrolled, of which 1375 (∼95%) volunteered to participate in the second EDIC phase [43–45]. In both phases, for each of these individuals either the time of the event (if it occurred) or the censoring time are recorded. Furthermore, a large number of risk factors are documented, including the few considered here.

Here, we consider the first occurrence of a CVD event among those in the DCCT/EDIC study. For each of the considered factors, we average the corresponding values during the first five years of the EDIC study to minimise the influence of the fluctuations. We then consider the longitudinal data since the beginning of year 6 of the EDIC study. Unfortunately, possibly due to censoring, we see a high imbalance between events and censoring after the year 12 of the EDIC study. Therefore, we limit our analysis to the 7 years following the first 5 years of the EDIC study. We do not consider, in the EDIC study, the small fraction of individuals (∼2%) with missing values for some factors and those that were censored before the start of our observation period. The final sample used here consists of 1293 subjects, of which 680 males (53%) and 613 females (47%) including 15% smokers. In Table 1, we show a summary of the factors included in the analysis. In addition, the empirical distributions of all the non-binary factors on the whole sample population are reported in Figure S3 of the Supplementary material.

4. Application

Here, we first describe the results obtained from the NIH real dataset of subjects with T1D described in the previous section. Then, we turn to those from a realistic simulated dataset. In both cases, we adopt the same BBN model described above.

4.1. Results on real data

We illustrate now some results from the real dataset about CVD in subjects with T1D [33] regarding both the estimation of the parameters and the performances of the method.

4.1.1. Parameter estimates

In the left panel of Fig. 3, we show the theoretical curve for the population risk, based on the estimated BBN model according to Eq. (15), superimposed to the empirical one, computed through the KM estimator. In the middle panel of the same figure, some individual theoretical risk curves are displayed corresponding to specific values of the covariates. We see quite a large spread of individual curves. However, when averaging them, we obtain a curve very close to the theoretical one derived from estimated BBN model, as shown in the right panel of the same figure. This increases the reliability of the estimated population risk curve. The relative β parameters of the linear combination of the covariates (see Eq. (5)) are estimated on the whole sample population and are reported in the last column of Table 1. The estimated values for the parameters of the baseline survival function (see Eq. (4)) result to be 137.7 for γ and 0.83 for κ.

4.1.2. Performance evaluation

We now present the performance evaluation results, first for individual risk prediction and then for the estimation of the probability of the output conditioned on one or two binary risk factors. In Table 2, we present the performance results relative to the complete model for both internal validation and 5-fold cross-validation. For 5-fold cross-validation, we use 100 random replications. The values obtained do not change when using 200 replications. Results for the case with missing values of LDL and HDL are also included in the same table. These results show a good ability of the methodology in predicting the risk, also in the case of missing data. The ROC curves relative to the internal validation for both the complete and the missing model are shown in Figure S4 of the Supplementary material.

To assess the goodness of the conditional probability estimation, we compare the theoretical risk curve (see Eq. (16)) with the corresponding empirical one, computed through the KM estimator. First, we condition on the single binary variable SM. The results are displayed in Fig. 4. Despite the small amount of smokers in the sample (about 15% of the total), the agreement between the two curves is good, with a mean discrepancy (in absolute value) equal to 0.003. However, this discrepancy increases significantly when conditioning on both SM and SEX, especially for female smokers where the mean discrepancy increases up to 0.01 (see Fig. 4(c)).

4.2. Results on simulated data

In order to have a realistic simulated dataset, we sample from the model described in the previous section corresponding to the parameters estimated from the real dataset described before. Then, we perform parameter estimation and we compare the estimated values with those used to simulate data. This allows us to demonstrate empirically the convergence of the parameter estimators to the true values. The scheme of the simulation process is the following:

- drawing the covariate values according to their joint probability distribution in Eq. (10);
- simulation of the survival times, given the simulated covariate values and according to the used survival function in Eq. (4);
- censoring of the data, based on the observed rate in the real dataset and a uniform distribution for the censoring times.

The same performance evaluation for the real data is applied also for simulated data. To evaluate the effect of the sample size m on the estimation of the parameters, we use five values of $m \in \{800, 1600, 3200, 6400, 12800\}$. Results on both parameter estimates and performance evaluation are illustrated below. To account for variability in the whole process, for each possible size of the sample, we simulate 100 different datasets and repeat the whole estimation process each time.

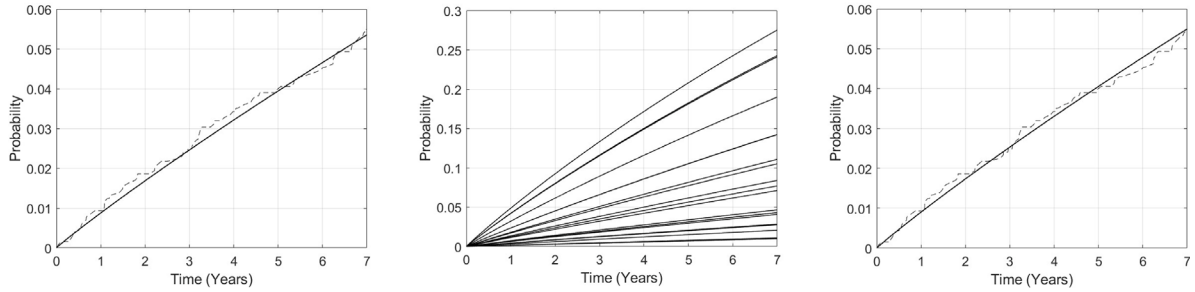

Fig. 3. Population and individual risk curves. On the left (right) panel, the theoretical population risk curve, in solid line, from estimated BBN model (average risk curves of the whole population) is shown over-imposed to the empirical one, in dashed line, computed through KM estimator. On the middle panel, some individual risk curves are shown for some specific values of the covariates.

Table 2
Performance results of the complete and missing models for the real data. In the second and third columns the values of the AUC and C-index for the internal validation are reported. In the last two columns, the mean of the values of 100 random replication of the 5-fold cross-validation are shown, together with their 95% quantile intervals in parenthesis.

Indicator	Internal validation		5-fold cross-validation	
	Complete model	Missing model	Complete model	Missing model
C-Index	0.75	0.73	0.73 (0.68–0.78)	0.71 (0.66–0.75)
AUC	0.76	0.74	0.74 (0.69–0.79)	0.71 (0.67–0.76)

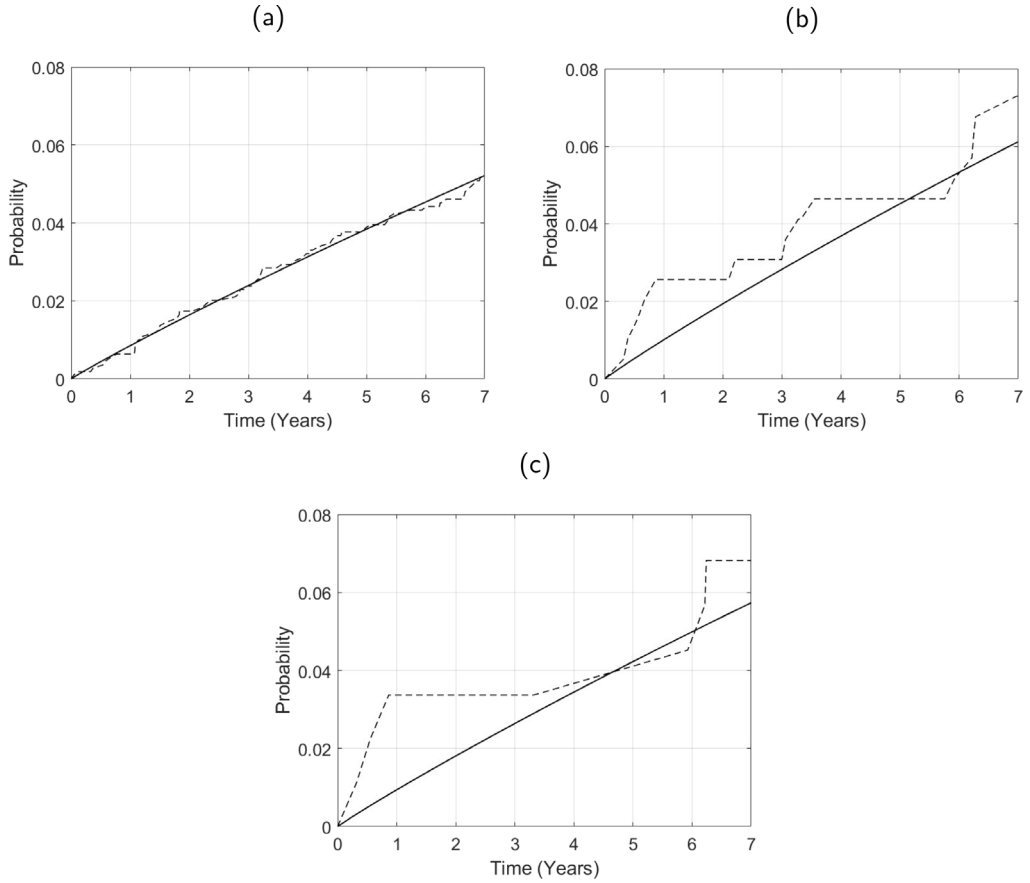

Fig. 4. Comparison between theoretical and empirical (KM estimator) conditional probability curves. On the top panels, we display the two probability curves obtained conditioning on SM alone: non smokers (**a**), smokers (**b**). In the bottom panel, we show the risk curves relative to female smokers, obtained conditioning on SM and SEX together. The corresponding mean absolute discrepancies between the three theoretical curves and the corresponding empirical ones are, respectively, 0.0007, 0.003 and 0.01.

4.2.1. Estimation of parameters

In Table 3, we present some results of the estimates of the β parameters appearing in the conditional CPH model of the output node (see Eq. (5)). These results are relative to $m = 1600$ individuals, which is comparable with the real sample (∼1300), and then to $m = 12800$. The results are good, with deviations of the mean values from the true

Table 3
Results of parameter estimation for simulated data with varying sample size. In the second column are shown the true model values used to simulate data. In the third column, we report the mean values of the parameters estimated from 100 independent samples with $m = 1600$ individuals. In the fourth column, the corresponding mean values of the difference in absolute value between true and estimated parameter values are shown. In the last column, we display the analogous results of the fourth column, for $m = 12\,800$.

Parameter	True value	m = 1600		m = 12 800
		Mean value	Mean difference	Mean difference
AGE	0.53	0.56	0.10	0.04
HBA1C	0.40	0.41	0.08	0.03
HDL	−0.37	−0.38	0.07	0.02
SBP	0.30	0.32	0.10	0.04
SEX	0.16	0.17	0.09	0.03
LDL	0.14	0.12	0.09	0.03
SM	0.04	0.04	0.10	0.03

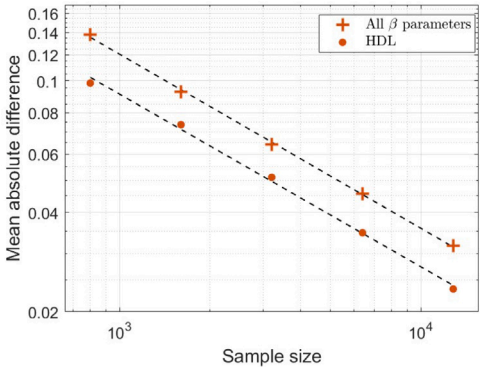

Fig. 5. Mean absolute difference between true (see Table 1) and estimated (see Eq. (5)) values for β parameters versus sample size of simulated data. The parameters are estimated through maximisation of the partial likelihood function (see Eqs. (6), (7)). The two sets of data points correspond to the estimated values for the HDL β parameter (•) and the average values over all the seven parameters (+). The values displayed correspond to the averages over 100 replications. We use a log–log scale and a linear fit is over-imposed to the data points.

ones within 10%. Looking at the mean absolute diffference columns, we notice that the values decrease by a factor 2–3, when increasing the sample size from 1600 to 12 800. The results relative to all the used sample sizes are reported in the Supplementary material, Table S1.

In Fig. 5, we show the decrease of the mean absolute error of the β parameters, when increasing the size of the simulated sample. The data points in the figure, represented in log–log scale, are well described by a linear model with slopes corresponding to −0.52 and −0.53, respectively, for the HDL curve and for the one corresponding to all the β parameters cumulated. This provides evidence that the error decreases as a power law, with the value of the exponent equal to the slope of the straight lines. We notice that the slopes estimated for the different βs are close enough to $-1/2$. Indeed, the mean value of those slopes is −0.53. The minimum value is −0.58 for HBA1C, followed by −0.56 of LDL and −0.53 of SBP. Both the two binary variables and HDL present a slope value equal to −0.52, while the value for AGE result to be the highest one, equal to −0.47. The corresponding curves for all the other β parameters are depicted in Figure S5 in the Supplementary material.

The values of the parameters γ and κ obtained from the simulated sample of 1600 individuals are 157.6 and 0.85, with an absolute mean error equal to 32.5 and 0.07, respectively. These two errors decrease to 10.4 and 0.03 for the 12 800 simulate sample, respectively. Looking at the mean absolute discrepancy between the true baseline survival and the one obtained from the simulated data when increasing the value of m, it decreases as a power law with exponent equal to −0.49. Details of the values estimated from the simulated dataset on varying the sample size are reported in Table S2 in the Supplementary material. In addition, in Figure S6 in the Supplementary material, we show the decrease of the mean absolute error of the two parameters γ and κ, when increasing the size of the simulated sample.

4.2.2. Evaluation of performances

In Table 4, we present the values of the two performance indicators for the simulated data with $m = 1600$ individuals. The results shown are relative to both the complete version of the model and to its variation with missing values of HDL and LDL. Also here the values obtained for the 5-fold cross-validation are relative to 100 random replications, with the same motivation for this choice as for the case of real data.

The results for the conditional probability estimation when $m = 1600$ individuals are comparable to those for real data. Indeed, the discrepancy for female smokers decreases slightly to 0.007 from 0.01 of the real data. This discrepancy depends on some features of the data and not on inadequacy of the model. In fact, increasing the simulated sample to $m = 50000$ individuals, the agreement between the two curves improves importantly and the mean discrepancy decreases to 0.002. Alternatively, increasing percentage of smokers from 15% to 50% and keeping $m = 1600$, the mean discrepancy decreases to 0.005.

5. Discussion

We first focus on the results relative to the real dataset. The estimated values of the $\boldsymbol{\beta}$ parameters, reported in the last column of Table 1, are ordered according to their absolute value. Except for HDL, they are all positive, which means that an increase of the value of the variable corresponds to an increase of the risk. The opposite is only true for HDL, which is a 'protective' factor. We remember that, in this case, for values above 50 mg/dL, the risk does not vary. Instead, for values below the threshold, the risk increases when HDL decreases [40]. We can see that AGE is the most important factor followed by HBA1C. Interestingly, the third most important factor is HDL, with an absolute value very close to the one of HBA1C. Immediately after HDL, we find SBP, followed by SEX and then LDL. These last two factors have comparable values of the parameters. At the end of the list there is SM, with value of the parameter one order of magnitude smaller than the others. However, we expect a large uncertainty in the estimates of β for this factor because of the small proportion of smokers in the sample population of the database analysed.

The results obtained from the real dataset, show a good ability of the BBN model in predicting the risk at the population level. In particular, looking at Fig. 3, we notice how the theoretical risk curve, obtained after model estimation by Eq. (15), fits well to the empirical one computed by KM estimator. From the same figure, we see that this is true also when the population risk is estimated by averaging the individual risk curves, although showing a slightly worse agreement.

The validation results in Table 2 show a good performance of the BBN model for prediction of the individual risk. The performances of individual risk estimation, as measured by the AUC and C-index from both the internal and 5-fold cross-validation, are about 0.75, for both real and simulated data with comparable sample sizes. As a comparison, when using the NORRISK2 risk calculator [13], the AUC for both internal and external validation is about 0.77 for men. Surprisingly, the AUC value obtained for women is about 0.83. The authors do not discuss the different values obtained for man and women. We expect that the higher values for the NORRISK2 when compared to the BBN ones, are mostly due to the large sample sizes in the former study. In fact, they are about one order of magnitude larger than the one used here. In addition, the NORRISK2 study is relative to the general population, while our is concerned with subjects with T1D. Finally, both studies consider CVD events but the specific endpoints (pathological conditions) are different. Another possible comparison can be made with the results obtained by Longato et al. [15], where a DL model is built to predict risk of cardiovascular complications in people

Table 4
Performance results of the complete and missing models for the simulated data with $m = 1600$ individuals. For both the internal and the 5-fold cross-validation, the evaluation is replicated 100 times. In the second and third columns the mean values of the AUC and C-index for the internal validation are reported. In the last two columns, the mean of the values of the 100 random replications of the 5-fold cross-validation for the same two indicators are displayed, together with their 95% quantile intervals in parenthesis.

Indicator	Internal validation		5-fold cross-validation	
	Complete model	Missing model	Complete model	Missing model
C-Index	0.75 (0.69–0.80)	0.71 (0.66–0.76)	0.72 (0.66–0.77)	0.69 (0.63–0.74)
AUC	0.75 (0.70–0.80)	0.72 (0.66–0.77)	0.73 (0.66–0.78)	0.70 (0.63–0.75)

with T1D from 1 to 5 years. The AUC and C-index values obtained for the 5-year prediction are equal to 0.79 and 0.77, respectively. The performances are comparable with NORRISK2 and slightly better than the ones obtained here. However, as already mentioned in the Introduction, the DL model used is trained based on a sample containing 214 676 individuals, that is much larger than the one used here (1293). Furthermore, the forecast time horizon used in the cited paper [15] is of 5 years, while the performance reported for the methodology proposed here are for forecast at 7 years.

We notice that the results of the performances obtained for the case of missing values show only a moderate loss with respect to those obtained for the complete one (see Table 2). Quantitatively the loss is in practice the same for both the internal and the 5-fold cross-validation. Focusing on the results for the conditional probabilities of the output, with respect to one single variable, we see that we can well describe it theoretically (see Fig. 4). We notice that, when conditioning on both SM and SEX together, we see a loss in ability of the theoretical model in describing the empirical one (see Fig. 4(c)). However, the results on simulated data suggest that this mainly depends on a limited sample size, as discussed later.

We turn now to the results of the simulation study, highlighting the unbiasedness and consistency of parameter estimators. Looking at Table 3, we notice that the mean estimated values for the $\boldsymbol{\beta}$ parameters are quite close to the true ones, also in the case of a simulated sample of the smallest size m. This shows that the estimation process is unbiased. The errors of both the parameters β and those of the baseline survival function decrease approximately as $\Theta(1/\sqrt{m})$ (see Fig. 5). This is true at level of single parameters and when the estimates of the $\boldsymbol{\beta}$ are cumulated with respect to the different factors. The highest percentage errors are for the β of SM. This could be explained by the fact that the proportion of smokers used here is quite low (∼15%). As said before, this is done to have a realistic dataset.

The results of the prediction performances obtained for the simulated data (see Table 4) are quite close to those for the real ones (see Table 2). This true for both the internal validation and the k-fold. However, the k-fold results present slightly worse values with respect to the real ones. These last results provide a better quantification of the performances. Indeed, in the real dataset, we perform the estimation only ones. Then, we take into account only the variation of the prediction ability due to different k-fold partition of the sample. Instead, for the simulation, we use several sets of original data and for each of them we perform parameter estimation. Therefore, in addition to the above variation we have also those induced by the data analysed. The performance loss when passing from complete model to the missing values case is almost the same for internal validation and 5-fold cross-validation.

We turn now to the results on the probability of the output (CVD), conditioned on a few factors. The discrepancy values between theoretical and empirical risk curve obtained with dimension $m = 1600$ individuals, which is comparable to the sample size of the real dataset, are similar to those relative to this last case. This is true both when conditioning on SM alone and when conditioning on SEX and SM together. Indeed, in the latter case, this discrepancy for female smokers decreases to 0.007 for the simulated dataset from 0.01 for the real one. Moreover, through this simulation, we can interpret the higher last value, when compared to those conditioning only on SM for the real dataset. In fact, increasing the sample size to $m = 50000$, the agreement improves significantly, for example, from 0.007 for female smokers to 0.002. Similarly, keeping the sample size at $m = 1600$, when the percentage of smokers is increased to 50%, the discrepancy decreases to 0.005.

From the overall results, we can see that the proposed BBN based method shows good ability in predicting CVD individual risks among the population with T1D. Furthermore, the method maintains good validation results also in case of missing values, with a relatively small loss in performance. This last allows a reliable risk estimation in a wider range of cases. The simulation study highlights the reliability of the estimation process and the ability of the method to properly retrieve the conditional probability of the output (CVD) given a few factors. Such information can be helpful when performing intervention.

The proposed methodology therefore represents a useful tool for individual CVD risk quantification, specifically for people with T1D. However, as the simulated results suggest, the method would benefit from the availability of a larger sample. In fact, this last improves the ability to correctly predict the event under study, which is expected to increase method performances, as measured, for example, by AUC. We could update the model by the planned acquisition of more data within the WARIFA project [39]. We notice that the parameter estimation has been performed on a dataset of relatively young subjects with T1D. In fact, age at the beginning of the observation period is ranging from 28 to 54 years and the mean is equal to 41.6 years. Therefore, some bias could be introduced when applying the methodology to subjects with T1D outside this range.

We remark that, unlike the CPH model, which is the basis of existing RCs, BBN models require a considerable implementation effort. Furthermore, some of the tasks, e.g. estimation of population risk, are often computationally intensive. In order to reduce computing time, continuous variables have been here discretised. However, our discretisation performed here seems to be adequate. In fact, using a higher number of factor levels does not change the results in practice. Finally, when compared to CPH model, BBN has an additional set of parameters to be estimated. For the BBN applied here, these parameters appear in three marginal models and four conditional ones.

Our methodology has been derived and applied to quantify the risk for some specific adverse events (CVD) in a given population (individuals with T1D). This is due to the fact that this research has been developed within the EU Horizon 2020 WARIFA project [39], which deals with this case. However, the methodology can be adapted to many other situations. For example, in a straightforward way it can be used to predict risk of CVD in the general population. This requires to reduce the DAG by deleting nodes and arrows specific to T1D. In addition, a suitable database for general population is needed to estimate model parameters. More in general, given the good behaviour of this approach, it could be developed and applied to other different pathological conditions, e.g. melanoma occurrence. Indeed, we have recently started to work on this pathology.

We notice that, despite the name of the method, the model developed here is not Bayesian. In fact, we do not use any a priori model for the model parameters. The historical reasons for this name can be found for example in Pearl [37,46]. However, a natural extension of this work includes the adoption of a real Bayesian approach. In this case, one possibility would be to use a multivariate non-informative prior (i.e. Dirichlet distribution) for the parameter vector.

6. Conclusions

The proposed BBN based methodology reveals a good ability in quantifying individual risk of CVD events in people with T1D. The performances remain good also when the values of some factors are missing. Empirical evidence of the estimator consistency has been provided through a simulation study. Results from both real and simulated data show that the method properly retrieves the conditional probability of the output (CVD) given a few factors. Such information can be helpful when performing intervention. The proposed methodology therefore represents a useful tool for computing the risk of CVD events in individuals with T1D. However, it can be extended with minor modifications to the general population, provided that a suitable database is available to perform parameter estimation. More in general, given the good behaviour of this approach, it could be developed and applied to other different pathological conditions, e.g. melanoma occurrence. Indeed, we have recently started to work on this pathology.

Acronyms not defined in the text

PA	Physical Activity
DRK	Drinking habits
TRT	Treatments
TRG	Triglyceride
BMI	Body-mass index
DBP	Diastolic blood pressure
FMH	Family History of CVD
DIA DUR	Diabetes duration
EDULV	Educational level
ETN	Ethnicity

CRediT authorship contribution statement

Ornella Moro: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Formal analysis, Data curation. **Inger Torhild Gram:** Writing – review & editing, Methodology, Conceptualization. **Maja-Lisa Løchen:** Writing – review & editing, Methodology, Conceptualization. **Marit B. Veierød:** Writing – review & editing, Methodology, Conceptualization. **Ana Maria Wägner:** Writing – review & editing, Methodology, Conceptualization. **Giovanni Sebastiani:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Formal analysis, Data curation, Conceptualization.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017385 (Art. 29.1 GA) [39].

Data acknowledgements

The Diabetes Control and Complications Trial (DCCT) and its follow-up the Epidemiology of Diabetes Interventions and Complications (EDIC) study were conducted by the DCCT/EDIC Research Group and supported by National Institute of Health grants and contracts and by the General Clinical Research Center Program, NCRR. The data from the DCCT/EDIC study were supplied by NIDDK Central Repository. This manuscript was not prepared under the auspices of the DCCT/EDIC study and does not represent analyses or conclusions of the DCCT/EDIC study group, NIDDK Central Repository, or NIH.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors are thankful to the anonymous reviewers for their useful comments and suggestions.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compbiomed.2025.109967.

References

[1] Richard I.G. Holt, et al., The management of type 1 diabetes in adults. A consensus report by the American diabetes association (ADA) and the European association for the study of diabetes (EASD), Diabetes Care 44 (11) (2021) 2589–2625, http://dx.doi.org/10.2337/dci21-0043.

[2] Muthiah Vaduganathan, et al., The global burden of cardiovascular diseases and risk: A compass for future health, J. Am. Coll. Cardiol. (ISSN: 0735-1097) 80 (25) (2022) 2361–2371, http://dx.doi.org/10.1016/j.jacc.2022.11.005.

[3] Gregory A. Roth, et al., Global burden of cardiovascular diseases and risk factors, 1990–2019: Update from the GBD 2019 study, J. Am. Coll. Cardiol. (ISSN: 0735-1097) 76 (25) (2020) 2982–3021, http://dx.doi.org/10.1016/j.jacc.2020.11.010.

[4] Dorte Vistisen, et al., Prediction of first cardiovascular disease event in type 1 diabetes mellitus, Circulation 133 (11) (2016) 1058–1066, http://dx.doi.org/10.1161/CIRCULATIONAHA.115.018844.

[5] Araz Rawshani, et al., Excess mortality and cardiovascular disease in young adults with type 1 diabetes in relation to age at onset: a nationwide, register-based cohort study, Lancet 392 (10146) (2018) 477–486.

[6] Frank L.J. Visseren, et al., 2021 ESC guidelines on cardiovascular disease prevention in clinical practice: Developed by the Task Force for cardiovascular disease prevention in clinical practice with representatives of the European society of cardiology and 12 medical societies with the special contribution of the European Association of Preventive Cardiology (EAPC), Eur. Heart J. (ISSN: 0195-668X) 42 (34) (2021) 3227–3337, http://dx.doi.org/10.1093/eurheartj/ehab484.

[7] Cristina Colom, et al., Cardiovascular disease in type 1 diabetes mellitus: Epidemiology and management of cardiovascular risk, J. Clin. Med. (ISSN: 2077-0383) 10 (8) (2021) http://dx.doi.org/10.3390/jcm10081798.

[8] Ik H. Teoh, Panchami Elisaus, Jonathan.D. Schofield, Cardiovascular risk management in type 1 diabetes., Curr. Diabetes Rep. 21 (9) (2021) 29, http://dx.doi.org/10.1007/s11892-021-01400-9.

[9] Bruno Vergès, Cardiovascular disease in type 1 diabetes, an underestimated danger: Epidemiological and pathophysiological data, Atherosclerosis 394 (2024) 117158, http://dx.doi.org/10.1016/j.atherosclerosis.2023.06.005.

[10] Diederik De Cock, et al., The effect of physical activity on glycaemic control in people with type 1 diabetes mellitus: A systematic literature review and meta-analysis, Diabetic Med. 41 (10) (2024) e15415, http://dx.doi.org/10.1111/dme.15415.

[11] Giovane R. Sousa, et al., Glycemic control, cardiac autoimmunity, and long-term risk of cardiovascular disease in type 1 diabetes mellitus, Circulation 139 (6) (2019) 730–743, http://dx.doi.org/10.1161/CIRCULATIONAHA.118.036068.

[12] Jonathan Schofield, Jan Ho, Handrean Soran, Cardiovascular risk in type 1 diabetes mellitus, Diabetes Ther. 10 (3) (2019) 773–789, http://dx.doi.org/10.1007/s13300-019-0612-8.

[13] Randi Selmer, et al., NORRISK 2: A norwegian risk model for acute cerebral stroke and myocardial infarction, Eur. J. Prev. Cardiol. (ISSN: 2047-4873) 24 (7) (2020) 773–782, http://dx.doi.org/10.1177/2047487317693949.

[14] David R. Cox, David Oakes, Analysis of Survival Data, Chapman and Hall/CRC, 2018.

[15] Enrico Longato, et al., A deep learning approach to predict diabetes' cardiovascular complications from administrative claims, IEEE J. Biomed. Heal. Inform. 25 (9) (2021) 3608–3617, http://dx.doi.org/10.1109/JBHI.2021.3065756.

[16] Flávio L. Seixas, et al., A Bayesian network decision model for supporting the diagnosis of dementia, alzheimer's disease and mild cognitive impairment, Comput. Biol. Med. 51 (2014) 140–158, http://dx.doi.org/10.1016/j.compbiomed.2014.04.010.

[17] Joseph W. Zabinski, Kelsey J Pieper, Jacqueline MacDonald Gibson, A Bayesian belief network model assessing the risk to wastewater workers of contracting ebola virus disease during an outbreak, Risk Anal. 38 (2) (2018) 376–391, http://dx.doi.org/10.1111/risa.12827.

[18] Evangelia Kyrimi, et al., A comprehensive scoping review of Bayesian networks in healthcare: Past, present and future, Artif. Intell. Med. (ISSN: 0933-3657) 117 (2021) 102108, http://dx.doi.org/10.1016/j.artmed.2021.102108.

[19] Biche Osong, et al., Bayesian network structure for predicting local tumor recurrence in rectal cancer patients treated with neoadjuvant chemoradiation followed by surgery, Phys. Imaging Radiat. Oncol. 22 (2022) 1–7, http://dx.doi.org/10.1016/j.phro.2022.03.002.

[20] Pilar Fuster-Parra, et al., Identifying risk factors of developing type 2 diabetes from an adult population with initial prediabetes using a Bayesian network, Front. Public Heal. (ISSN: 2296-2565) 10 (2023) http://dx.doi.org/10.3389/fpubh.2022.1035025.

[21] Aisha Hikal, et al., A treatment decision support model for laryngeal cancer based on Bayesian networks, Biomedicines (ISSN: 2227-9059) 11 (1) (2023) http://dx.doi.org/10.3390/biomedicines11010110.

[22] Alexander Stojadinovic, et al., Development of a Bayesian belief network model for personalized prognostic risk assessment in colon carcinomatosis, Am. Surg.™ 77 (2) (2011) 221–230, http://dx.doi.org/10.1177/000313481107700225.

[23] Mervyn Stone, Cross-validatory choice and assessment of statistical predictions, J. R. Stat. Soc. Ser. B Stat. Methodol. (ISSN: 0035-9246) 36 (2) (2018) 111–133, http://dx.doi.org/10.1111/j.2517-6161.1974.tb00994.x.

[24] Seymour Geisser, The predictive sample reuse method with applications, J. Amer. Statist. Assoc. (350) (1975) 320–328, http://dx.doi.org/10.1080/01621459.1975.10479865.

[25] James A. Hanley, Barbara J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve., Radiology 143 (1) (1982) 29–36, http://dx.doi.org/10.1148/radiology.143.1.7063747.

[26] Kalia Orphanou, Athena Stassopoulou, Elpida Keravnou, DBN-extended: A dynamic Bayesian network model extended with temporal abstractions for coronary heart disease prognosis, IEEE J. Biomed. Heal. Inform. 20 (3) (2016) 944–952, http://dx.doi.org/10.1109/JBHI.2015.2420534.

[27] Jose M. Ordovas, et al., A Bayesian network model for predicting cardiovascular risk, Comput. Methods Programs Biomed. (ISSN: 0169-2607) 231 (2023) 107405, http://dx.doi.org/10.1016/j.cmpb.2023.107405.

[28] Olivier Pourret, Patrick Na, Bruce Marcot, Bayesian Networks: a Practical Guide to Applications, John Wiley & Sons, 2008.

[29] Leila Yousefi, et al., Predicting comorbidities using resampling and dynamic Bayesian networks with latent variables, in: 2017 IEEE 30th International Symposium on Computer-Based Medical Systems, CBMS, 2017, pp. 205–206, http://dx.doi.org/10.1109/CBMS.2017.32.

[30] Leila Yousefi, et al., Predicting disease complications using a stepwise hidden variable approach for learning dynamic Bayesian networks, in: 2018 IEEE 31st International Symposium on Computer-Based Medical Systems, CBMS, 2018, pp. 106–111, http://dx.doi.org/10.1109/CBMS.2018.00026.

[31] Yuval Shahar, Mark A. Musen, RÉSUMÉ: A temporal-abstraction system for patient monitoring, Comput. Biomed. Res. (ISSN: 0010-4809) 26 (3) (1993) 255–273, http://dx.doi.org/10.1006/cbmr.1993.1018.

[32] Jie Cheng, David A. Bell, Weiru Liu, An algorithm for Bayesian network construction from data, in: David Madigan, Padhraic Smyth (Eds.), Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics, in: Proceedings of Machine Learning Research, vol. R1, PMLR, 1997, pp. 83–90, URL https://proceedings.mlr.press/r1/cheng97a.html.

[33] Diabetes control and complications trial / epidemiology of diabetes interventions and complications (dcct/edic) (version 21) [dataset] niddk central repository, 2021, http://dx.doi.org/10.58020/mn8k-ms49.

[34] Edward L. Kaplan, Paul Meier, Nonparametric estimation from incomplete observations, J. Amer. Statist. Assoc. 53 (282) (1958) 457–481, http://dx.doi.org/10.1080/01621459.1958.10501452.

[35] Enrico Colosimo, Flavio Ferreira, Maristela Oliveira, Cleide Sousa, Empirical comparisons between Kaplan-Meier and nelson-aalen survival function estimators, J. Stat. Comput. Simul. 72 (4) (2002) 299–308, http://dx.doi.org/10.1080/00949650212847.

[36] Paul Meier, Theodore Karrison, Rick Chappell, Hui Xie, The price of Kaplan–Meier, J. Amer. Statist. Assoc. 99 (467) (2004) 890–896, http://dx.doi.org/10.1198/016214504000001259.

[37] Judea Pearl, Causality: Models, Reasoning, and Inference, Cambridge University Press, New York, 2009.

[38] Judea Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan kaufmann, 1988.

[39] Watching the risk factors: Artificial Intelligence (AI) and the prevention of chronic conditions, https://www.warifa.eu/ [Accessed 02 April 2024].

[40] Mohammad Al Zein, et al., Revisiting high-density lipoprotein cholesterol in cardiovascular disease: Is too much of a good thing always a good thing? Prog. Cardiovasc. Dis. (ISSN: 0033-0620) 87 (2024) 50–59, http://dx.doi.org/10.1016/j.pcad.2024.10.009, Assorted Topics III.

[41] Marcel Wolbers, et al., Prognostic models with competing risks: Methods and application to coronary risk prediction, Epidemiology 20 (4) (2009) 555–561, http://dx.doi.org/10.1097/EDE.0b013e3181a39056.

[42] Michael W. Browne, Cross-validation methods, J. Math. Psych. 44 (1) (2000) 108–132, http://dx.doi.org/10.1006/jmps.1999.1279.

[43] Diabetes Control, Complications Trial (DCCT) Research Group, Design and methodologic considerations for the feasibility phase. The DCCT research group, Diabetes 35 (5) (1986) 530–545.

[44] Epidemiology of diabetes interventions and complications (EDIC). Design, implementation, and preliminary results of a long-term follow-up of the diabetes control and complications trial cohort, Diabetes Care (ISSN: 0149-5992) 22 (1) (1999) 99–111, http://dx.doi.org/10.2337/diacare.22.1.99.

[45] The Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications Research Group, Retinopathy and nephropathy in patients with type 1 diabetes four years after a trial of intensive therapy, N. Engl. J. Med. 342 (6) (2000) 381–389, http://dx.doi.org/10.1056/NEJM200002103420603.

[46] Judea Pearl, Bayesian networks: A model of self-activated memory for evidential reasoning, in: Proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine, CA, USA, 1985, pp. 15–17.