BMC Public Health

RESEARCH

Open Access

Causal analysis of traditional and environmental risk factors for long-term development of type 2 diabetes using a conditional survival Bayesian network: evidence from the Korean Genome and Epidemiology Study

Ji Young Choi[1] and Man-Suk Oh[1*]

Abstract

Background Over the past decade, traditional demographic, lifestyle, and metabolic factors, along with air pollutants, have increasingly been recognized as key contributors to type 2 diabetes (T2D). However, the comprehensive causal structure among these factors and their individual and interacting interventional effects have seldom been characterized in long-term population studies.

Methods Using 11-year follow-up data from 2,102 adults without T2D in the Ansan cohort of the Korean Genome and Epidemiology Study (KoGES), we investigated causal pathways among demographic, lifestyle, metabolic factors, and multiple ambient air pollutants leading to long-term T2D incidence. We employed a Conditional Survival Bayesian Network (CSBN), which integrates survival analysis with Bayesian network modeling to accommodate censored and incomplete data, to visualize the causal structure among risk factors, and to estimate both individual and joint (interaction) interventional effects.

Results The CSBN depicted a holistic causal structure showing how multiple risk factors jointly shape T2D development over the 11-year follow-up and helped distinguish putative direct/indirect pathways from associations likely reflecting confounding. Interventional analysis quantified each factor's causal contribution to the 11-year T2D incidence. Obesity produced the largest individual effect: setting BMI to the obese category approximately doubled 11-year T2D risk compared with normal weight. High alanine aminotransferase (ALT) and older age increased risk by about 40–50%, while family history of T2D, dyslipidemia, overweight, $PM_{2.5}$, and gaseous pollutants had intermediate effects. Furthermore, the CSBN uncovered synergistic interactions mainly among metabolic factors. In particular, ALT

*Correspondence:
Man-Suk Oh
msoh@ewha.ac.kr

Full list of author information is available at the end of the article

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

with family history, dyslipidemia, or obesity displayed strong additive interactions. By contrast, air pollutants were found to influence T2D independently rather than through interactions with other risk factors.

Conclusion These findings underscore the importance of integrated public health strategies targeting multiple risk factors to effectively curb T2D incidence. The CSBN's capability to explicitly model complex causal interactions highlights the necessity for advanced epidemiological analyses to inform targeted preventive measures and efficient resource allocation.

Keywords Type 2 diabetes, Air pollution, Bayesian network, Survival analysis, Conditional survival Bayesian network, Korean Genome and Epidemiology Study, Causal inference

Introduction

Type 2 diabetes (T2D) represents a critical global health challenge characterized by insulin resistance and impaired glucose regulation, contributing significantly to severe health complications such as cardiovascular disease, kidney failure, diabetic retinopathy, and limb amputations [1–3]. Its prevalence has risen particularly in rapidly industrializing countries, including South Korea, driven by economic development, dietary shifts towards Western patterns, and increasingly sedentary lifestyles [4, 5]. Recent data indicate that approximately 15.5% of Korean adults aged 30 and older have T2D, highlighting an urgent need for comprehensive preventative strategies [6].

Although traditional risk factors for T2D, including obesity, dyslipidemia, hypertension, and advancing age, are well established [7, 8], emerging evidence emphasizes ambient air pollution as a significant environmental risk factor [9–11]. Epidemiological studies have demonstrated associations between exposure to particulate matter ($PM_{2.5}$) and gaseous pollutants such as carbon monoxide (CO), nitrogen dioxide (NO_2), and sulfur dioxide (SO_2) and increased T2D risk [12–14].

However, prior research has largely focused on single pollutants or simple additive models and often overlooked the interactive effects of pollutants combined with traditional risk factors [9, 15, 16]. To address these limitations, we employed a Conditional Survival Bayesian Network (CSBN) to comprehensively evaluate the interplay of air pollutants and traditional risk factors on T2D development over an 11-year follow-up period, utilizing data from the Ansan cohort of the Korean Genome and Epidemiology Study (KoGES) [17].

The CSBN method integrates Bayesian network (BN) modeling and survival analysis, enabling the simultaneous assessment of direct, indirect, and confounded effects among multiple covariates while robustly managing censored and incomplete data [18, 19]. Its visual probabilistic structure transparently depicts structural dependencies among risk factors, facilitating intuitive understanding of complex causal mechanisms [20]. Additionally, the CSBN method flexibly accommodates non-linear relationships and interactions, surpassing the limitations of traditional survival models [19].

The study demonstrates that, in addition to traditional risk factors—age, obesity, regular exercise, smoking, family history of T2D, elevated alanine aminotransferase (ALT), and dyslipidemia—both $PM_{2.5}$ and gaseous pollutants are important determinants of long-term T2D development. Furthermore, it identifies interactions among these factors that significantly amplify T2D risk beyond additive effects, underscoring the need for integrated public health interventions.

Data

KoGES data

Study population and data sources

This study utilized data from the Ansan cohort of the KoGES, a nationwide prospective cohort designed to identify genetic and environmental risk factors for chronic diseases commonly observed in Koreans, such as diabetes, hypertension, and hyperlipidemia. Data collection included lifestyle surveys, clinical examinations, and biospecimen analysis under informed consent [17, 21]. The Ansan cohort of the KoGES study began in 2001 with approximately 5,000 participants aged 40–69 years and has been followed up biennially. In this study, we used data from the 4th survey (2009–2010) as baseline and followed participants through the 9th survey (2019–2020), focusing on an 11-year observation period.

Ansan, recognized as South Korea's largest cluster of small and medium-sized enterprises and a major industrial complex in the Seoul metropolitan area, exhibits relatively high concentrations of gaseous pollutants and particulate matter compared to other regions in the country [22]. This provides researchers with a unique opportunity to investigate the health impacts associated with complex air pollution, closely reflecting real-world exposure scenarios typical of rapidly industrializing and developing regions.

Outcome and participant selection

Incident T2D was defined according to the diagnostic criteria of the American Diabetes Association. A diagnosis of T2D was made if any one of the following criteria

was met: (1) fasting plasma glucose ≥ 126 mg/dL (7.0 mmol/L) after at least 8 hours of no caloric intake; (2) 2-hour plasma glucose ≥ 200 mg/dL (11.1 mmol/L) during a 75-g oral glucose tolerance test (OGTT); or (3) HbA1c $\geq 6.5\%$ [23].

Among 2,903 individuals with information at the 4th follow-up survey (which served as the baseline for this study), we excluded those who did not participate in any subsequent follow-up survey, those with a history of diabetes before baseline, and those with T2D at baseline. We further excluded participants with missing T2D status at the final follow-up (incomplete follow-up). Finally, we excluded one participant with missing ALT values, given that ALT is an established risk factor for T2D. After these exclusions, the final analytic sample comprised 2,102 participants.

Variables and preprocessing

We selected variables from the raw dataset that were identified as having clinical or biological relevance to T2D pathogenesis based on a comprehensive literature review [8, 24–27]. We then excluded variables with an absolute pairwise correlation coefficient exceeding 0.7 with any other selected variable. After this data-cleaning process, 13 variables—age (AGE), sex (SEX), family history of diabetes (FDM), monthly household income (INCOME), partnership status (PARTNER), smoking status (SMOKE), alcohol consumption (DRINK), regular physical activity (EXER), body mass index (BMI), hypertension (HTN), dyslipidemia (DYSLIP), aspartate aminotransferase (AST), ALT—were retained as candidate risk factors for T2D. These variables encompass a broad spectrum of domains, including demographic and socioeconomic characteristics, metabolic parameters, and lifestyle factors.

Continuous variables among these candidate predictors were discretized prior to BN learning to improve computational simplicity, model stability, and interpretability. BMI, a standardized screening indicator for obesity calculated as weight(kg)/[height(m)]^2, was categorized according to WHO criteria [28]. Hypertension status (HTN) was classified based on [29] using blood pressure (bp) measurements in the higher arm during the second seated measurement. Dyslipidemia status (DYSLIP) thresholds were determined by 2022 Korean Guidelines for the Management of Dyslipidemia criteria [30]. ALT and AST thresholds were based on CLSI Guideline C28-A3c [31].

Air pollution data

Exposure estimation

To address growing concerns about air pollution's health effects, Korea Disease Control and Prevention Agency (KDCA) has developed a method for individualized air pollution exposure assessment for participants in KoGES [17, 21]. This approach estimates nationwide air pollution levels using the Community Multiscale Air Quality (CMAQ) model, incorporating meteorological data, emission inventories, satellite-derived aerosol optical depth, and the normalized difference vegetation index to enhance accuracy and spatial resolution [32–34]. The modeled concentrations of $PM_{2.5}$, PM_{10}, CO, NO_2, SO_2, and O_3 are then linked to KoGES participants' geocoded residential addresses, assigning exposure levels using 1 km grids for particulate matter and 9 km grids for gaseous pollutants. When a participant's address differed between two consecutive examinations, the KDCA assumed that the move occurred at the midpoint of the interval between the two surveys and allocated yearly concentrations accordingly.

To capture chronic environmental burden, we used these data to construct five-year average concentrations for each pollutant in the period preceding the study baseline (2009–2010). This long-term exposure metric is intended to reflect the cumulative impact of ambient air pollution on metabolic and inflammatory pathways implicated in the development of T2D.

Data processing

Among the six air pollutants, PM_{10} was highly correlated with $PM_{2.5}$ (Pearson r = 0.81), and O_3 was strongly and inversely correlated with NO_2 (r = -0.89). To mitigate multicollinearity in the survival models and to obtain a parsimonious and interpretable BN structure, we excluded PM_{10} and O_3 from the analysis.

Because CO, SO_2, and NO_2 predominantly arise from common combustion-related sources (traffic, residential heating, and industrial emissions), they are highly correlated in our data. In a causal BN framework, arcs among these pollutants are therefore difficult to interpret, because their dependencies largely reflect shared sources rather than meaningful causal pathways between the pollutants themselves. Additionally, in survival models that included both CO and NO_2, which were highly correlated (r=0.71), only CO showed a statistically significant association with incident T2D, whereas NO_2 did not. To address these issues, we constructed a composite gas index (GASmix), defined as the average of the three gaseous pollutants after standardizing each to have a mean 0 and a standard deviation 1. This approach captures the overall burden of combustion-related gaseous pollution, rather than attempting to disentangle correlated individual effects with limited interpretability, and yields a more parsimonious and interpretable BN by replacing three correlated exposure variables with a single mixture node.

The two pollutant variables $PM_{2.5}$ (variable name PM2.5) and GASmix were discretized for BN learning. The WHO 2021 Air Quality Guidelines recommend an

annual limit of 5 $\mu g/m^3$ for $PM_{2.5}$. In our cohort, however, the five-year average $PM_{2.5}$ exposures substantially exceeded this guideline value. No corresponding guideline is currently available for GASmix. In this study, we therefore dichotomized $PM_{2.5}$ and GASmix at their 75th percentiles to examine health effects associated with very high pollution exposure.

Final analytic data

The KoGES data and air pollution data were merged by participants, resulting in 15 covariates (13 covariates from KoGES + 2 pollutants). Figure 1 displays the flow diagram of participant selection and data processing, along with construction of the CSBN. Table 1 presents a detailed overview of all study variables, including their names, descriptions, thresholds, frequency distributions, and statistical significance (p-values from chi-square tests) in relation to incident T2D over the 11-year follow-up.

To evaluate potential selection and censoring bias, we compared baseline characteristics between participants included in the final analysis and those excluded because of incomplete follow-up (Table S1). Participants with incomplete follow-up (n = 282, 12%) were generally similar to those with complete follow-up (n = 2,102, 88%) in terms of BMI, hypertension, dyslipidemia, liver enzymes, physical activity, and air pollution exposures. However, they were more likely to be in the highest age category, to be female, to have lower household income, to be non-smokers and non-drinkers, and to live without a partner. Overall, these patterns suggest that although some sociodemographic differences were observed, key metabolic profiles and air pollution exposures were comparable, reducing—though not eliminating—the likelihood of substantial selection or censoring bias.

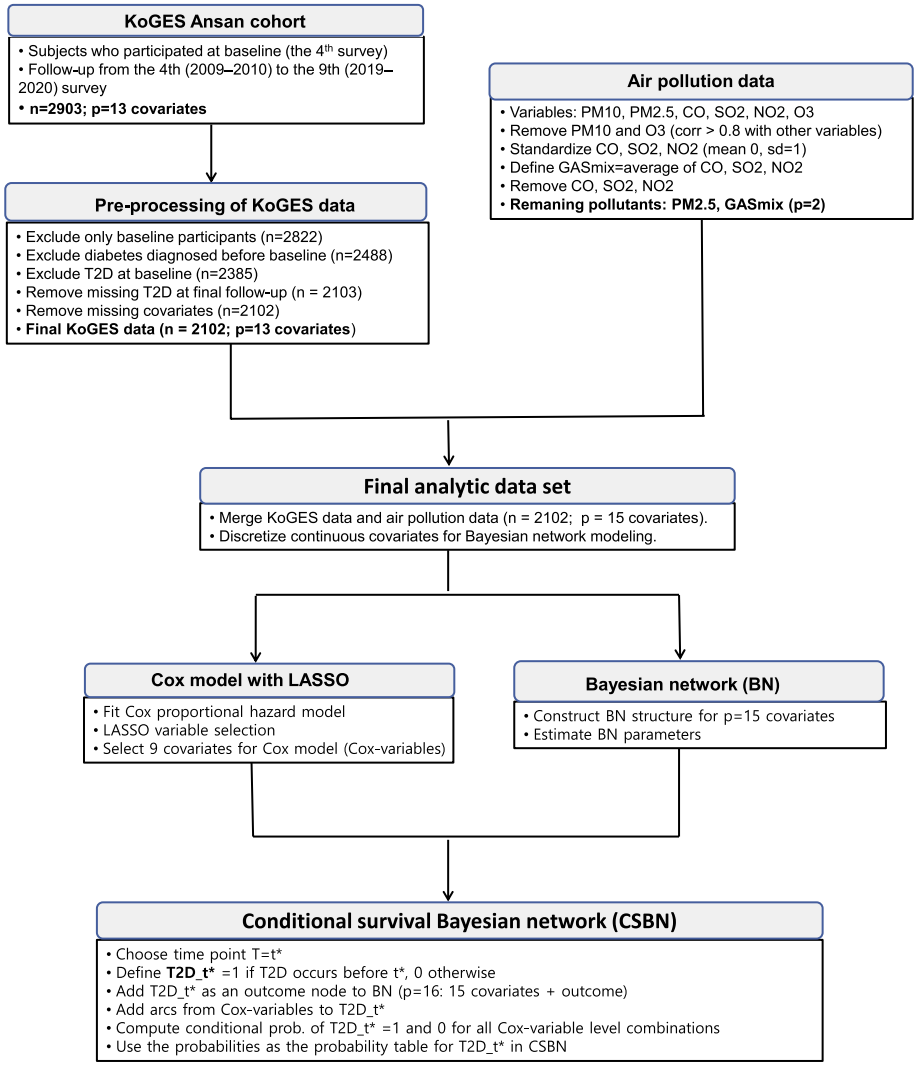


Fig. 1 Study flow diagram illustrating participant selection, data processing, and CSBN development

Table 1 Descriptions and summary statistics of the variables at baseline

Variable		Category		Total	No incident T2D	Incident T2D	χ^2 test
Name	**Description**	**Level**	**Threshold**	**N**	**N(%)**	**N(%)**	***p*-value**
AGE	Age at baseline (years)	Low	40–52	734	608 (82.8)	126 (17.2)	0.0029
		Mid	53–56	690	551 (79.9)	139 (20.1)	
		High	≥57	678	512 (75.5)	166 (24.5)	
SEX	Sex	Male		1081	861 (79.6)	220 (20.4)	0.9010
		Female		1021	810 (79.3)	211 (20.7)	
FDM	Diabetes diagnosis among first-degree relatives	No	Absent	1997	1594 (79.8)	403 (20.2)	0.1387
		Yes	Present	105	77 (73.3)	28 (26.7)	
INCOME	Monthly household income (million KRW)	Low	<1.5	333	255 (76.6)	78 (23.4)	0.2480
		Mid	1.5–4	1052	835 (79.4)	217 (20.6)	
		High	≥4	717	581 (81.0)	136 (19.0)	
PARTNER	Partnership status	Yes	Ever married or Cohabiting	1931	1537 (79.6)	394 (20.4)	0.7763
		No	Otherwise	171	134 (78.4)	37 (21.6)	
SMOKE	Lifetime consumption of ≥ 400 cigarettes	No	Non-smoker	1252	999 (79.8)	253 (20.2)	0.7364
		Former	Former smoker	548	437 (79.7)	111 (20.3)	
		Current	Current smoker	302	235 (77.8)	67 (22.2)	
DRINK	Alcohol consumption	No	Non-drinker	1017	808 (79.4)	209 (20.6)	0.7947
		Former	Former drinker	45	34 (75.6)	11 (24.4)	
		Current	Current drinker	1040	829 (79.7)	211 (20.3)	
EXER	Exercise	Yes	Regular vigorous physical activity	1033	835 (80.8)	198 (19.2)	0.1504
		No	Otherwise	1069	836 (78.2)	233 (21.8)	
BMI	Body mass index (kg/m^2)	Normal	<25	1223	1011 (82.7)	212 (17.3)	< 0.0001
		Over	25≤ <30	810	617 (76.2)	193 (23.8)	
		Obese	≥30	69	43 (62.3)	26 (37.7)	
HTN	Hypertension (mmHg)	Yes	Diagnosis, Systolic bp≥140, or Diastolic bp ≥90	332	255 (76.8)	77 (23.2)	0.2120
		No	Otherwise	1770	1416 (80.0)	354 (20.0)	
DYSLIP	Dyslipidemia (mg/dL)	Yes	Diagnosis, Total chol ≥240, TG ≥200, or HDL ≤40	1120	843 (75.3)	277 (24.7)	< 0.0001
		No	Otherwise	982	828 (84.3)	154 (15.7)	
AST	Aspartate aminotransferase (U/L)	Normal	AST ≤ 40	2009	1605 (79.9)	404 (20.1)	0.0509
		High	AST > 40	93	66 (71.0)	27 (29.0)	
ALT	Alanine aminotransferase (U/L)	Normal	ALT ≤ 40	1943	1566 (80.6)	377 (19.4)	< 0.0001
		High	ALT > 40	159	105 (66.0)	54 (34.0)	
PM2.5	Fine particulate matter $\leq 2.5\mu m$ ($\mu g/m^3$)	Low	$< 28.4\,(Q3)$	1577	1269 (80.5)	308 (19.5)	0.0638
		High	$\geq 28.4\,(Q3)$	525	402 (76.6)	123 (23.4)	
GASmix	Averaged z-score of CO, SO_2, NO_2	Low	$< 0.0760(Q3)$	1623	1297 (79.9)	326 (20.1)	0.4183
		High	$\geq 0.0760(Q3)$	479	374 (78.1)	105 (21.9)	

Methods

Basic Bayesian network

In the CSBN framework, it is necessary to perform structure learning for a basic BN to integrate it effectively with a survival model. Prior to conducting structure learning, we specified a blacklist to constrain possible relationships among variables. The blacklist is defined as a set of arcs representing relationships known to be impossible, which must be excluded from the learning model. By setting a blacklist, the number of candidate DAGs (Directed Acyclic Graphs) to explore during the learning process can be reduced, and clinically implausible relationships are prevented from being included as arcs in the BN [20].

Table 2 Variable layers used to impose blacklist in a basic Bayesian network

Layer	Category	Variables
Layer 1-1	Demographics	SEX, AGE, FDM
Layer 1–2	Air pollutants	PM2.5, GASmix
Layer 2	Lifestyle factors	INCOME, PARTNER, SMOKE, DRINK, EXER
Layer 3	Metabolic health	BMI, HTN, DYSLIP, AST, ALT

The blacklist used in this study followed the layered architecture summarized in Table 2. Variables in lower-numbered layers were allowed to influence only variables in higher-numbered layers, thereby ruling out reverse-causal paths. Additional constraints were imposed as follows: (1) no arcs were permitted between demographic

variables (Layer 1-1) and air pollutants (Layer 1–2), treating them as parallel root causes; (2) demographic variables in Layer 1-1 (SEX, AGE, FDM) were not allowed to influence one another, preserving them as mutually independent baseline characteristics; (3) air pollutants (Layer 1–2) were restricted from influencing lifestyle factors (Layer 2), although they may affect metabolic factors (Layer 3); and (4) within the metabolic health layer, BMI was protected from incoming edges from other metabolic variables, reflecting its role as an upstream determinant of hypertension, dyslipidemia, and liver enzyme abnormalities rather than their consequence.

The resulting hierarchical structure imposed by these blacklist constraints is illustrated in Fig. 2. With this layered blacklist, most of the causal structure was specified a priori, based on temporal ordering and clinical knowledge [35–38], and the data-driven learning component played a minor role.

To learn the structure of the basic BN, the score-based Tabu Search algorithm [20] was used. This algorithm starts with a random structure and iteratively modifies the current structure, evaluates the model's performance, and identifies the optimal structure. The Bayesian Information Criterion (BIC) was used as the model evaluation score [39, 40].

A single dataset offers limited evidence for the stability of relationships between nodes in a BN, so we employed bootstrap aggregation (bagging) to bolster confidence [41]. From the dataset, we created N new datasets (called bootstrap samples) of the same size as the original dataset by randomly sampling with replacement. For each bootstrap sample, the basic BN structure was constructed. From the resulting N basic BNs, arc strength and direction were computed for each arc. The arc strength is the relative frequency of the arc's presence across the networks, regardless of direction. The arc direction is the relative frequency of the arc's specific direction across the networks. Both metrics range from 0 to 1, with values close to 1 indicating strong support for the presence or direction of the arc [41].

Using model averaging, a consensus network was constructed by including only arcs with arc strengths exceeding a specified threshold to eliminate weak or spurious connections and retain only robust relationships. Higher thresholds correspond to sparser networks.

Survival analysis

We used the Cox proportional hazards (CPH) model for survival analysis in this study, although other survival models could also be applied. Cox [42] first proposed CPH model, which has become one of the most widely used methods in survival analysis. The model is semi-parametric: it does not specify a particular functional form for the baseline hazard over time, but assumes a log-linear relationship between covariates and the hazard. Thus, it combines a nonparametric component (the unspecified baseline hazard) with a parametric component (the regression coefficients).

Let T denote survival time and $X = (X_1, X_2, \ldots, X_p)'$ the covariate vector. In the CPH model, the hazard function at time t given covariates X is

$$h(t \mid X) = h_0(t) \exp(X'\beta),$$

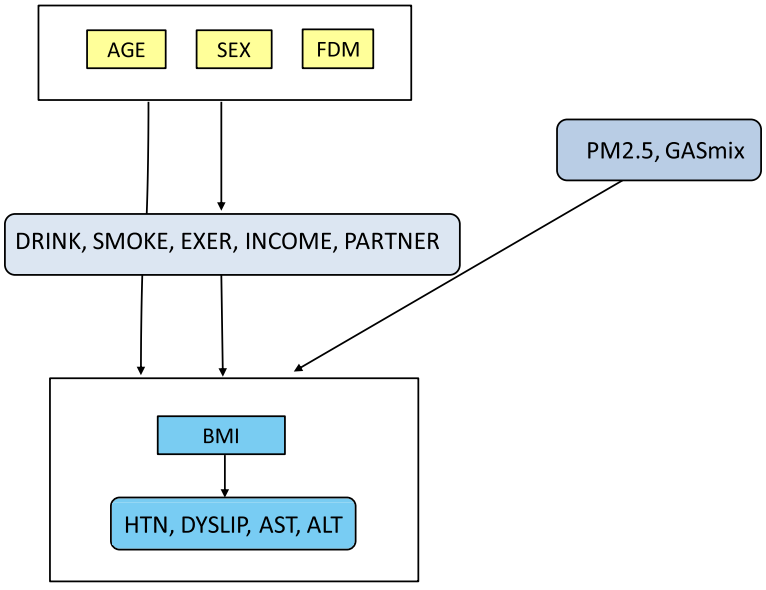


Fig. 2 Allowed arcs in the baseline Bayesian network structure under the specified blacklist. Arcs among variables within each rounded box are also permitted

where $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ is the vector of regression coefficients and $h_0(t)$ is the baseline hazard function, representing the hazard at time t when all covariates are zero. The regression coefficients β are estimated by maximizing the partial likelihood [43].

Let $\widehat{H}_0(t)$ denote the estimated baseline cumulative hazard function. For an individual with covariate vector x, the estimated survival probability at time t is

$$\widehat{S}(t \mid \mathrm{x}) = \exp\left\{-\widehat{H}_0(t)\exp(\mathrm{x}'\widehat{\beta})\right\},$$

which represents the estimated probability that an individual with covariates x remains free of the event up to time t.

To enhance the survival model's applicability, we applied a variable selection to systematically identify the most influential predictors, achieving a parsimonious model while preserving predictive performance [44].

Conditional survival Bayesian network

BNs serve as versatile tools in medical decision-making due to their non-parametric learning capabilities [45]. Their key advantage lies in learning probabilistic relationships directly from observed data patterns, without requiring pre-specified probability distributions or functional forms that characterize traditional statistical methods. This flexibility allows BNs to capture complex, real-world medical relationships that might be overlooked by more rigid approaches.

However, a significant challenge emerges when applying BNs to survival analysis, particularly with right-censored data – cases where we only know that a patient's survival time extends beyond a certain point. Standard BN learning algorithms struggle to effectively incorporate this type of incomplete information, potentially limiting their application in long-term patient outcome analysis [18]. This limitation becomes especially relevant in clinical settings, where right-censored data is common due to factors like patient dropout or study conclusion before the event of interest occurs.

To address this limitation, [18] proposed CSBN that combines a BN with a CPH model. Subsequently, the authors demonstrated that the CSBN is also effective for handling missing values in survival data [46]. Inspired by this, [19] utilized the CSBN to develop a survival prediction model and applied it to data from lung cancer patients in China, demonstrating its superior predictive performance and advantages.

CSBN integrates the basic BN and the survival model to construct an extended BN that incorporates (i) the dependence structure among covariates learned from the basic BN and (ii) the causal relationships leading to T2D onset within a specified time horizon, as estimated from the survival model.

The CSBN is constructed in the following steps:

- Select a time point t^* (e.g., 5 or 11 years) and define a binary outcome variable $\mathrm{T2D}_t^*$, which takes value 1 if incident T2D occurs before time t^* and 0 otherwise.
- Fit a CPH model for time to incident T2D, using a variable selection to identify a parsimonious set of important predictors (hereafter, Cox-variables).
- Treat $\mathrm{T2D}_t^*$ as a special outcome node and add it to the basic BN as a terminal child node. Draw directed arcs from each Cox-variable to $\mathrm{T2D}_t^*$, while keeping all other arcs among the covariates identical to those in the basic BN.
- For all combinations of Cox-variable levels, compute the probability of developing T2D by time t^* from the fitted Cox model using the estimated baseline survival function and regression coefficients. Use these probabilities to construct the conditional probability table (CPT) of $\mathrm{T2D}_t^*$ node in the extended BN.

This procedure yields a CSBN that preserves the complex dependence structure among risk factors learned from the basic BN, while simultaneously encoding their causal effects on T2D development over the chosen time horizon through the survival model–based CPT of the $\mathtt{T2D_}t^*$ node.

Note that the continuous survival time (time-to-event) T is discretized into a binary indicator of T2D occurrence by a prespecified time horizon t^*. This formulation allows the CSBN to directly estimate the risk of developing T2D within t^* in a clinically interpretable way. It also enables the use of standard BN software and facilitates computation of individual and joint interventional effects.

Results

Development of the conditional survival Bayesian network model

We learned the structure of the basic BN for the 15 covariates in the final dataset while enforcing the blacklist constraints in Table 2. For bagging, we generated N=1,500 bootstrap replicates and applied the data-driven arc strength threshold proposed by [47]. This criterion selects the cutoff that minimizes the L_1 distance between the empirical cumulative distribution function of the observed arc strengths and its asymptotic counterpart; in our data, the resulting threshold was approximately 0.5. This value coincides with the commonly used conventional threshold of 0.5. The resulting basic BN structure is shown as black arcs in the upper panel of Fig. 3.

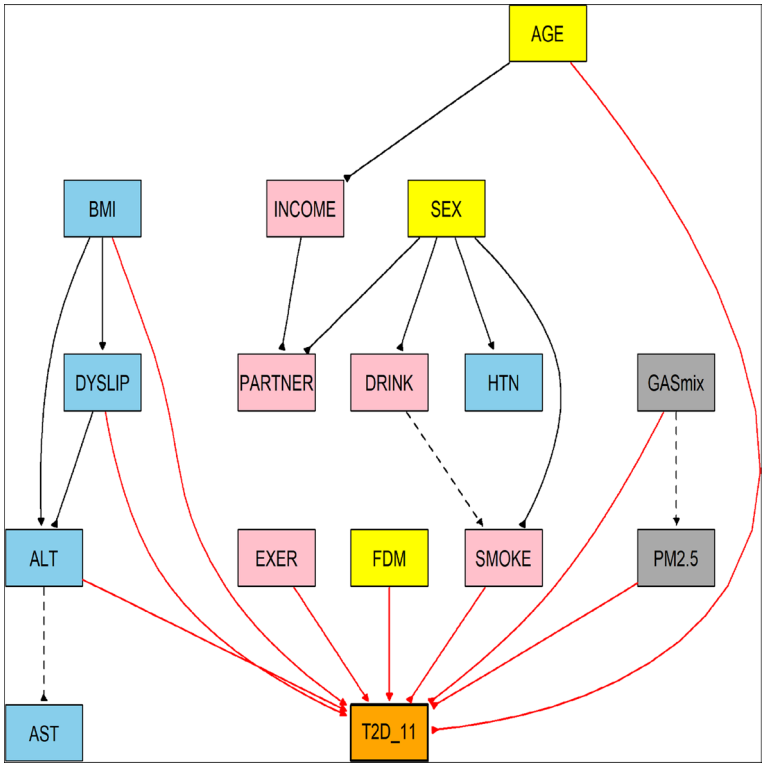


Fig. 3 Conditional Survival Bayesian Network (CSBN) for 11-year incident T2D. Nodes are colored by domain (demographic, lifestyle, metabolic health, and air pollutants). Black arcs represent the learned basic BN structure, and red arcs indicate additional links to the `T2D_11` outcome derived from the survival component. Dashed edges (ALT–AST, DRINK–SMOKE, and PM2.5–GASmix) denote plausibly associational links whose directions should not be interpreted causally

Table 3 Survival analysis results using the Cox proportional hazards model

Variable	Category	Hazard Ratio	95% CI	*p*-value
AGE	Mid	1.207	(0.956, 1.526)	0.1142
AGE	High	1.600	(1.271, 2.013)	<0.0001
BMI	Over	1.310	(1.078, 1.591)	0.0067
BMI	Obese	2.0	(1.314, 3.032)	0.0012
ALT	High	1.58	(1.190, 2.109)	0.0016
FDM	Yes	1.46	(1.004, 2.110)	0.0477
DYSLIP	Yes	1.46	(1.198, 1.775)	0.00017
PM2.5	High	1.27	(1.030, 1.579)	0.0258
GASmix	High	1.26	(1.011, 1.564)	0.0393
SMOKE	Former	0.95	(0.761, 1.183)	0.6410
SMOKE	Current	1.13	(0.867, 1.474)	0.3658
EXER	No	1.09	(0.901, 1.312)	0.3817

We also fitted a CPH model using the `coxph` function in R, estimating the baseline hazard by the Breslow estimator. To assess the stability of variable selection, we conducted AIC- and BIC-based stepwise selection and LASSO. BIC retained only three variables, whereas AIC retained seven (AGE, BMI, ALT, FDM, DYSLIP, PM2.5, and GASmix); LASSO additionally selected SMOKE and EXER. Because our primary aim was to identify a comprehensive risk-factor network rather than to build a maximally parsimonious prediction model, we carried forward the more inclusive set of variables suggested by LASSO. The nine selected variables and their corresponding survival model results are summarized in Table 3.

We assessed the proportional hazards assumption using Schoenfeld residuals [48] via the `cox.zph` function in R. BMI (p = 0.014) and FDM (p = 0.039) showed some statistical evidence against proportional hazards, and the global test p-value was 0.078. However, visual inspection of the Schoenfeld residual plots (Fig. S1) did not reveal clear systematic time trends for any covariate, so we judged any departures from proportionality to be minor and retained the proportional hazards specification.

The Cox model demonstrated moderate overall discrimination, with Harrell's C-index [49] of 0.63. Time-dependent AUCs [50] at t=60, 84, 108, 120, and 132 months were 0.64, 0.61, 0.64, 0.68, and 0.73, respectively, indicating moderate discriminatory performance throughout follow-up, with particularly good discrimination at later time points. Calibration of the 11-year risk predictions was evaluated by dividing participants into deciles of predicted risk, computing Kaplan–Meier estimates of the observed 11-year risk within each group,

and plotting observed versus predicted values [51]. The calibration plot (Fig. S2) showed good overall agreement between predicted and observed risks across most of the risk spectrum.

In this study, we set $t^* = 11$ years to focus on long-term T2D risk; accordingly, we defined the outcome node as T2D_11. Air pollution exposure is a chronic, cumulative environmental risk factor that generally requires prolonged periods to exert measurable effects on metabolic health; an 11-year follow-up window is therefore likely sufficient to capture the latency between sustained air pollution exposure and the biological processes leading to T2D onset.

The variables in Table 3 were then specified as parent nodes of the outcome T2D_11, and the conditional probability table for T2D_11 was derived from the fitted CPH model, thereby completing the construction of the CSBN.

To interpret a DAG as encoding causal relationships, standard causal identification assumptions are required, including *causal sufficiency, consistency, and positivity* [52, 53]. In observational studies, the most problematic of these is often causal sufficiency, that is, the absence of unmeasured confounding factors. Because this assumption is difficult to justify in practice, it is crucial to incorporate domain knowledge when specifying the network structure. In this study, we used a blacklist to encode the major causal backbone of the DAG, as specified in Table 2 and Fig. 2. Accordingly, only the arcs among air pollutants, lifestyle factors, and metabolic factors other than BMI were learned from the data. The resulting CSBN structure is broadly consistent with temporal ordering and clinical knowledge, suggesting that most arcs are reasonable for causal interpretation, except for the ALT → AST, DRINK → SMOKE, and PM2.5 → GASmix links, which are likely associational.

ALT and AST arise from a common underlying process—liver injury—and are therefore strongly correlated. In our Cox model, AST was no longer statistically significant once ALT was included, and only ALT retained an independent association with T2D. Because the arcs into the T2D_11 outcome node and its conditional probability table in the CSBN were derived from this Cox model, the CSBN likewise did not show a direct effect of AST on T2D. In other words, for the ALT–AST pair, the CSBN effectively omitted their latent common cause (liver injury), so the causal identification assumptions were not fully satisfied and the separate causal contributions of ALT and AST could not be reliably distinguished.

Similarly, DRINK and SMOKE often co-occur and may be influenced by shared unmeasured predispositions or social contexts. Likewise, PM2.5 and GASmix partially share the common combustion-related sources. As with the ALT–AST pair, the DRINK–SMOKE and PM2.5–GASmix links likely reflect association rather than directionality, so the arc directions for these pairs should not be interpreted causally.

To reflect these limitations, we depicted the ALT–AST, DRINK–SMOKE, and PM2.5–GASmix arcs as dashed edges in the DAG, indicating that they should be viewed as associational links rather than direct causal effects. Accordingly, we did not attempt to interpret causal effects of AST and DRINK, which were not directly linked to T2D_11.

Figure 3 displays the CSBN constructed in this study, with nodes colored by category—demographic factors, lifestyle factors, metabolic health, and air pollutants—as summarized in Table 2. Arcs inherited from the basic BN are shown in black, whereas arcs added from the survival model that link covariates to the T2D_11 outcome node are shown in red. To reflect potential source-driven or co-occurrence-driven dependencies that are more plausibly associational than directional, the ALT–AST, DRINK–SMOKE, and PM2.5–GASmix links are depicted as dashed edges; the directions of these pairs should not be interpreted as direct causal effects.

Causal pathways in the conditional survival Bayesian network

Demographic factors

Age and family history of diabetes (FDM) both exert direct influences on the T2D_11 outcome node, reflecting well-established evidence that older age and genetic predisposition substantially increase diabetes risk. Sex is not directly linked to the T2D_11 node; instead, its influence appears to be mediated through smoking, rather than acting as an independent predictor in this cohort.

Metabolic factors

BMI, dyslipidemia (DYSLIP), and elevated ALT all show direct arcs to the T2D_11 outcome node, highlighting adiposity, abnormal lipid metabolism, and impaired liver function as key metabolic determinants of long-term T2D risk. In contrast, hypertension (HTN) is not directly connected to T2D_11, suggesting that its independent contribution is weak once BMI, DYSLIP, and ALT are taken into account.

Lifestyle and social factors

Physical activity (EXER) and smoking status (SMOKE) each exhibit direct connections to the T2D_11 node. Lower physical activity and current smoking are associated with higher 11-year risk of T2D, consistent with their roles as modifiable behavioral risk factors. Income affects partnership status but does not show direct arcs to T2D_11.

Air pollutants

Both the combined gaseous mixture index (GASmix) and PM2.5 have direct arcs to the T2D_11, indicating that higher long-term exposure to fine particulate matter and traffic-related gases contributes to increased 11-year risk of incident T2D.

Individual effects

The CSBN structure shown in Fig. 3 is broadly consistent with established domain knowledge and is suitable for causal interpretation of the relationships among the variables, except for the dashed edges, which represent associational rather than causal relationships.

Based on this causal CSBN structure, we estimated interventional probabilities of developing T2D over the 11-year follow-up, $P(\text{T2D by 11 years} \mid \text{do}(X = x))$, for each variable X using the do-calculus framework of Pearl [53] and the potential outcomes formulation of Hernán and Robins [52]. This approach conceptually fixes the value of a given risk factor (e.g., obesity vs. normal BMI) while detaching it from its parent nodes, thereby removing confounding influences and yielding causal risk estimates under hypothetical interventions.

For root nodes such as AGE, BMI, PM2.5, and GASmix—variables without parents in the DAG—the interventional probability of incident T2D coincides with the corresponding conditional probability $P(\text{T2D by 11 years} \mid X = x)$, because their marginal distributions are not altered by intervention.

In contrast, for variables that are confounded (e.g., income, partnership status, hypertension), the interventional probability may differ substantially from the conditional probability. These variables have no direct arc to the T2D_11 outcome node, so setting them to a given level while cutting incoming arcs yields negligible interventional probabilities, even if their conditional probabilities are sizeable. For example, INCOME is confounded by AGE: the conditional probability of 11-year incident T2D among participants with low income was around 26%, but once we intervened on INCOME while holding the rest of the network structure fixed, its interventional probability was negligible. This contrast underscores the importance of focusing on interventional rather than purely conditional associations.

Figure 4a and b show, respectively, the differences and ratios in the interventional probabilities of developing T2D over an 11-year period for each level of variables that directly or indirectly affect incident T2D, compared with their base levels. The corresponding interventional probabilities are provided in Fig. S3. Obesity emerged as the strongest individual risk factor: intervening to set BMI to the obese category increased the 11-year probability of T2D by roughly 20 percentage points (%p) relative to normal BMI (causal risk ratio ≈ 2.0). Elevated ALT and high age showed the next largest effects, each raising T2D risk by about 8–10 %p (risk ratios ≈ 1.4–1.5). Family history of T2D, dyslipidemia, and overweight status (BMI Over) also made substantial contributions, with risk differences of about 5–9 %p (risk ratios ≈ 1.3). For environmental exposures, high $\text{PM}_{2.5}$ and a high gaseous pollution mixture (GASmix) increased 11-year T2D risk by approximately 3–5 %p (risk ratios ≈ 1.2), indicating that both particulate and gaseous pollution contribute meaningfully to disease burden in this cohort. Current smoking, and lack of regular exercise had small positive effects (risk ratios ≈ 1.1), whereas current drinking and male sex showed little interventional impact, and former smoking was slightly protective.

Interaction effects

We next examined how pairs of risk factors jointly influence the 11-year risk of T2D. For each pair, we computed the interventional probabilities of developing T2D when both risk factors were elevated (p_{11}), only the first or second factor was elevated (p_{10}, p_{01}), or neither factor was elevated (p_{00}). From these joint probabilities, we derived standard measures of additive interaction—the relative excess risk due to interaction (RERI), the attributable proportion due to interaction (AP), and the synergy index (S)—following Andersson et al. [54].

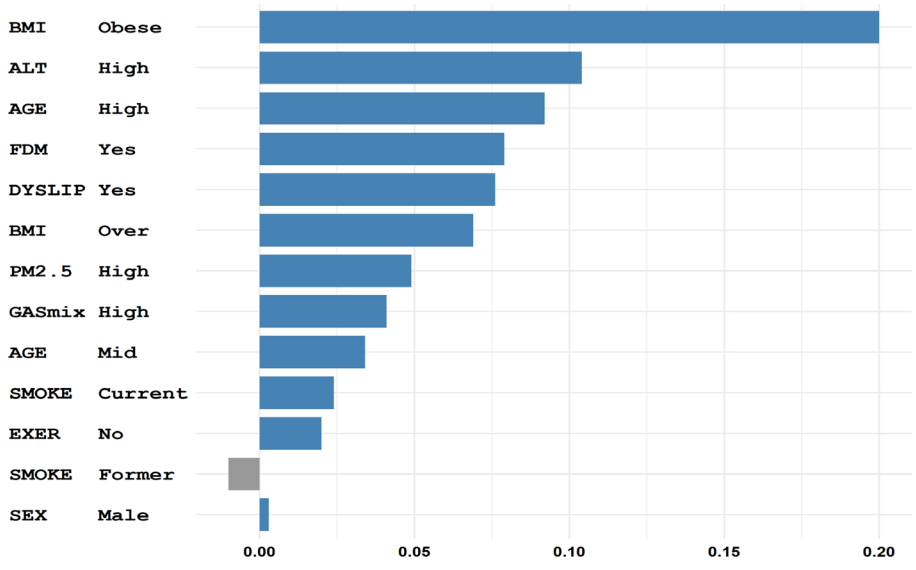
Formally, defining $RR_{11} = p_{11}/p_{00}, \quad RR_{10} = p_{10}/p_{00}, \quad RR_{01} = p_{01}/p_{00}$, we obtain

$$\begin{aligned}
\text{RERI} &= RR_{11} - RR_{10} - RR_{01} + 1 = (p_{11} - p_{10} - p_{01} + p_{00})/p_{00}, \\
\text{AP} &= \text{RERI}/RR_{11} = (p_{11} - p_{10} - p_{01} + p_{00})/p_{11}, \\
S &= (RR_{11} - 1)/[(RR_{10} - 1) + (RR_{01} - 1)] \\
&= (p_{11} - p_{00})/[(p_{10} - p_{00}) + (p_{01} - p_{00})].
\end{aligned}$$

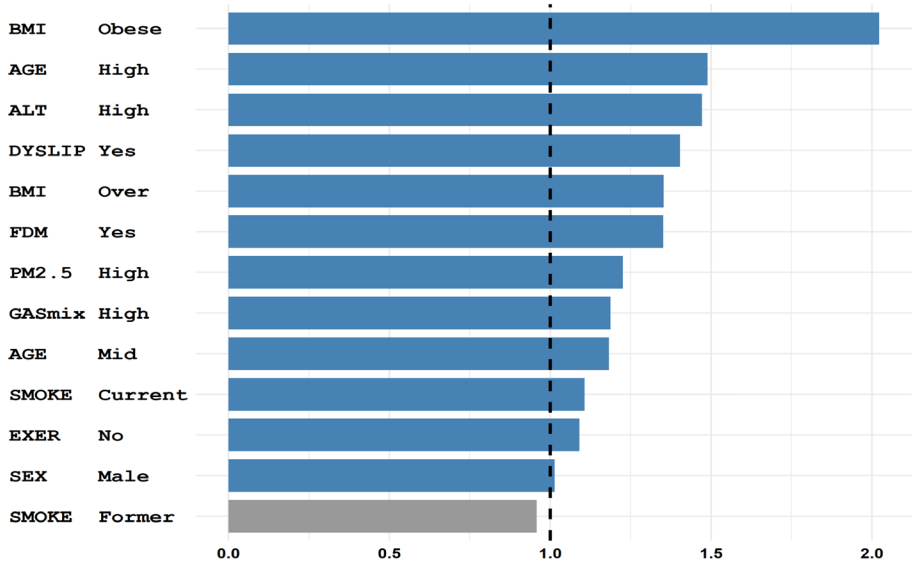
Table 4 summarizes these measures for pairs of risk factors that exhibited significant positive interactions at the 5% significance level. Overall, combinations of metabolic risk factors showed the strongest positive interactions. Pairs involving elevated ALT with family history of T2D, dyslipidemia, or obesity yielded more-than-additive joint effects. Other metabolic combinations, such as DYSLIP with FDM, BMI with DYSLIP, and BMI with FDM, also exhibited positive RERI and AP values and synergy indices above 1. These findings suggest that co-occurring metabolic abnormalities jointly elevate long-term T2D risk beyond what would be expected from their individual effects.

Short-term (5-year) T2D risk analysis

While the CSBN model was initially applied to an 11-year horizon ($t^* = 11$ years) to evaluate long-term T2D risk, its flexibility allows prediction of survival continuously,



(a) Absolute risk differences (interventional risk minus baseline risk).



(b) Interventional risk ratios (elevated vs reference levels).

Fig. 4 Interventional effects of risk factors on 11-year T2D risk in the CSBN. **a** Absolute risk differences (interventional risk minus baseline risk). **b** Interventional risk ratios (elevated vs reference levels)

Table 4 Interaction measures for combinations of variables that exhibited statistically significant interaction at the 5% significance level for 11-year incident T2D

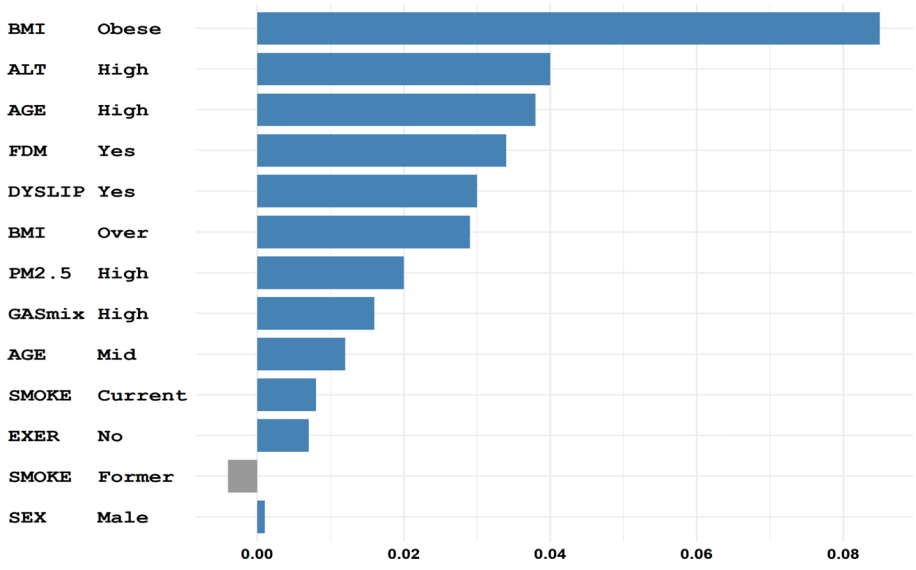
Var1	Var2	p_{11}	p_{10}	p_{01}	p_{00}	RERI	AP	S
ALT	FDM	0.428	0.317	0.296	0.216	0.142	0.072	1.170
ALT	DYSLIP	0.369	0.271	0.254	0.182	0.142	0.070	1.161
ALT	BMI	0.368	0.283	0.252	0.192	0.127	0.066	1.160
DYSLIP	FDM	0.353	0.261	0.256	0.184	0.112	0.059	1.140
BMI	DYSLIP	0.298	0.212	0.233	0.164	0.108	0.059	1.151
BMI	FDM	0.353	0.261	0.265	0.193	0.105	0.057	1.146

accommodating various time horizons. To explore short-term dynamics, we conducted a reanalysis using a 5-year horizon ($t^* = 5$ years).

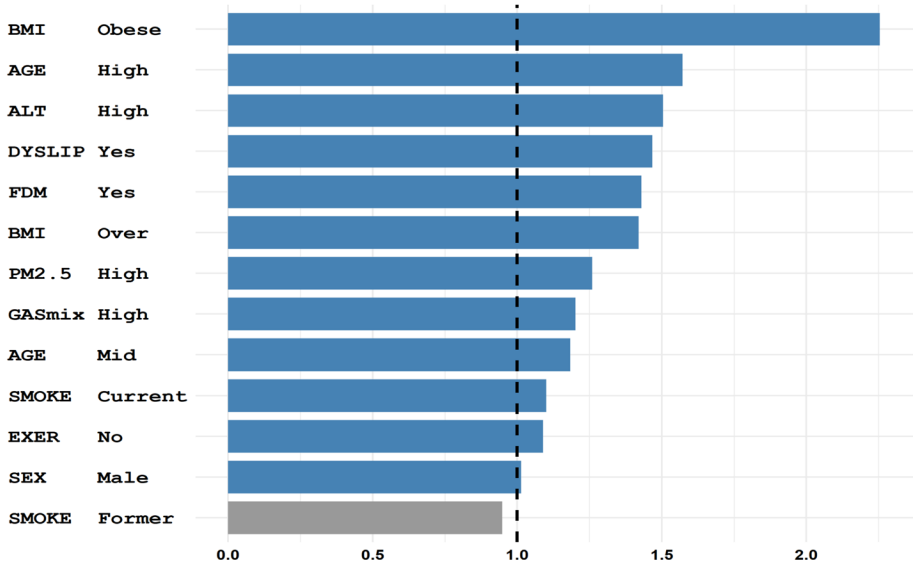
In fact, the CSBN structure for $t^* = 5$ years was identical to that for $t^* = 11$ years, suggesting that the key causal pathways linking demographic, metabolic, lifestyle, and environmental factors to T2D had already been established over the shorter horizon and remained stable over time, with changes occurring mainly in the magnitude of interventional probabilities rather than in the network topology itself.

Figure 5a and b show, respectively, the differences and ratios of interventional probabilities of developing T2D over 5 years for each level of variables that directly or indirectly affect incident T2D, relative to their respective base levels. The corresponding interventional probabilities are provided in Fig. S4.

Across both the 5-year and 11-year horizons, the ranking of key risk factors was essentially unchanged: obesity was by far the dominant driver, followed by high ALT, older age, family history of T2D, dyslipidemia, overweight, and then high $PM_{2.5}$ and high GASmix. The



(a) Absolute risk differences (interventional risk minus baseline risk).



(b) Interventional risk ratios (elevated vs reference levels).

Fig. 5 Interventional effects of risk factors on 5-year T2D risk in the CSBN

Table 5 Interaction measures for combinations of variables listed in Table 4 for 5-year incident T2D

Var1	Var2	p_{11}	p_{10}	p_{01}	p_{00}	RERI	AP	S
ALT	DYSLIP	0.140	0.098	0.089	0.063	0.240	0.109	1.247
ALT	FDM	0.167	0.118	0.109	0.075	0.212	0.096	1.211
ALT	BMI	0.141	0.104	0.091	0.066	0.187	0.088	1.201
BMI	FDM	0.134	0.094	0.095	0.067	0.183	0.092	1.225
DYSLIP	FDM	0.133	0.093	0.092	0.064	0.179	0.086	1.200
BMI	DYSLIP	0.110	0.075	0.082	0.056	0.148	0.076	1.184

main difference lay in the *magnitude* of effects. Over 5 years, absolute risk differences were smaller (e.g., obesity increased 5-year T2D risk by < 0.10 vs. ≈ 0.20 over 11 years), but the corresponding risk ratios were already large (BMI Obese ≈ 2.0 at both horizons), indicating that the relative effect of obesity emerged early and stayed stable, whereas the cumulative absolute burden grew over time.

To compare early-stage interaction dynamics with long-term dynamics, we estimated interaction measures based on the interventional probabilities of developing T2D within 5 years for the variable combinations listed in Table 4. The results are summarized in Table 5. Comparison of Table 4 (11-year follow-up) and Table 5 (5-year follow-up) showed that the same core pairs of risk factors drove positive additive interaction. However, interaction measures were generally stronger over 5 years. In the 5-year analysis, RERI values were larger (approximately 0.11–0.24 vs. 0.07–0.14 at 11 years), AP reached about 0.10–0.11 (vs. 0.04–0.07), and the synergy index ranged around 1.18–1.25 (vs. 1.10–1.17). These findings suggest that more-than-additive effects of clustered metabolic risk factors were already pronounced in the short term, whereas over the longer 11-year horizon the same pairs remained important but their interactions appeared attenuated relative to the growing overall risk.

Discussion

Study design strengths

Our research design offers several advantages for understanding T2D risk. First, the 11-year longitudinal follow-up provides sufficient time to capture the chronic effects of environmental exposures on diabetes development. Second, we simultaneously considered traditional demographic, lifestyle, and metabolic risk factors together with multiple air pollutants, enabling assessment of their relative contributions and interactions within a unified analytical framework. Third, the study was conducted in an industrial region (Ansan) with persistently high levels of multiple air pollutants, providing a natural laboratory to investigate the long-term impact of elevated environmental exposures in combination with metabolic and lifestyle factors.

Methodological advantages

The CSBN approach offers substantial methodological advantages compared with traditional BNs or survival analysis alone. First, CSBN integrates the strengths of BN modeling and survival analysis, allowing right-censoring and incomplete data to be handled within a coherent probabilistic framework. Second, the approach yields a global view of the causal structure among all risk factors and supports intuitive visualization of complex relationships. Third, it can distinguish direct, indirect, and confounded associations, thereby clarifying the pathways through which risk factors contribute to long-term incident T2D. Fourth, CSBN enables estimation of both individual and joint (interaction) interventional effects, revealing synergistic relationships that might be overlooked by conventional models. Finally, the framework accommodates both long-term and short-term risk horizons, increasing its flexibility for different clinical and research applications.

Principal findings and practical implications

Our analysis identified BMI, ALT, family history of T2D, age, and dyslipidemia as key risk factors consistently important in both long-term (11-year) and short-term (5-year) T2D survival analyses. This consistency across different temporal scales emphasizes these factors as fundamental pathogenic mechanisms that should be prioritized in both immediate and long-term prevention strategies. The persistent importance of these factors suggests that interventions targeting weight management, liver function optimization, and lipid control remain cornerstone approaches for diabetes prevention regardless of the intervention timeline. These findings are consistent with previous studies highlighting the central importance of adiposity, liver function, and lipid abnormalities in T2D etiology [8, 26, 27, 55].

The presence of direct arcs from both $\mathrm{PM}_{2.5}$ and GASmix to the T2D_11 outcome node, and their roughly 20% increases in interventional risk, indicate that in this industrial setting both fine particles and a combustion-related gaseous mixture make meaningful contributions to diabetes incidence. This is in line with accumulating evidence that long-term exposure to air

pollution promotes insulin resistance, systemic inflammation, and T2D [56–61]. These findings support the inclusion of air-quality improvement as part of diabetes-prevention policy. Reductions in combustion-related emissions may not only prevent respiratory and cardiovascular disease but also help slow the rising incidence of T2D, particularly in densely populated industrial regions.

In this study, variables such as hypertension, household income, and partnership status—although associated with T2D in Korean populations in conventional analyses [62–64]—showed no direct causal effect on long-term T2D risk once the network structure and confounding pathways were accounted for. This illustrates how CSBN can help differentiate factors that act primarily as markers of underlying risk from those that lie on causal pathways to disease.

Beyond individual effects, the CSBN highlighted that clustered metabolic disturbances markedly amplify T2D risk. The largest additive interactions were observed for combinations involving elevated ALT with family history of T2D, dyslipidemia, or obesity, and for pairs of metabolic risk factors such as obesity with dyslipidemia or family history. These patterns were evident for both 11-year and 5-year horizons, with somewhat stronger interaction measures at 5 years. Prioritizing individuals who simultaneously exhibit these metabolic risk factors for intensive lifestyle interventions and metabolic screening could yield disproportionate benefits.

Limitations and future directions

Several limitations should be acknowledged. A key limitation is the snapshot nature of the baseline covariates, particularly lifestyle and socioeconomic factors such as exercise, smoking, alcohol consumption, income, partnership status, and hypertension. Although extensive research has shown that income, partnership status, and hypertension influence T2D onset and progression [8, 65–67], they did not exhibit strong direct causal effects on long-term T2D risk in our CSBN. This discrepancy likely reflects behavioral and socioeconomic changes over the 11-year follow-up that are not captured by baseline measurements. Applying dynamic or longitudinal BNs could further clarify how trajectories of body weight, liver function, and environmental exposures interact over time to determine diabetes risk. Incorporating longitudinal covariate data or using time-varying coefficient models would allow researchers to identify critical intervention windows, capture temporal variation in risk factor effects, and provide a more comprehensive understanding of their evolving roles in T2D progression.

Using the Ansan cohort provided both distinctive advantages and important limitations. Ansan's dual role as an industrial hub and a recipient of transboundary pollution from China resulted in an extremely high exposure profile involving both gaseous pollutants and particulate matter, providing a unique opportunity to examine chronic disease risk and associations with specific pollutants under real-world, high-pollution conditions. Asian regions differ markedly from Western regions in terms of health status, healthcare infrastructure, pollution exposure patterns, and socioeconomic circumstances, yet studies on the health impacts of pollution exposure in these Asian contexts remain scarce, highlighting a strong need for more comprehensive studies in understudied Asian regions [68–71]. Therefore, this study makes a significant contribution by investigating how high pollution levels in an Asian context influence diabetes risk, providing valuable insights to inform preventive strategies and public health policy decisions across Asia. However, the uniformly high pollution exposure in this setting also presents limitations, as it reduces the availability of a true low-exposure reference group, thereby constraining the exposure contrast needed to robustly evaluate dose-response relationships.

The need to discretize continuous variables for BN learning may have obscured nonlinear dose–response relationships and reduced statistical power. We initially used the 75th percentiles as cut-offs for discretizing air pollutants. To assess the impact of this choice, we repeated the analysis using the 50th and 25th percentiles as alternative cut-offs. When using the 50th percentiles (median splits), the magnitude of the effect of $PM_{2.5}$ was attenuated, although the overall CSBN structure remained unchanged. However, when using the 25th percentiles as cut-offs, the direct arc from $PM_{2.5}$ to the T2D outcome node disappeared. This pattern suggests that the elevated T2D risk is driven primarily by the highest levels of $PM_{2.5}$ exposure, and that moderate elevations are less clearly distinguishable from the reference category. It also highlights that CSBN inferences for weaker environmental effects can be sensitive to discretization choices, underscoring the need for cautious interpretation and, where possible, complementary analyses based on multi-category discretization.

Causal interpretation of arcs in a DAG requires causal identification assumptions, including (at minimum) a reasonable degree of *causal sufficiency*. In this study, we judged causal sufficiency to be unlikely for three specific pairs—ALT–AST, PM2.5–GASmix, and DRINK–SMOKE—because plausible shared unobserved determinants may underlie these associations. We therefore excluded causal interpretation for these links and conducted causal inference only for the remaining relationships. Nevertheless, our analysis remains subject to residual confounding from unmeasured or imperfectly measured factors—such as more detailed dietary information and occupational or indoor air-pollution exposures—which cannot be fully ruled out.

Finally, the analysis is restricted to the KoGES Ansan cohort, an industrial urban population that may not be globally representative. The relatively high levels of air pollution and the specific demographic and socioeconomic profile of this cohort mean that our findings are most directly generalizable to similar industrialized urban settings. External validation of the CSBN structure and interventional contrasts in other cohorts or independent populations is warranted and should be interpreted with caution.

Future work should extend the CSBN approach to larger and more diverse populations and incorporate time-varying covariates, richer lifestyle and occupational histories, and more detailed pollutant mixtures. As data linkages to mortality and hospital records become available, competing-risk frameworks could be integrated into the CSBN to address the joint dynamics of T2D onset, complications, and death.

Conclusion

Using an 11-year follow-up of the KoGES Ansan cohort, we applied the CSBN to disentangle how demographic, lifestyle, metabolic, and environmental factors jointly shape the development of T2D. The CSBN framework combined a BN with a Cox model, allowing us to translate complex dependence structures into interpretable interventional risks over both long-term (11-year) and short-term (5-year) horizons. This approach provided a coherent way to accommodate censoring, distinguish direct from confounded pathways, and evaluate both individual and joint causal effects within a single graphical model.

Substantively, our analysis showed that obesity, elevated ALT, dyslipidemia, older age, and family history of diabetes are the dominant causal determinants of long-term T2D incidence in this industrial urban population, with smoking and low physical activity contributing additional risk. Long-term exposure to fine particulate matter and a combustion-related gaseous mixture also increased T2D risk, even after accounting for traditional risk factors, underscoring the role of ambient air pollution as an important environmental driver of diabetes. The CSBN further highlighted that clustered metabolic abnormalities—particularly combinations involving obesity, dyslipidemia, elevated ALT, and family history—exert more-than-additive effects, with evidence of stronger interactions already apparent within five years.

These findings highlight the value of considering risk factors simultaneously and support the development of targeted prevention strategies for industrialized populations facing combined pollution and cardiometabolic burdens.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12889-026-26358-9.

Supplementary Material 1.

Acknowledgements
The authors sincerely thank the editor and the two referees for their constructive and important comments, which greatly strengthened the quality and clarity of this article. Data used in this study were provided by the Clinical & Omics Data Archive (CODA), the Korea Disease Control and Prevention Agency (KDCA) under approval (CODA_S2400014-01).

Authors' contributions
JC: Conceptualization, Formal analyses, Methodology, Visualization. MO: Conceptualization, Methodology, Writing, Supervision. All authors read and approved the final manuscript.

Funding
This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (NRF-2022R1A2C1091256).

Data availability
De-identified data underlying this study are available from the Clinical & Omics Data Archive (CODA) of the Korea Disease Control and Prevention Agency (KDCA; https://www.kdca.go.kr/) under controlled access (project code: CODA_S240001401). The authors accessed these data under an approved data-use agreement and are not permitted to publicly share them. Qualified researchers may request access through KDCA/CODA; the authors will provide study materials and guidance upon reasonable request and subject to KDCA authorization.

Declarations

Ethics approval and consent to participate
This study utilized data from the Korean Genome and Epidemiology Study (KoGES), which was approved by the Institutional Review Board (IRB) of the Korea Disease Control and Prevention Agency (KDCA). All participants provided written informed consent at enrollment. All methods were carried out in accordance with the ethical standards outlined in the 1964 Declaration of Helsinki and its later amendments.
The present secondary data analysis was reviewed and granted an exemption by the Institutional Review Board of Ewha Womans University (IRB No. 2024-0079), as it involved anonymized data with no direct participant interaction or identifiable information.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Statistics, Ewha Womans University, 52 Ewha-Yeo-Dae-Gil, Sodaemun-Gu, Seoul 36074, Korea

Received: 18 July 2025 / Accepted: 15 January 2026
Published online: 21 January 2026

References

1. Kahn SE, Cooper ME, Del Prato S. Pathophysiology and treatment of type 2 diabetes: perspectives on the past, present, and future. Lancet. 2014;383(9922):1068–83. https://doi.org/10.1016/S0140-6736(13)62154-6.
2. Punthakee Z, Goldenberg R, Katz P. Definition, classification and diagnosis of diabetes, prediabetes and metabolic syndrome. Can J Diabetes. 2018;42:S10-5. https://doi.org/10.1016/j.jcjd.2017.10.003.

3. International Diabetes Federation. IDF Diabetes Atlas. 2015. 7th ed. Available from: http://www.diabetesatlas.org. Accessed 12 Jul 2025.
4. Choi HH, Choi G, Yoon H, Ha KH, Kim DJ. Rising incidence of diabetes in young adults in South Korea: a national cohort study. Diabetes Metab J. 2022;46(5):803–7. https://doi.org/10.4093/dmj.2021.0236.
5. Kim YJ, Yeom HY. Evaluating the Predictive Efficiency of Obesity-related Factors for Type 2 Diabetes: A Panel Study Using KoGES Data. J Health Inform Stat. 2024;49(1). https://doi.org/10.21032/jhis.2024.49.1.54.
6. Park SE, Ko SH, Kim JY, Kim K, Moon JH, Kim NH, et al. Diabetes fact sheets in Korea 2024. Diabetes Metab J. 2025;49(1):24–33. https://doi.org/10.4093/dmj.2024.0818.
7. Ismail L, Materwala H, Al KJ. Association of risk factors with type 2 diabetes: a systematic review. Diabetes Res Clin Pract. 2021;174:108799.
8. Bellou V, Belbasis L, Tzoulaki I, Evangelou E. Risk factors for type 2 diabetes mellitus: an exposure-wide umbrella review of meta-analyses. PLoS One. 2018;13(3):e0194127. https://doi.org/10.1371/journal.pone.0194127.
9. Li Y, Xu L, Shan Z, Teng W, Han C. Association between air pollution and type 2 diabetes: an updated review of the literature. Diabetes Res Clin Pract. 2019;148:94–103.
10. Yang BY, Fan S, Thiering E, Seissler J, Nowak D, Dong GH, et al. Ambient air pollution and diabetes: a systematic review and meta-analysis. Environ Res. 2020;180:108817. https://doi.org/10.1016/j.envres.2019.108817.
11. Liu P, Zhang Z, Luo M. Relationship between air pollution exposure and insulin resistance in Chinese middle-aged and older populations: evidence from Chinese cohort. Front Public Health. 2025;13:1551851. https://doi.org/10.3389/fpubh.2025.1551851.
12. Eze IC, Schaffner E, Fischer E, Schikowski T. Long-term air pollution exposure and diabetes: a systematic review and meta-analysis. Environ Int. 2015;70:14–22. https://doi.org/10.1016/j.envint.2014.05.014.
13. Feizi A, Shahraki PK, Najafabadi AM, Iraj B, Abyar M, Amini M, et al. The association of exposure to air pollution with changes in plasma glucose indices, and incidence of diabetes and prediabetes: a prospective cohort of first-degree relatives of patients with type 2 diabetes. J Res Med Sci. 2023;28(1):21. https://doi.org/10.4103/jrms.jrms_477_22.
14. Bonanni LJ, Wittkopp S, Long C, Aleman JO, Newman JD. A review of air pollution as a driver of cardiovascular disease risk across the diabetes spectrum. Front Endocrinol. 2024;15:1321323. https://doi.org/10.3389/fendo.2024.1482063.
15. Blangiardo M, Pirani M, Kanapka L, Hansell A, Fuller G. A hierarchical modelling approach to assess multi pollutant effects in time-series studies. PLoS One. 2019;14(3):e0212565. https://doi.org/10.1371/journal.pone.0212565.
16. Englert J, Ebelt S, Chang H. Modeling Joint Health Effects of Environmental Exposure Mixtures with Bayesian Additive Regression Trees. 2024. arXiv preprint arXiv:2411.09025.
17. Kim Y, Han BG. Cohort profile: the Korean Genome and Epidemiology Study (KoGES). Int J Epidemiol. 2017;46(2):e20–e20. https://doi.org/10.1093/ije/dyv316.
18. Rabinowicz S, Hommersom A, Butz RS, Williams M. A Prognostic Model of Glioblastoma Multiforme Using Survival Bayesian Networks. In: ten Teije A, Popow C, Holmes JH, Sacchi L, editors. Artificial Intelligence in Medicine: 16th Conference on Artificial Intelligence in Medicine (AIME 2017), Vienna, Austria, June 21–24, 2017, Proceedings. vol. 10259 of Lecture Notes in Computer Science. Cham: Springer International Publishing; 2017. pp. 81–85. https://doi.org/10.1007/978-3-319-59758-4_9.
19. Zhong L, Yang F, Sun S, Wang L, Yu H, Nie X, et al. Predicting lung cancer survival prognosis based on the conditional survival Bayesian network. BMC Med Res Methodol. 2024;24(1):16. https://doi.org/10.1186/s12874-023-02043-y.
20. Scutari M, Denis JB. Bayesian Networks: With Examples in R. 2nd ed. Boca Raton: Chapman and Hall/CRC; 2021. https://doi.org/10.1201/9780429347436.
21. Woo HD, Song DS, Choi SH, Park JK, Lee K, Yun HY, et al. Integrated dataset of the Korean Genome and Epidemiology Study cohort with estimated air pollution data. Epidemiol Health. 2022;44:e2022071. https://doi.org/10.4178/epih.e2022071.
22. Song C, Lim CC, Gurmu BL, Kim S, Lee M, Park J, et al. Comparison of Personal or Indoor $PM_{2.5}$ Exposure Level to That of Outdoor Over Four Seasons in Selected Urban, Industrial, and Rural Areas of South Korea: K-IOP Study. Int J Environ Res Public Health. 2023;20(17):6684. https://doi.org/10.3390/ijerph20176684.
23. American Diabetes Association. 2. classification and diagnosis of diabetes: standards of medical care in diabetes—2021. Diabetes Care. 2021;44(Supplement 1):S15–33. https://doi.org/10.2337/dc21-S002.
24. Fletcher B, Gulanick M, Lamendola C. Risk factors for type 2 diabetes mellitus. J Cardiovasc Nurs. 2002;16(2):17–23. https://doi.org/10.1097/00005082-200201000-00003.
25. Ley SH, Meigs JB, In: Bonora E, DeFronzo RA, editors. Epidemiology and Risk Factors of Type 2 Diabetes. Cham: Springer International Publishing; 2018. pp. 55–80. https://doi.org/10.1007/978-3-319-45015-5_5.
26. Bi Y, Wang T, Xu Y, Xu M, Chen Y, Liu Y, et al. Advanced research on risk factors of type 2 diabetes. Diabetes Metab Res Rev. 2012;28(S2):32–9. https://doi.org/10.1002/dmrr.2352.
27. Singh A, Kumar R, Sharma P, Gupta N. Deciphering the complex interplay of risk factors in type 2 diabetes mellitus: a comprehensive review. Metab Open. 2024;22:100287. https://doi.org/10.1016/j.metop.2024.100287.
28. World Health Organization. Obesity: Preventing and Managing the Global Epidemic: Report of a WHO Consultation, Technical Report vol. 894. Geneva: World Health Organization; 2000.
29. Unger T, Borghi C, Charchar F, Khan NA, Poulter NR, Prabhakaran D, et al. 2020 International Society of Hypertension Global Hypertension Practice Guidelines. Hypertension. 2020;75(6):1334–57. https://doi.org/10.1161/HYPERTENSIONAHA.120.15026.
30. Korean Society of Lipid and Atherosclerosis (KSoLA) Committee of Clinical Practice Guideline. Korean guidelines for the management of dyslipidemia. KSoLA Seoul; 2022. Available from: https://www.lipid.or.kr/reference/guideline.php. Accessed 12 Jul 2025.
31. Clinical and Laboratory Standards Institute. Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory; Approved Guideline. 3rd ed. Wayne: CLSI; 2010. CLSI document C28-A3c. Available from: https://clsi.org/standards/products/method-evaluation/documents/ep28/. Accessed 12 Jul 2025.
32. Byun D, Schere KL. Review of the governing equations, computational algorithms, and other components of the Models-3 Community Multiscale Air Quality (CMAQ) Modeling System. Appl Mech Rev. 2006;59(2):51–77. https://doi.org/10.1115/1.2128636.
33. Appel KW, Bash JO, Fahey KM, Foley KM, Gilliam RC, Hogrefe C, et al. The Community Multiscale Air Quality (CMAQ) model versions 5.3 and 5.3.1: system updates and evaluation. Geosci Model Dev. 2021;14(5):2867–97. https://doi.org/10.5194/gmd-14-2867-2021.
34. Choi DR, Jung J, Lee K, Park S, Kim J, Park RJ, et al. High-Resolution Daily PM2.5 Exposure Concentrations in South Korea Using CMAQ Data Assimilation with Surface Measurements and MAIAC AOD (2015–2021). Atmosphere. 2024;15(10):1152. https://doi.org/10.3390/atmos15101152.
35. Lu X, Xie Q, Pan X, Zhou B, Wang L, et al. Type 2 diabetes mellitus in adults: pathogenesis, prevention and therapy. Signal Transduct Target Ther. 2024;9(1):262. https://doi.org/10.1038/s41392-024-01951-9.
36. Galaviz KI, Narayan KV, Lobelo F, Weber MB. Lifestyle and the prevention of type 2 diabetes: a status report. Am J Lifestyle Med. 2018;12(1):4–20. https://doi.org/10.1177/1559827615619159.
37. Halpern A, Mancini MC, Nery M. Metabolic syndrome, dyslipidemia, hypertension and type 2 diabetes in youth: from diagnosis to treatment. Diabetol Metab Syndr. 2010;2(1):55. https://doi.org/10.1186/1758-5996-2-55.
38. Rajagopalan S, Brook RD. Air pollution and type 2 diabetes: mechanistic insights. Diabetes. 2012;61(12):3037–45. https://doi.org/10.2337/db12-0190.
39. Schwarz G. Estimating the dimension of a model. Ann Stat. 1978;6(2):461–4.
40. Geiger D, Heckerman D, Meek C. Asymptotic Model Selection for Directed Networks with Hidden Variables. In: Jordan MI, editor. Learning in Graphical Models. Dordrecht, Netherlands: Kluwer Academic Publishers; 1998. pp. 461–477. https://doi.org/10.1007/978-94-011-5014-9_16.
41. Friedman N, Goldszmidt M, Wyner A. Data Analysis with Bayesian Networks: A Bootstrap Approach. In: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence. 1999. pp. 196–205. arXiv:1301.6695.
42. Cox DR. Regression models and life-tables. J Roy Stat Soc: Ser B (Methodol). 1972;34(2):187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x.
43. Kleinbaum DG, Klein M. Survival Analysis: A Self-Learning Text. 3rd ed. New York: Springer; 2012. https://doi.org/10.1007/978-1-4419-6646-9.
44. Harrell FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. 2nd ed. New York: Springer; 2015. https://doi.org/10.1007/978-3-319-19425-7.
45. Lucas PJ. Biomedical applications of Bayesian networks. In: Advances in probabilistic graphical models. Cham: Springer; 2021. pp. 333–58. https://doi.org/10.1007/978-3-540-68996-6_16.
46. Rabinowicz S, Butz R, Hommersom A, Williams M. CSBN: A Hybrid Approach For Survival Time Prediction With Missing Data. In: ECML/PKDD Workshop on Advanced Analytics and Learning on Temporal Data. 2018. Available from: https://project.inria.fr/aaldt18/files/2018/08/aaldt18_csbn.pdf. Accessed 12 Jul 2025.

47. Scutari M, Nagarajan R. Identifying significant edges in graphical models of molecular networks. Artif Intell Med. 2013;57(3):207–17. https://doi.org/10.1016/j.artmed.2012.12.006.
48. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika. 1994;81(3):515–26. https://doi.org/10.1093/biomet/81.3.515.
49. Harrell FEJ, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15(4):361–87. https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.
50. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Bossuyt P, et al. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230. https://doi.org/10.1186/s12916-019-1466-7.
51. Crowson CS, Atkinson EJ, Therneau TM. Assessing calibration of prognostic risk scores. Stat Methods Med Res. 2016;25(4):1692–706. https://doi.org/10.1177/0962280213497434.
52. Hernan MA, Robins JM. Causal Inference: What If. Boca Raton: Chapman and Hall/CRC; 2020.
53. Pearl J. Causal inference in statistics: an overview. Stat Surv. 2009;3:96–146. https://doi.org/10.1214/09-SS057.
54. Andersson T, Alfredsson L, Källberg H, Zdravkovic S, Ahlbom A. Calculating measures of biological interaction. Eur J Epidemiol. 2005;20(7):575–9. https://doi.org/10.1007/s10654-005-7835-x.
55. Oh TJ, Moon JH, Choi SH, Cho YM, Park KS, Cho NH, et al. Development of a clinical risk score for incident diabetes: a 10-year prospective cohort study. J Diabetes Investig. 2021;12(4):610–8. https://doi.org/10.1111/jdi.13382.
56. Park SK, Wang W. Ambient air pollution and type 2 diabetes mellitus: a systematic review of epidemiologic research. Curr Environ Health Rep. 2014;1(3):275–86. https://doi.org/10.1007/s40572-014-0017-9.
57. Wang B, Xu D, Jing Z, Liu D, Yan S, Wang Y. Mechanisms in endocrinology: effect of long-term exposure to air pollution on type 2 diabetes mellitus risk: a systemic review and meta-analysis of cohort studies. Eur J Endocrinol. 2014;171(5):R173-82. https://doi.org/10.1530/EJE-14-0365.
58. Liu F, Chen G, Huo W, Wang C, Liu S, Li N, et al. Associations between long-term exposure to ambient air pollution and risk of type 2 diabetes mellitus: a systematic review and meta-analysis. Environ Pollut. 2019;252:1235–45. https://doi.org/10.1016/j.envpol.2019.06.033.
59. Lao XQ, Guo C, Ly C, Bo Y, Zhang Z, Chuang YC, et al. Long-term exposure to ambient fine particulate matter (PM2. 5) and incident type 2 diabetes: a longitudinal cohort study. Diabetologia. 2019;62(5):759–69. https://doi.org/10.1007/s00125-019-4825-1.
60. Hwang MJ, Kim JH, Koo YS, Yun HY, Cheong HK. Impacts of ambient air pollution on glucose metabolism in Korean adults: a Korea National Health and Nutrition Examination Survey study. Environ Health. 2020;19(1):70. https://doi.org/10.1186/s12940-020-00623-9.
61. Shin MK, Kim KN. Association between long-term air pollution exposure and development of diabetes among community-dwelling adults: modification of the associations by dietary nutrients. Environ Int. 2023;174:107908. https://doi.org/10.1016/j.envint.2023.107908.
62. Kim MJ, Lim NK, Choi SJ, Park HY. Hypertension is an independent risk factor for type 2 diabetes: the Korean Genome and Epidemiology Study. Hypertens Res. 2015;38(11):783–9. https://doi.org/10.1038/hr.2015.72.
63. Hwang J, Shon C. Relationship between socioeconomic status and type 2 diabetes: results from Korea National Health and Nutrition Examination Survey (KNHANES) 2010–2012. BMJ Open. 2014;4(8):e005710. https://doi.org/10.1136/bmjopen-2014-005710.
64. Nam GE, Kim W, Han K, Jung JH, Han B, Kim J, et al. Association between living alone and incident type 2 diabetes among middle-aged individuals in Korea: a nationwide cohort study. Sci Rep. 2021;11:3659. https://doi.org/10.1038/s41598-021-82868-z.
65. Kolb H, Martin S. Environmental/lifestyle factors in the pathogenesis and prevention of type 2 diabetes. BMC Med. 2017;15(1):131. https://doi.org/10.1186/s12916-017-0901-x.
66. Zheng Y, Ley SH, Hu FB. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. Nat Rev Endocrinol. 2018;14(2):88–98. https://doi.org/10.1038/nrendo.2017.151.
67. Cloostermans L, Wendel-Vos W, Gea D. Independent and combined effects of physical activity and body mass index on the development of type 2 diabetes – a meta-analysis of 9 prospective cohort studies. Int J Behav Nutr Phys Act. 2015;12:147. https://doi.org/10.1186/s12966-015-0304-3.
68. Speizer FE, Cohen A, Asia MS. Changing times and changing problems. 2008. https://doi.org/10.1289/ehp.11856.
69. Yang Z, Mahendran R, Yu P, Xu R, Yu W, Godellawattage S, et al. Health effects of long-term exposure to ambient PM2. 5 in Asia-Pacific: a systematic review of cohort studies. Curr Environ Health Rep. 2022;9(2):130–51. https://doi.org/10.1007/s40572-022-00344-w.
70. Paoin K, Ueda K, Ingviya T, Buya S, Phosri A, Seposo XT, et al. Long-term air pollution exposure and self-reported morbidity: a longitudinal analysis from the Thai cohort study (TCS). Environ Res. 2021;192:110330. https://doi.org/10.1016/j.envres.2020.110330.
71. Downward G, Hystad P, Tasmin S, Abe S, Saito E, Rahman M, et al. Long-term exposure to particulate matter and all-cause and cause-specific mortality in an analysis of multiple Asian cohorts. Environ Int. 2024;189:108803. https://doi.org/10.1016/j.envint.2024.108803.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.