A Bayesian network for model for predicting cardiovascular risk

ICMAT DataLab

ICMAT INSTITUTO DE CIENCIAS MATEMÁTICAS

A little bit of context ...

In Europe:

- Cardiovascular diseases (CVD) are leading death cause in Europe
  - 3.9 millions deaths per year
  - 45% of all deaths
- Annual CVD treatment > 210 billion €

IMPORTANT: Cardiovascular risk prediction for CVD management and control

Research methodology and objectives

Bayesian network (BN) implementation

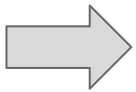
- Modifiable and non-modifiable cardiovascular risk factors (CVRFs) and medical conditions
- Large dataset + expert knowledge to build underlying model

BN model provides:

- Interpretable inference and prediction on CVRFs
- Decision-support tool to suggest diagnosis, treatment, policy, and research actions

Data collection and preprocessing

- Annual health (2012 - 2016) assessments from insurance company.
- Complemented with census information to infer socioeconomic status and education level.
- Removal of outliers, duplicates, misrecorded and missing values.
- Retain the most recent assessment of each individual.



Final dataset contains 205,087 health assessments.

Relevant variables

CVRFs = Cardiovascular risk factors

4 Non-modifiable CVRFs

6 modifiable CVRFs

4 Medical conditions

Table 1
Variables in model.

Variable	Definition	Levels
v_1	Sex	{*female, male*}
v_2	Age	(24,34], (34,44], (44,54], (54,64], (64,74]
v_3	Education level	{1,2,3}
v_4	Socioeconomic status	{1,2,3}
v_5	Body mass index	{*underw., normal, overw., obese*}
v_6	Physical activity	{*insufficiently active (1), regularly active (2)*}
v_7	Sleep duration	{*short, normal, excessive*}
v_8	Smoker profile	{*non-smoker, ex-smoker, smoker*}
v_9	Anxiety	{*yes, no*}
v_{10}	Depression	{*yes, no*}
v_{11}	Hypertension	{*yes, no*}
v_{12}	Hypercholesterolemia	{*yes, no*}
v_{13}	Diabetes	{*yes, no*}

Marginal distributions of each variable

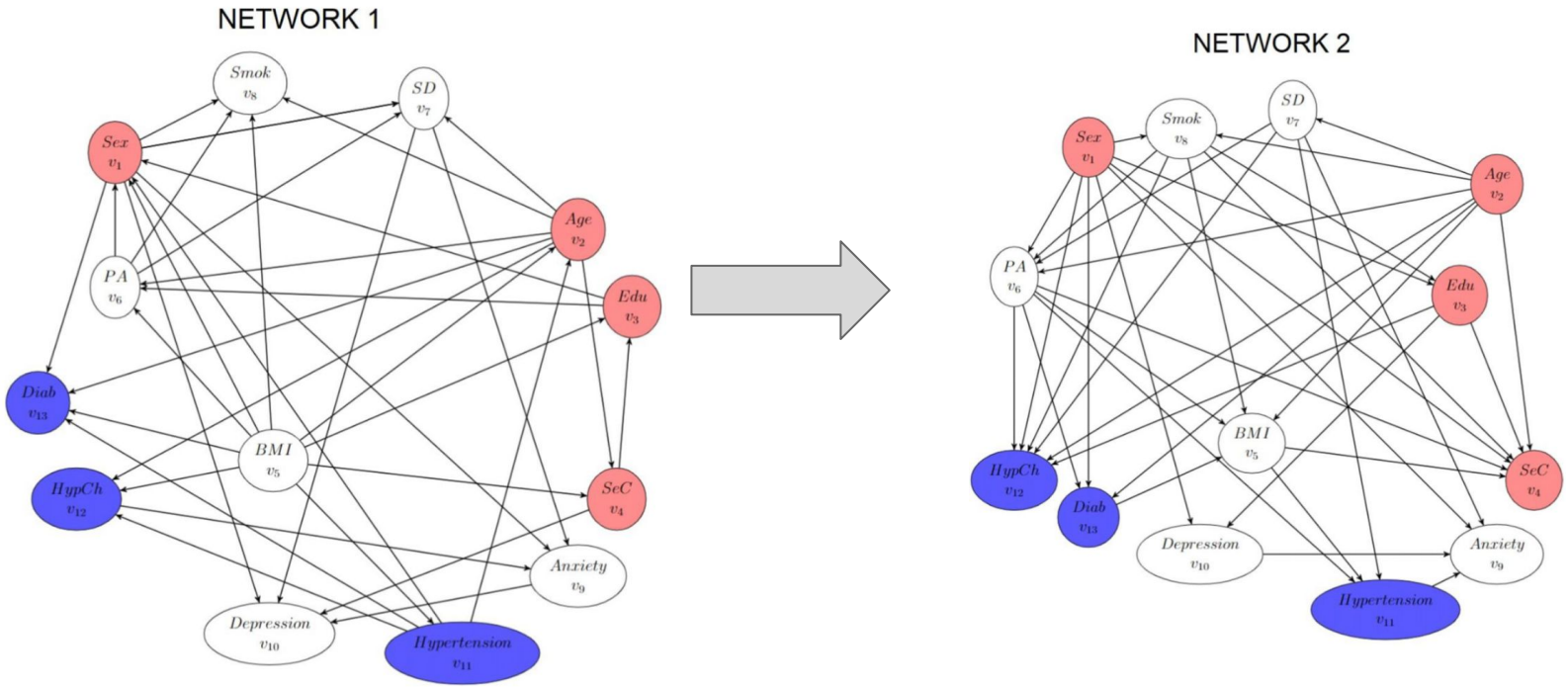
Reflect standard structure of Spanish labors markets (with few exceptions).

Table 2
Percentage of observations at each class.

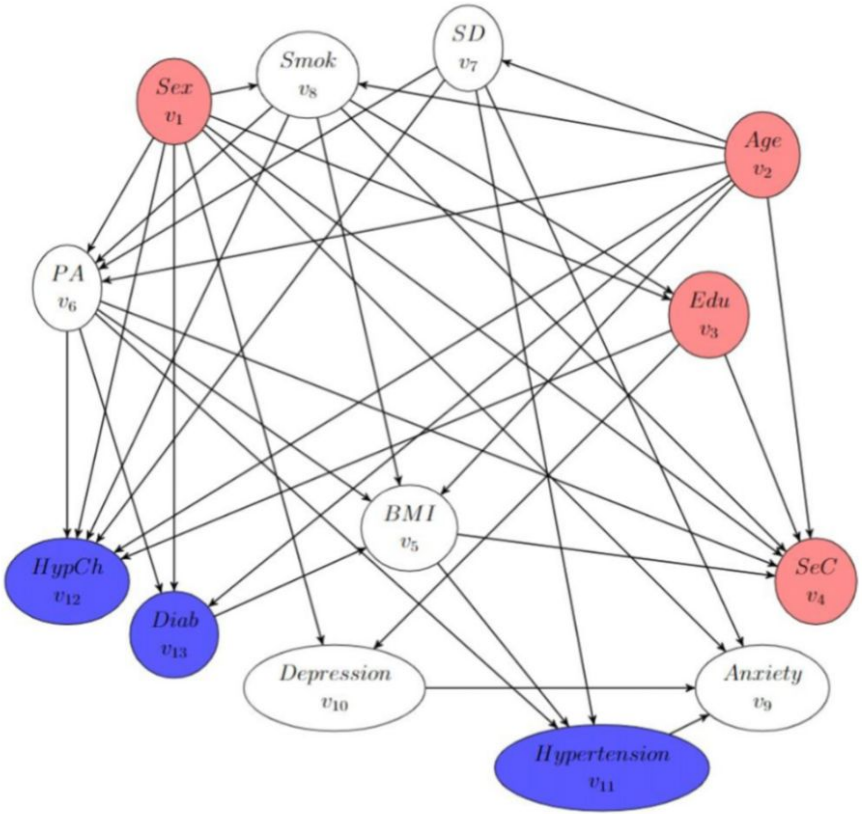
Variable	States	Marginal	Variable	States	Marginal
BMI	Underweight	1.16%	Diabetes	Yes	2.51%
	Normal	40.74%		No	97.49%
	Overweight	40.31%	Hypertension	Yes	15.05%
	Obese	17.79%		No	84.95%
Sex	Female	32.10%	Hypercholest.	Yes	30.53%
	Male	67.90%		No	69.47%
Smoker	Non-Smoker	50.14%	Physical Act.	1	75.63%
profile	Ex-Smoker	20.49%		2	24.37%
	Smoker	29.37%			
Age(y)	[18,24]	0.71%	Educ. lev.	1	0.21%
	(24,34]	17.10%		2	76.65%
	(34,44]	36.81%		3	23.14%
	(44,54]	30.37%	Sleep Dur.	< 6h	11.66%
	(54,64]	14.83%		(6h-9 h)	88.25%
	(64,74]	0.18%		> 9h	0.09%
Socioeconomic	1	37.83%	Depression	No	99.52%
status	2	35.15%		Yes	0.48%
	3	27.02%			
Anxiety	Yes	2.68%			
	No	97.32%			

Bayesian network construction

- Network 1: Greedy Thick Thinning (GTT)
- Network 2: From network 1, expert modifications
  - 15 edges added and 7 reversed



Bayesian network



$$
\begin{aligned}
p(v_1,\ldots,v_{13}) = {} & [p(v_1)p(v_2)p(v_3 \mid v_1, v_8)p \\
& (v_4 \mid v_1, v_2, v_3, v_5, v_6, v_8)] \\
& \times [p(v_5 \mid v_2, v_6, v_8)p(v_6 \mid v_1, v_2, v_7, v_8)p(v_7 \mid v_2) \\
& p(v_8 \mid v_1, v_2)p(v_9 \mid v_1, v_7, v_{10}, v_{11})p(v_{10} \mid v_1, v_3)] \\
& \times [p(v_{11} \mid v_5, v_6, v_7)p(v_{12} \mid v_1, v_2, v_3, v_6, v_7, v_8) \\
& p(v_{13} \mid v_1, v_2, v_6)],
\end{aligned}
$$

Diagnosis and evidence propagation

- Individual/ set of individuals with Age>= 45, BMI = Overweight, SD>= 6 and Anxiety = Yes

$$\Pr(v_{11} = y \mid v_1 = \text{male}, v_2 \geq 45, v_5 = \text{overw.}, v_6 = 1, v_7 =< 6h), v_9 = y)$$

$$= \frac{\Pr(v_1 = \text{male}, v_2 \geq 45, v_5 = \text{overw.}, v_6 = 1, v_7 =< 6h), v_9 = y, v_{11} = y)}{\Pr(v_1 = \text{male}, v_2 \geq 45, v_5 = \text{overw.}, v_6 = 1, v_7 =< 6h), v_9 = y)} =$$

25.26 % > 15.05% (marginal probability)

- Individual should be informed of a high probability of having hypertension.

Table 4
Probability of developing hypertension given various patient conditions for age greater than 44, poor sleeping level and anxiety.

BMI	Physical activity	Probability Male	Probability Female
Overw.	1	25.26	26.34
Overw.	2	19.79	20.70
Obese	1	45.54	46.95
Obese	2	34.49	35.78
Overw., obese	1	32.85	33.82
Overw., obese	2	22.90	23.85

- Positive impact of PA .

Health research through hypothesizing evidence

- Hypothesize evidence to address various research issues

Table 6
Probability of BMI population structure given socioeconomic status.

Status	Underweight	Normal	Overweight	Obese
3	1.18	42.53	39.63	16.66
2	1.22	41.54	40.09	17.15
1	1.10	38.70	41.01	19.19

  - Slightly worse BMI structure for lower-income population

- Conclusions should be ascertained through hypothesis testing.

Health research through hypothesizing evidence

- Impact of socio-economic status on health conditions
  - Slightly increase of hypertension as SeC decreases
  - Higher prevalence of HypCh and Diabetes with lower SeC
  - Slight decrease of depression with higher SeC
  - Decrease of anxiety with lower SeC

Table 5
Probability of health conditions given socioeconomic status.

Status	Anxiety	Depression	Diabetes	Hypercholesterolemia	Hypertension
3	2.75	0.45	2.38	29.71	14.65
2	2.70	0.49	2.49	30.61	14.90
1	2.61	0.49	2.62	31.03	15.48

Therapies through influential findings

Individual s.t. Sex = Male, Age = (44, 54], Edu = 3, SeC = 3, BMI = obese, PA = inactive, Smok = non smoker, SD = short, Anxiety = yes, Depression = No

- Prob. of developing Hypertension = 45.63%.

Table 7
Probability of hypertension given improved conditions for a male aged 44–54, education and socioeconomic status 3, obese, low PA, insufficient sleep, with anxiety but no depression.

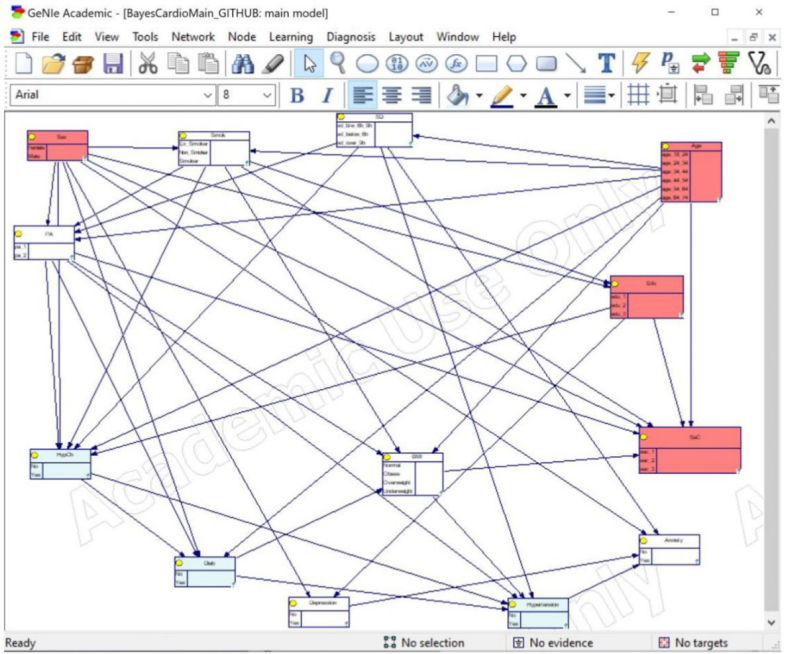
MCVRF	Level	Probability
BMI	Normal	11.30
Physical activity	Regularly active	34.57
Sleep	Normal	39.69
Anxiety	No	37.02

- Priority should be to improve BMI.
- If all the MCRF are improved, prob. decreases to 4.80.

Software

- GeNie model (Academic use)

  https://datalab-icmat.github.io/software.html



Discussion

- BN construction based on large dataset and expert knowledge.
- Decision support tool for public health, policy, and diagnosis.
- Combine with utility function to select best recommendations.

- Some limitations:
    - The dataset has different structure to Spanish population (Healthy worker effect)
    - Some data were self-reported
    - No explicit data concerning diet
    - Only predictive claims, no causal

More information on …

Ordovas, J. M., Insua, D. R., Santos-Lozano, A., Lucia, A., Torres, A., Kosgodagan, A., & Camacho, J. M. (2023). A Bayesian network model for predicting cardiovascular risk. *Computer Methods and Programs in Biomedicine*, 107405