

Contents lists available at [ScienceDirect](https://www.sciencedirect.com)

Computer Methods and Programs in Biomedicine

journal homepage: <https://www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine>



Colorectal cancer risk mapping through Bayesian networks

D. Corrales^{a,*}, A. Santos-Lozano^{b,c}, S. López-Ortiz^c, A. Lucia^{b,d}, D. Ríos Insua^a

^a Inst. Math. Sciences, CSIC, 28049 Madrid, Spain

^b Research Institute of Hospital 12 de Octubre ('imas12'), 28041 Madrid, Spain

^c i+HeALTH Strategic Research Group, Miguel de Cervantes European University, 47012 Valladolid, Spain

^d Faculty of Sport Sciences, Universidad Europea de Madrid, Spain

ARTICLE INFO

Keywords:

Colorectal cancer

Bayesian network

Risk mapping

Modifiable risk factors

Health policy

ABSTRACT

Background and Objective: Only about 14% of eligible EU citizens finally participate in colorectal cancer (CRC) screening programs despite it being the third most common type of cancer worldwide. The development of CRC risk models can enable predictions to be embedded in decision-support tools facilitating CRC screening and treatment recommendations. This paper develops a predictive model that aids in characterizing CRC risk groups and assessing the influence of a variety of risk factors on the population.

Methods: A CRC Bayesian Network is learnt by aggregating extensive expert knowledge and data from an observational study and making use of structure learning algorithms to model the relations between variables. The network is then parametrised to characterize these relations in terms of local probability distributions at each of the nodes. It is finally used to predict the risks of developing CRC together with the uncertainty around such predictions.

Results: A graphical CRC risk mapping tool is developed from the model and used to segment the population into risk subgroups according to variables of interest. Furthermore, the network provides insights on the predictive influence of modifiable risk factors such as alcohol consumption and smoking, and medical conditions such as diabetes or hypertension linked to lifestyles that potentially have an impact on an increased risk of developing CRC.

Conclusion: CRC is most commonly developed in older individuals. However, some modifiable behavioral factors seem to have a strong predictive influence on its potential risk of development. Modeling these effects facilitates identifying risk groups and targeting influential variables which are subsequently helpful in the design of screening and treatment programs.

1. Introduction

Colorectal cancer (CRC) is the third most common type of cancer worldwide, making up for about 10% of all cases [1] and being accountable for around 12% of all deaths due to cancer. In 2020, there were 1.9 million new cases and 930,000 associated deaths. It is more common in developed countries, where more than 65% of the cases are found. Despite this, as an example, only about 14% of susceptible EU citizens participate in screening programs, at the moment mostly based on fecal testing and colonoscopy. Hence, there is a need for accurate, non-invasive, cost-effective screening tests based on novel technologies and raise further awareness of the disease and its detection. Moreover, more personalized screening approaches are required to consider genetic and socioeconomic variables as well as environmental stressors that potentially lead to different onsets of the disease [2]. A particular line of action is the development of predictive models that facilitate

CRC predictions, the subject of this paper, possibly embedded in decision support tools that aid in the advice on screening and treatment recommendations.

The epidemiology of CRC and its most important risk factors (CR-CRF) are discussed, among others, in Marley and Nan [3] and Sawicki et al. [4]. These factors are defined as measurable characteristics associated with increased CRC incidence and considered to be significant independent predictors of increased risk of the disease. They are qualified as modifiable or not. Non-modifiable ones are factors over which the individual has no control, including genetics, age, or gender. In contrast, modifiable ones cover behavioral factors that can evolve through individual action, including physical activity (PA), or tobacco use. Most CRC development does not have a genetic burden, but is linked to lifestyle and environmental factors [4] and thus the identification of the impact of the modifiable factors in individuals is key to reducing CRC incidence.

* Corresponding author.

E-mail address: daniel.corrales@icmat.es (D. Corrales).

<https://doi.org/10.1016/j.cmpb.2024.108407>

Received 22 March 2024; Received in revised form 30 July 2024; Accepted 3 September 2024

Available online 6 September 2024

0169-2607/© 2024 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).

The purpose of this paper is to provide a Bayesian network (BN) [5,6] that facilitates the prediction of CRC risks and their mapping. The network will be built from extensive expert judgment and data and illustrated through two relevant use cases referring to CRC risk mapping and CRC influential finding identification; other uses will be sketched in the conclusion. Interest in BNs in the healthcare community has increased over the last decade as for diagnosis and prognosis BNs represent a natural framework to analyze dependence among risk factors. Furthermore, they can aggregate knowledge from experts, which is especially relevant in contexts in which data might be limited, and still provide meaningful and accurate decision support [7]. Relevant work in the field includes Wang et al. [8], who propose a BN model for cancer treatment assessment and development monitoring; Jang et al. [9], who use a BN model together with expert knowledge to analyze the disease burden of breast cancer and the risks and benefits of radiation therapy; and Liu et al. [10] who use BNs to analyze the most influential factors in breast cancer diagnosis. Regarding CRC, Myte et al. [11] build a BN to analyze the possible impact of one-carbon metabolites in relation to CRC, also considering genetic information and environmental factors in the study; Sieswerda et al. [12] leverage BN structure learning algorithms and expert knowledge to create causal models to estimate treatment effectiveness in colon cancer therapies; and Osong et al. [13] make use of BNs for predicting local tumor recurrence in rectal cancer patients after treatment and surgery.

In contrast, the approach proposed in this paper aims to build a representative probabilistic model of the interactions between several variables in a general population setting, including non-modifiable and modifiable risk factors, to analyze their influence in the development of CRC. Major advantages of BN models, that we shall draw upon, are their use for generative purposes and their ability to propagate the evidence along the network to obtain representative probabilities based on this evidence. Thus, the model built in this paper is intended to serve as a quantitative guideline for the CRC risk assessment of different segments of a population, as it manages to maintain representative proportions and imbalances of the different variables found in the data set. Hence, the conclusions reached through the model will be representative (at least for the population set taken into account) and actions taken could be modeled to obtain a meaningful approximation of their influence. Furthermore, the characterization of segments of the population with a higher risk of developing CRC would be of interest for screening purposes as targeting these groups would yield more cases per screening test performed and increase a screening program's effectiveness [14].

The rest of the paper is structured as follows. We first describe how the BN was built taking into account the data and knowledge available; this entails discovering the structure of the network and building the corresponding tables of probabilities. We next deal with two important use cases: the first one refers to building risk maps depending on key features of individuals; the second one, refers to reporting key factors in developing CRC. A final section summarizes results, discusses limitations, suggests additional use cases, and sketches future work. Importantly, for reasons outlined in this last section, we prevent from making causal claims for our BN and just pursue predictive claims as in Hernan and Robins [15]; Scutari and Denis [6] provide further insights regarding causality and BNs. For reproducibility purposes, software for the full model, as well as for the use cases presented, is available in https://github.com/DanielCorralesAlonso/CRC_Risk_BN.

2. Materials and methods

This section describes the process used to build our BN for CRC risk predictions. It is divided into five parts characterizing the work pipeline adopted: collection of available knowledge, data gathering and processing, network structure discovery, estimation of probabilities, and validation.

2.1. Materials

2.1.1. Prior available knowledge

The data used in this project were extracted from an observational study covering annual health assessments of adult workers affiliated with a private health insurance provider in Spain, from 2012 to 2016. After conveniently securitizing the data, they were enriched with census information from the Spanish National Statistics Institute (INE) based on postal code, allowing us to infer their socioeconomic status and educational level. This led to an initial dataset with about 2.4 million records and 66 variables.

In order to compile relevant knowledge about CRC, we performed exhaustive searches through scientific and medical databases with the expressions '*causal inference and CRC*'; '*probabilistic networks, Bayesian networks, influence diagrams and CRC*'; '*Data mining and CRC*'; '*Risk factors and CRC*'; '*CRC epidemiology*'; '*causes of CRC*'. We also queried ChatGPT with the prompts '*What are the risk factors in the development of CRC*' and '*What are the modifiable risk factors in the development of CRC*'. Additionally, relevant information from a previous network developed concerning cardiovascular disease (CVD) risk factors [16] was considered.

2.1.2. Available data

The list of relevant variables, together with the background information mined, was submitted to a team of expert clinicians who, through a consensus session, suggested to retain from the original database the fourteen variables presented in Table 6 in Appendix. They also grouped the variables as follows:

- Non-modifiable CRCRFs: *sex, age, and socioeconomic status.*
- Modifiable CRCRFs: *body mass index (BMI), physical activity (PA), sleep duration (SD), alcohol consumption, smoking profile, anxiety, and depression.*
- Medical conditions: *hypertension, hypercholesterolemia, and diabetes.*
- Target variable: *presence of CRC.*

An intensive exploratory data analysis focused on detecting outliers and misrecorded values, duplicates, and missing values.

In particular, for originally continuous unimodal approximately symmetrical variables around the mean, we considered the standard rule of treating as outliers those data points whose values were further from the marginal distribution's mean by three standard deviations [17], with 230,841 data points meeting these criteria. These were assumed to come from measuring or recording mistakes; we removed them from the training phase, assuming that model performance would not be affected.¹ As an example, the case of a patient whose record showed that their height was 160 cm and their weight 342 kg, was eliminated. We also discarded for training purposes 325,147 data points with a missing value in any variable, as given the size of the final dataset, we would have enough training data. Note that there was no evidence suggesting any missing not at random (MNAR) scenario which would have prevented us from discarding these data points.²

Finally, we retained a total of 1,778,270 health assessments which were split according to the date of the recording with the motivation of updating the parameters of our model every year based on information from previous years and reserving those of year 2016 for validation purposes.

Table 7 in the Appendix lists the proportion of cases in various marginal categories reflecting, by and large, the standard structure of the Spanish labor force. We performed this exploratory analysis for

¹ Importantly, they were not used for training purposes, but we used them for validation purposes in the sense of Section 2.2.3.

² Again we used them for validation purposes, Section 2.2.3.

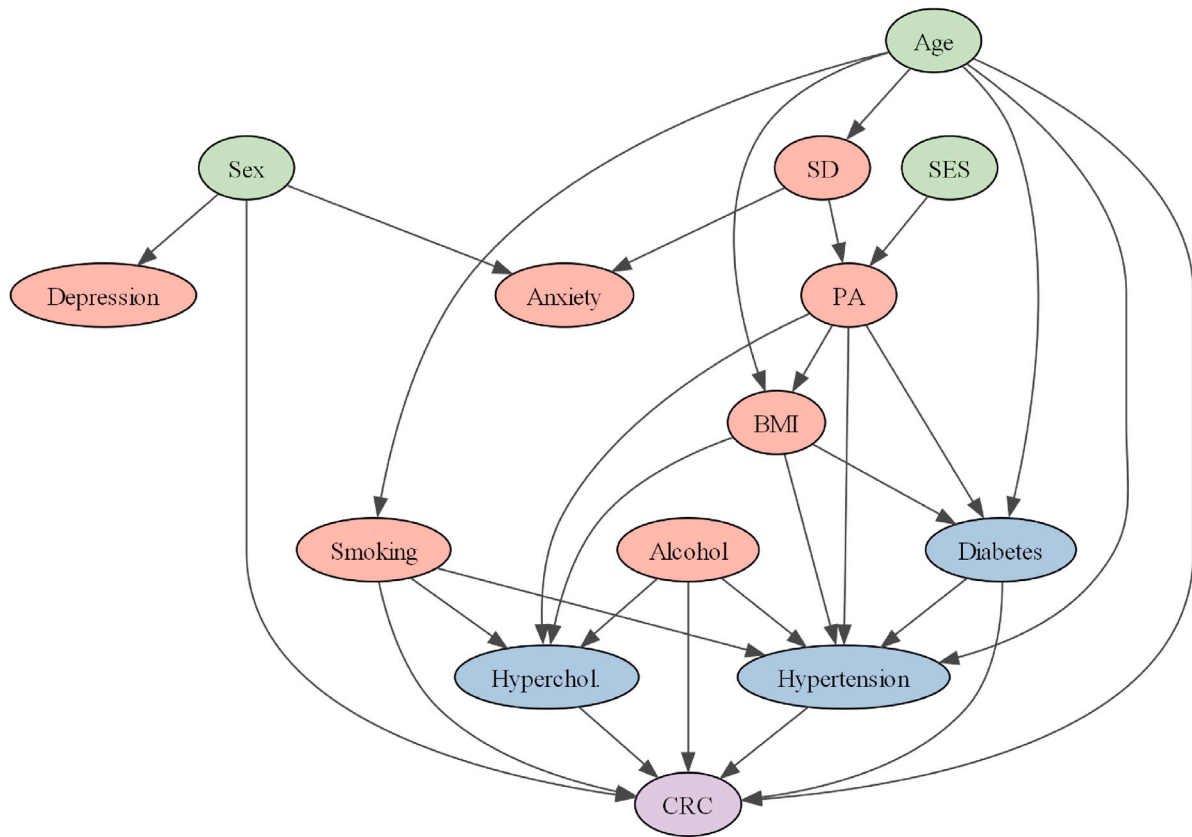


Fig. 1. Initial BN structure (network 1) coding knowledge available for CRC taking into account available variables. Forbidden arcs not included for clarity.

each of the years as an exploratory sensitivity analysis check, revealing just minor differences over the years.

Similarly, we explored the impact of spatial effects based on postcodes, finding no evidence of spatial correlations for all variables considered except the socioeconomic situation, in which spatial information is encoded by definition.

2.2. Methods

2.2.1. Structure discovery

Once the data was collected and processed we built a discrete BN to estimate the underlying joint distribution, which served as the basis to make inferences and predictions on CRC risk cases of interest. The selected variables were coded as described by Table 6 in Appendix. A two-stage procedure was used to learn the BN structure.

First, based on the information described in Sections 2.1.1 and 2.1.2, specially the causal suggestions from our medical experts, we obtained an initial description of the network describing proposed and forbidden arcs, summarizing their knowledge, as agreed with the team of experts. Fig. 1 provides the initial network where, to facilitate visualization, we do not include forbidden arcs. As an example, (Hypercholesterolemia, Age) would be a forbidden arc as the former cannot affect the latter in any possible way. Different color codes were used for the four types of variables mentioned above.

Such structure was used as the initial network to several structure discovery algorithms and software. There are numerous procedures available for the purpose of building a network based on relations in the data summarized e.g. in [18,19], who also mention related software solutions. In particular, we used the algorithms available in GeNIe Modeler [20], and the Python libraries *pyAgrum* [21] and *pgmpy* [22]. The solutions arrived at with various algorithms were analyzed by three experts in the CRC domain who revised the additional arcs reasoning in terms of plausible predictive relationships. This process

led to the final BN structure shown in Fig. 2 where new data-based arrows are displayed in red. To specifically obtain such a network, we employed the *greedy hill-climbing* algorithm, a local optimization algorithm that maximizes a predefined score at each step and adds an edge between nodes until the score cannot be maximized [23]. For our network structure discovery, we used the Bayesian Dirichlet sparse (BDs) score defined in Scutari [24] and implemented in *pgmpy*, as it is argued [25] that BDs seems to provide better accuracy in structure learning, specially with sparse data. As a consequence of the chosen graphical representation, the underlying suggested probabilistic model over the variables is characterized through the following expression:

$$\begin{aligned}
 & p(v_{sex}, \dots, v_{depression}) \\
 &= \left[p(v_{sex})p(v_{age})p(v_{SES}|v_{sex}, v_{age}) \right] \times \\
 & \left[p(v_{SD}|v_{sex}, v_{age})p(v_{PA}|v_{sex}, v_{age}, v_{SD}, v_{SES})p(v_{depr}|v_{sex}, v_{age}, v_{SES}) \right. \\
 & p(v_{smok}|v_{sex}, v_{age}, v_{PA})p(v_{alc}|v_{sex}, v_{age}, v_{smok}) \\
 & \left. p(v_{BMI}|v_{sex}, v_{age}, v_{PA}, v_{smok})p(v_{anx}|v_{sex}, v_{SD}, v_{smok}, v_{depr}) \right] \times \\
 & \left[p(v_{hypchol}|v_{sex}, v_{age}, v_{PA}, v_{smok}, v_{BMI}, v_{alc})p(v_{diab}|v_{sex}, v_{age}, v_{PA}, v_{BMI}) \right. \\
 & \left. p(v_{hyprien}|v_{age}, v_{PA}, v_{smok}, v_{BMI}, v_{alc}, v_{diab}) \right] \times \\
 & p(v_{CRC}|v_{sex}, v_{age}, v_{alc}, v_{smok}, v_{hypchol}, v_{hyprien}, v_{diab}) \quad (1)
 \end{aligned}$$

where, to facilitate reading and reasoning, we have separated the products into the four blocks of variables considered.

2.2.2. Probabilities discovery

Once with the structure, the next stage was to learn the associated probability tables drawing on the data D available. We estimated them using standard multinomial-Dirichlet models [23,26]. Let X be a network variable, U its parent variables, and u one of its instantiations. In general, if $p(\theta_{X|u})$ is a Dirichlet prior distribution with hyperparameters

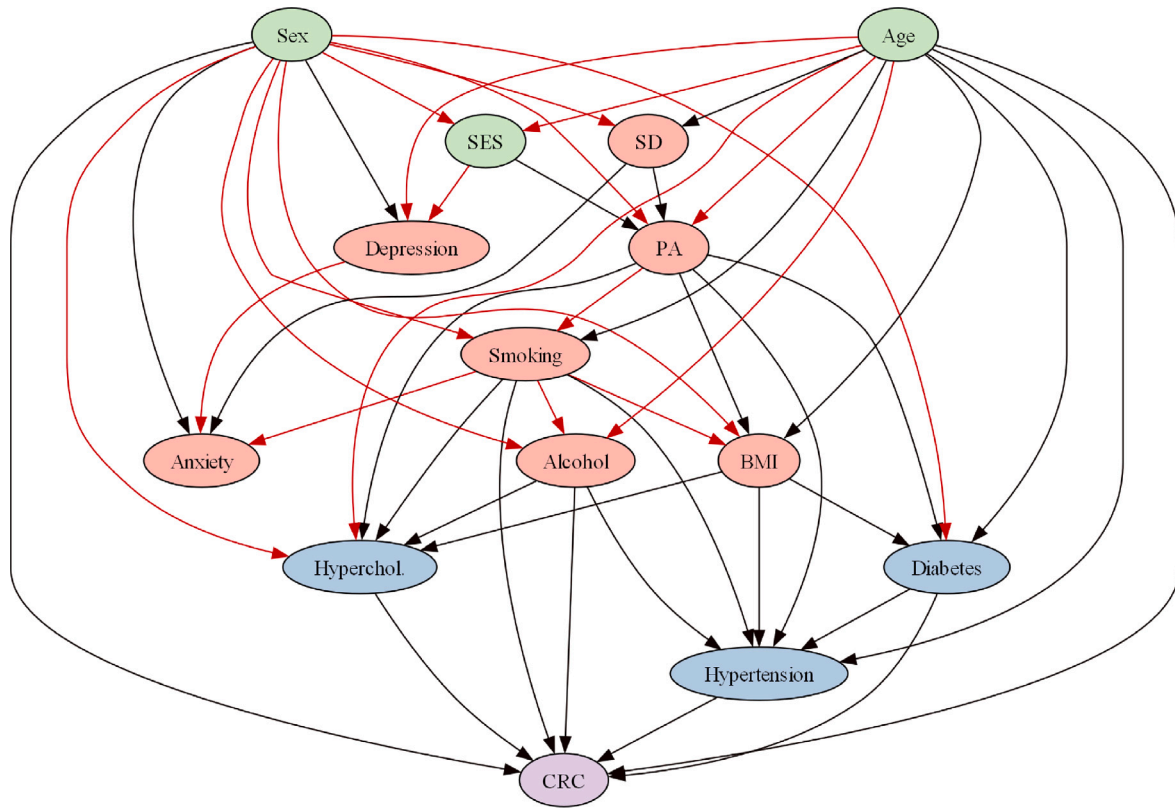


Fig. 2. Final BN structure (network 2) coding knowledge and data available for CRC taking into account relevant available variables and enhanced through the database.

$\alpha_{x^1|u}, \dots, \alpha_{x^K|u}$, the posterior $p(\theta_{X^i|u}|D)$ will be a Dirichlet distribution with hyperparameters $\alpha_{x^1|u} + m[u, x^1], \dots, \alpha_{x^K|u} + m[u, x^K]$, where $m[u, x^i]$ is the number of times that instance (u, x^i) appears in the dataset. In particular, the estimate of the parameter $\theta_{X=x^i|u}$ based on the posterior mean would be

$$\hat{\theta}_{X=x^i|u} = \frac{\alpha_{x^i|u} + m[u, x^i]}{\sum_i (\alpha_{x^i|u} + m[u, x^i])}$$

A potential problem with our BN structure is that, due to the many connections arriving at some of the nodes some of the columns in the tables might receive relatively little data. In particular, minority classes in highly imbalanced variables, e.g. the CRC positive class in the CRC node, are affected by this issue. In that case, the corresponding posterior distributions would essentially coincide with the priors, therefore demanding care in assessing such priors, notwithstanding the related problem of the large number of priors to be chosen for some of the variables considered in the model.

Uniform priors are largely used in scenarios where no prior information is available. One example is the prior defined for the Bayesian Dirichlet equivalent uniform (BDeu) score [27], which assumes complete ignorance about the parameters of the network and thus at each node each class has the same probability [28,29]. In the case of the prior for the BDs score used for structure learning, it follows an empirical Bayes approach by giving prior uniform probability to the classes that appear at least once in the dataset, and zero prior probability to the classes that do not appear in the dataset [24]. Still, in medical practice the lack of prior information is rare, and a carefully defined informative prior may be more meaningful than a uniform one. As a consequence, the following approach was employed to build the priors for estimating the parameters. Table 7 in Appendix provides the marginal empirical distributions for all variables, which we use as prior means for the corresponding conditionals, whatever the conditioning values are, as a means of characterizing prior knowledge. We multiply them by a factor α interpreted as the relative weight of the prior with

respect to the data in the calculation of the posterior distribution. After cross-validating [30,31] this parameter by trying several values in a grid, based on classification performance (Section 2.2.3) and quality of inference (Sections 3.1 and 3.2), α was set to the number of patients considered divided by 10000, that is $\alpha \approx 31.69$, as it entailed a reasonable influence of the prior knowledge in the above-mentioned cases with few data when a variable is conditioned by many others. Other α values were tried varying the denominator in powers of 10; some of their multiples resulted in poorer performance, in classification terms, in the extreme cases in which α was too small or too large; other intermediate candidate values resulted in more similar performance to the α selected. A limitation of this approach is that the parameter α will have a different impact on the parametrization of each variable depending on its skewness or uniformity. This has been further analyzed in the literature, see [32,33]. However, it simplifies the prior characterization by only determining a single free parameter.

This quantity α will determine the uncertainty for all probability distributions at all nodes. Then, as discussed, if there is sufficient data for each of the combinations of conditioning variables, the uncertainty for the distributions will be reduced and the posterior means will shift depending on the conditioning variables. For cases with less data, the posterior distributions will be more similar to each other but will entail a larger uncertainty of the approximation. This process is repeated over several years, from 2012 to 2015, using the posterior distribution of the previous year as the prior for the next one, appraising the value of data from previous years. As an example, Table 1 provides the prior means for the distribution of the variable *SD* conditional on its two antecedents, *Age* and *Sex*, based on the marginals in Table 7 for each of the three categories Short (S), Normal (N), and Excessive (E), which, as described above, coincide.

Tables 2 and 3 respectively provide the posterior conditional mean and 0.9 posterior predictive intervals after processing the data from year 2012. Observe that there has been a reasonable change in the posterior conditional probability table when compared with the prior

Table 1
Prior mean probability for SD given Sex and Age.

| | [24, 34] | | [34, 44] | | [44, 54] | | [54, 64] | |
|-----------|----------|--------|----------|--------|----------|--------|----------|--------|
| | Man | Woman | Man | Woman | Man | Woman | Man | Woman |
| Short | 0.1024 | 0.1024 | 0.1024 | 0.1024 | 0.1024 | 0.1024 | 0.1024 | 0.1024 |
| Normal | 0.8963 | 0.8963 | 0.8963 | 0.8963 | 0.8963 | 0.8963 | 0.8963 | 0.8963 |
| Excessive | 0.0011 | 0.0011 | 0.0011 | 0.0011 | 0.0011 | 0.0011 | 0.0011 | 0.0011 |

Table 2
Posterior mean probability for SD given Sex and Age in 2012.

| | [24, 34] | | [34, 44] | | [44, 54] | | [54, 64] | |
|-----------|----------|--------|----------|--------|----------|--------|----------|--------|
| | Man | Woman | Man | Woman | Man | Woman | Man | Woman |
| Short | 0.0600 | 0.0711 | 0.0897 | 0.1039 | 0.1211 | 0.1581 | 0.1386 | 0.2256 |
| Normal | 0.9384 | 0.9264 | 0.9092 | 0.8952 | 0.8778 | 0.8407 | 0.8604 | 0.7737 |
| Excessive | 0.0016 | 0.0025 | 0.0011 | 0.0009 | 0.0012 | 0.0012 | 0.0010 | 0.0007 |

table, effectively addressing the differences among the states and the conditional states of the variables considered.

Table 4 provides the evolution of the SD probability distribution for a man in the age range [24, 34] after processing the data from years 2012 to 2015, displaying the mean and the 0.9 posterior predictive intervals for each year. The aforementioned change in the posterior probabilities is more subtle after 2012 as the prior information of the previous years was already informative. Nevertheless, this approach is able to detect subtle changes in the distribution over the years which can be highly useful in certain contexts.

The proposed method seeks to address the aforementioned problem related to the priors through the implementation of informative and representative priors, which in the cases where none or few data are collected avoids assessing uniform probabilities to combinations of variables that are so rare that might not even appear in the data. A uniform probability prior in this scenario would represent exactly the opposite of what we infer from the data as would characterize these combinations of variables with around a probability of $1/k$ (for k -valued categorical variables) in the conditional distributions of the model, resulting in a poor and misleading probability assessment. By acknowledging these situations, we reduce the uncertainty surrounding the less frequent values and we shall better characterize the risk assessments to be performed.

2.2.3. Validation through classification

Once the model has been built, and before illustrating relevant use cases in Sections 3.1 and 3.2, a core issue is to validate it. A natural way to do it in a probabilistic setting, see e.g. [30,31,34], is to conceive the network as a classifier and assess its performance over various nodes with a number of classification metrics. We undertook extensively this approach suggesting good results.

Let us illustrate the process with two variables, CRC and Diabetes. In the first case, we set CRC as the target variable that we would like to classify using the available instances for 2016, and those related to outliers and missing values. Note that the problem we are dealing with in this case is a highly imbalanced problem (1:1500 approx) which entails a major challenge for classifiers [35]. As an example, using the BN built, we classify the data set for the 2016 patients and aim to maximize the *G-mean*, the root of the product between sensitivity and specificity [36]. Recall that the major interest will be to detect as many CRC positives as possible without falsely classifying CRC negatives as positives. Table 5a presents the confusion matrix achieved in the classification of the whole data set, achieving a sensitivity of 0.68 and a specificity of 0.72. The corresponding AUC score is 0.76, which, incidentally, surpasses the values reported by other CRC studies with similar datasets, population imbalance characteristics, and calibration results [37]. In the case of diabetes, the classification problem is much less imbalanced (1:30). Table 5b provides the confusion matrix achieved, with a sensitivity of 0.73 and a specificity of 0.76.

Besides the usual classification metrics, we paid special attention to their calibration, in line with recent discussions in the medical literature [38]. This is of vital importance in risk prediction models as it has a great impact on the usefulness of these decision-support aspects. Fig. 3 displays the calibration curves obtained through quantile binning for the cases of Diabetes and CRC over the relevant ranges for both diseases.³ Quantile binning [39] creates bins with an equal number of samples based on the distribution of the data instead of bins with equal width. Thus, the number of predictions is larger on the lower end of the distribution in the cases of imbalanced data and fewer predictions are made on the upper end of the distribution. The resulting curves suggest a good calibration with a slight tendency to overestimate in the final relevant bins.

3. Results

3.1. Use case: CRC risk mapping

Once the BN has been built, parametrised, and validated, we proceed to exploit some of its properties and functionalities. The first use case for our model is the production of risk maps or tables that reflect the risk of a person suffering CRC assuming certain conditions c (e.g., this person is a man who is a smoker) as other features b vary (e.g., his age and drinking status) of interest. The motivation behind this use case is the well-evidenced assumption that different conditions have non-identical CRC effects in distinct segments of the population [3]. Furthermore, eventual tendencies could be broadly characterized through the use of risk maps.

The basic ingredient for the design of this tool would be the probabilities $p(CRC|c, b, q)$ of a person having CRC given that it has features b and c , as b adopts values in a set B , when q are the parameter values adopted for the probability tables, which are computed from the BN model with standard Bayesian computations [26]. To facilitate interpretation, we perform a comparison against the baseline of not having the information b , computing the differences in log probabilities

$$r(b, q) = \log(p(CRC|c, b, q)) - \log(p(CRC|c, q)),$$

and display graphically such quantities as a function of b . Recall though that we have uncertainty about q and thus we have to reflect it, for example through an interval $i(b) = [lr(b, q), ur(b, q)]$ of high posterior predictive probability for $r(b, q)$. For that, an iterative sampling approach is followed to generate posterior predictive estimates for the probabilities of interest. The uncertainty is then reflected through the, e.g., 0.9 posterior predictive interval of the desired quantity and, essentially, we would declare that if:

- $0 \in i(b)$ there is no sufficient evidence for an increase in risk with respect to the baseline;
- $0 < lr(b, q)$, there is an increase in risk; and, finally,
- $0 > ur(b, q)$ there is a reduction in risk.

After several design and visualization tests, we decided to display the risk maps as follows:

- Condition b would refer to one or two criteria, leading to uni- or bi-dimensional risk maps.
- We use $r(b, \hat{q})$ as reference for graphical purposes, where \hat{q} is the posterior mean of q , but additionally include $i(b)$.
- A color scheme based on $r(b, \hat{q})$ is used and displayed together with the whole interval $i(b)$. We avoid colors typically used in risk matrices [40] (red, yellow, green) to mitigate cultural biases.
- The size of the representation associated with the variation of risk in the segment b should reflect the size of the corresponding population.

³ The empirical marginal of diabetes in 2016 is 0.0336 and that of CRC is 0.00064.

Table 3
0.9 posterior predictive interval for SD probability given Sex and Age after processing 2012 data.

| | [24, 34] | | [34, 44] | | [44, 54] | | [54, 64] | |
|---|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| | Man | Woman | Man | Woman | Man | Woman | Man | Woman |
| S | [.0583, .0617] | [.0686, .0737] | [.0881, .0914] | [.1013, .1064] | [.1189, .1232] | [.1543, .1619] | [.1347, .1425] | [.2175, .2338] |
| N | [.9367, .9401] | [.9238, .929] | [.9076, .9108] | [.8926, .8978] | [.8756, .88] | [.8369, .8445] | [.8565, .8643] | [.7654, .7818] |
| E | [.0013, .0019] | [.002, .003] | [.0009, .0013] | [.0007, .0012] | [.0009, .0014] | [.0008, .0015] | [.0007, .0014] | [.0003, .0013] |

Table 4
Evolution of the SD probability distribution for a man in the age range [24–34] over years 2012–2015.

| | Prior | 2012 | 2013 | | 2014 | | 2015 | | |
|-----------|-------|-------|----------------|-------|----------------|-------|----------------|-------|----------------|
| Short | .1024 | .0600 | [.0583, .0617] | .0600 | [.0581, .0613] | .0595 | [.0579, .0612] | .0608 | [.0591, .0626] |
| Normal | .8963 | .9384 | [.9367, .9401] | .9388 | [.9372, .9404] | .9389 | [.9373, .9406] | .9378 | [.9360, .9396] |
| Excessive | .0011 | .0016 | [.0013, .0019] | .0015 | [.0013, .0018] | .0015 | [.0013, .0018] | .0013 | [.0011, .0016] |

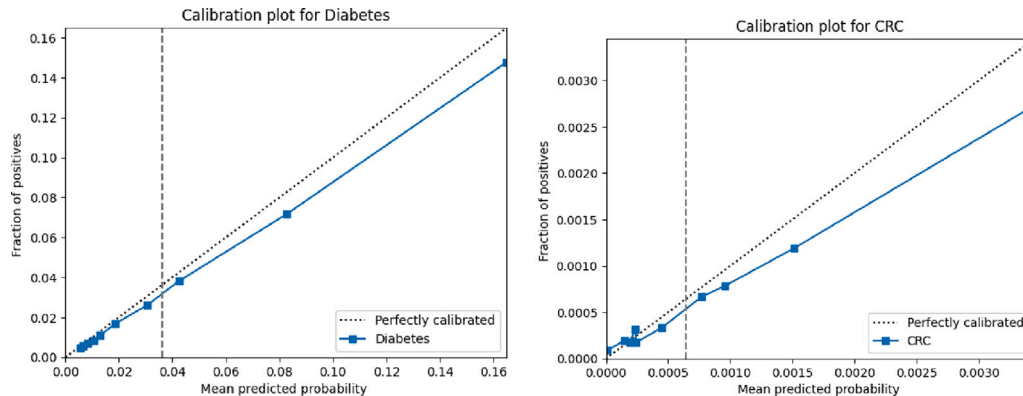


Fig. 3. Calibration curves for diabetes and CRC.

Table 5
Confusion Tables for BN validation.

| (a) CRC confusion table | | | | (b) Diabetes confusion table | | | |
|-------------------------|-------------|--------|--------------|------------------------------|--------|--|--|
| | Pred. label | | | Pred. label | | | |
| | 0 | 1 | | 0 | 1 | | |
| True label 0 | 243 326 | 96 163 | True label 0 | 249 937 | 78 361 | | |
| True label 1 | 70 | 148 | True label 1 | 3118 | 8291 | | |

predictive intervals. Therefore, we would conclude that SD is not a variable that fundamentally increases the risk of CRC on its own. Observe that the normal SD group is the largest one, followed by a smaller group with shorter SD. Note also that the smaller the population group, the larger the uncertainty as shown by the lower and upper bounds of the reported 0.9 posterior predictive intervals. Δ

Example 2. Fig. 5 provides a risk map when $c = man$, taking into account that $b = (Age, BMI)$ with age varying in the x -axis and BMI in the y -axis. Thus, the reference probabilities are $p(CRC|man, (Age, BMI), q)$.

Observe that CRC risk increases as both BMI and Age increase. However, age is the variable that has a larger impact, as colors are more similar column-than row-wise. We state that there is a smaller risk of CRC development with respect to the baseline for patients with ages lower than 44 and a bigger risk for patients older than 54.

In turn, Fig. 6 provides a risk map for $c = man$ taking into account $b = (BMI, Alcohol)$ with BMI in x -axis and $Alcohol$ in y -axis, with reference probabilities defined through $p(CRC|man, (BMI, Alcohol), q)$.

In this case, higher alcohol consumption always induces an increased CRC risk which accentuates greatly with age. Moreover, alcohol consumption seems to influence CRC risk more than BMI. Δ

3.2. Use case: influential findings

Risk maps provide visual comparisons of population groups in terms of different risk factors. An additional useful approach to the analysis of the factors potentially affecting the development of CRC would be to examine the variables that had the largest impact on patients diagnosed with CRC. In line with Section 3.1 and earlier work in determining influential findings in BNs, e.g. [41], we propose an approach to characterize the predictive power of each class and variable in the network. In our analysis, the variables will be modified independently among all the possible values for each risk factor and the difference in risk will be

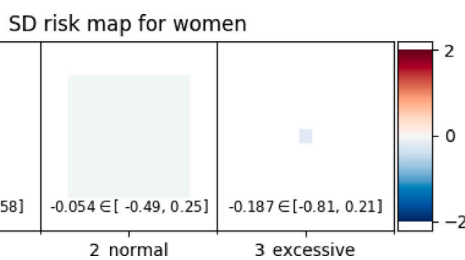


Fig. 4. Risk map for sleep duration (SD) for women.

We provide now several examples of risk maps based on the previous guidelines.

Example 1. The first example, Fig. 4, provides a risk map when $c = woman$, taking into account $b = (SD)$ reflected in the x -axis, that is, we want to display the CRC risk variation depending on the sleep duration (short, normal, excessive) in women. Therefore, the reference probabilities are $p(CRC|woman, SD, q)$.

In this case, shorter sleep duration seems to be related to an increase in CRC risk as shown by the point-wise estimations reflected in the colors and the first quantity in each of the cells. However, the interval estimates do not confirm this finding as 0 belongs to all the 0.9 posterior

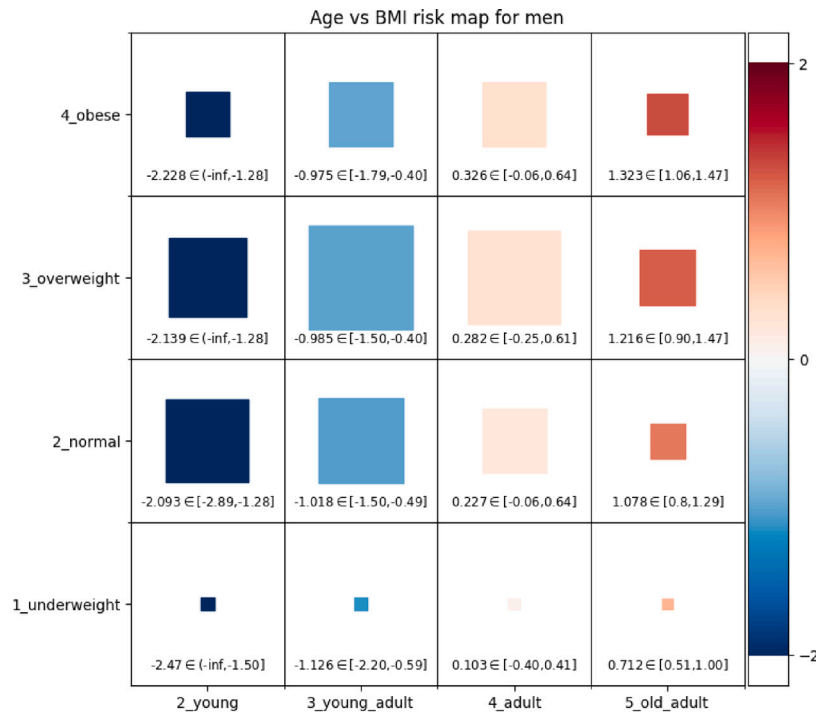


Fig. 5. Risk map for Age and BMI for men.

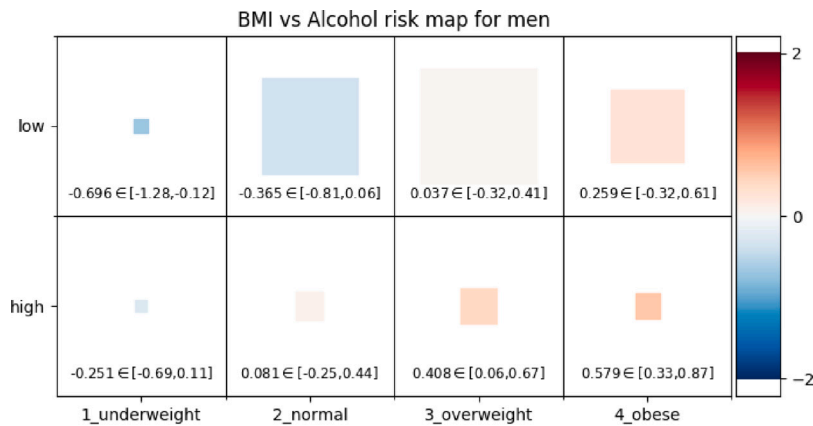


Fig. 6. Risk map for BMI and alcohol for men.

assessed. Repeating this with all CRC-positive patients in the database, we obtain an estimation of the strength of the predictive influence for each of the risk factors. As mentioned in the introduction, it is important, though, to remark that the influence of the variables depends on the model’s graphical structure, and any causality claim should be carefully analyzed before taking it for granted, see our final discussion. This prevents us from employing standard causal evaluations of effect sizes through interventions/do-calculus or counterfactuals.

In detail, we proceed as follows, where Algorithm 1 summarizes the method used. First, the entire information of each CRC-positive patient is recovered from the database. The order of the evidence available for a patient is randomized and set variable by variable. At each step, the

relative risk variation is calculated, which is quantified as the relative change in the difference of logarithms of the mean probabilities of developing CRC conditioned on the added evidence, similarly to the approach in Section 3.1. That is,

$$RRV(i, j) = \frac{\log(p_{model}(CRC|ev_j)) - \log(p_{model}(CRC|ev_{j-1}))}{\log(p_{model}(CRC|ev_{j-1}))} \times 100,$$

where $RRV(i, j)$ refers to the relative risk variation for patient i and variable j , and ev_j represents the values of the first j conditioning variables.

The reason for randomizing the evidence is that, when the evidence of the parents of the target node is fully set, the remaining variables

have no effect on the target node as the entire probability distribution is determined by the parents of such node, due to the local Markov property [23]. Thus, the order in which the evidence is set may have an impact on how certain variables seem to influence the prediction on the model target. Recording the relative variations in probability corresponding to the set of new evidence for each variable will assess the relative impact of the variable instance in the determination of the final probability. Randomizing the order of the evidence and repeating the process several times would provide a better understanding of the predictive influence of all the variables on the target node.

Algorithm 1: Pseudo code to determine influential findings

```

Data: Dataset, model, target
Result: diff_vect
for n iterations do
  for row in rowsDataset do
    evidence = Dataset[row,:] #Take variable information as evidence.
    p_model(target|evidence)
    shuffled_evid = random.shuffle(evidence)
    for j ← 1 to len(shuffled_evid) do
      partial_evid = Dataset[row, shuffled_evid[0:j-1]]
      new_evid = Dataset[row, shuffled_evid[j]]
      relative_risk_variation[row, j] =
        (log(p_model(target|partial_evid+new_evid)) - log(p_model(target|partial_evid))) /
        log(p_model(target|partial_evid))
        × 100
    end
  end
  Average along the data set rows
end
Average along all iterations

```

Fig. 7 reflects an average of the positive and negative predictive influence that different variables have on the risk of developing CRC. The standard deviations of the predictions are also provided. Our conclusions seem to agree with GBD 2019 Colorectal Cancer Collaborators [42] and Marley and Nan [3], which state that countries in Western Europe are prone to an increased consumption of alcohol and tobacco that highly contributes to CRC DALYs (Disability Adjusted Life Years). Furthermore, high fasting plasma glucose is one of the major contributors to CRC DALYs in Western European women and our analysis coincides with this by showing how diabetes is one of the main influential factors in the development of CRC. Although not modifiable, age is certainly the most significant factor influencing the risk of developing CRC as about 90% of the new cases occur in individuals over 50 years old [4]. Moreover, a larger BMI seems to affect also the risk of developing CRC.

The influence of smoking in our model is interesting as it would seem that it is better to be a smoker than to quit tobacco and become an ex-smoker. This appears to be related to the fact that the effects of smoking on CRC are mainly observed in the long run. People tend to be smokers when they are young and quit tobacco when they become older or are diagnosed with some condition for which tobacco is known to be a risk factor. Furthermore, as we are in the context of an observational study, we cannot discard the possibility that heavy smokers may have died earlier due to other conditions not recorded in the study. Thus, it is being an ex-smoker that would determine the risk of smoking in this case. However, further analysis would have to be done to reach a definitive conclusion.

Similar studies could also be performed using just certain segments of the CRC-positive population, which could target more precisely the influence of relevant factors in a specific group.

4. Discussion

The proposed BN associates relevant medical conditions and CR-CRFs in relation to CRC. We used expert opinion to get its initial

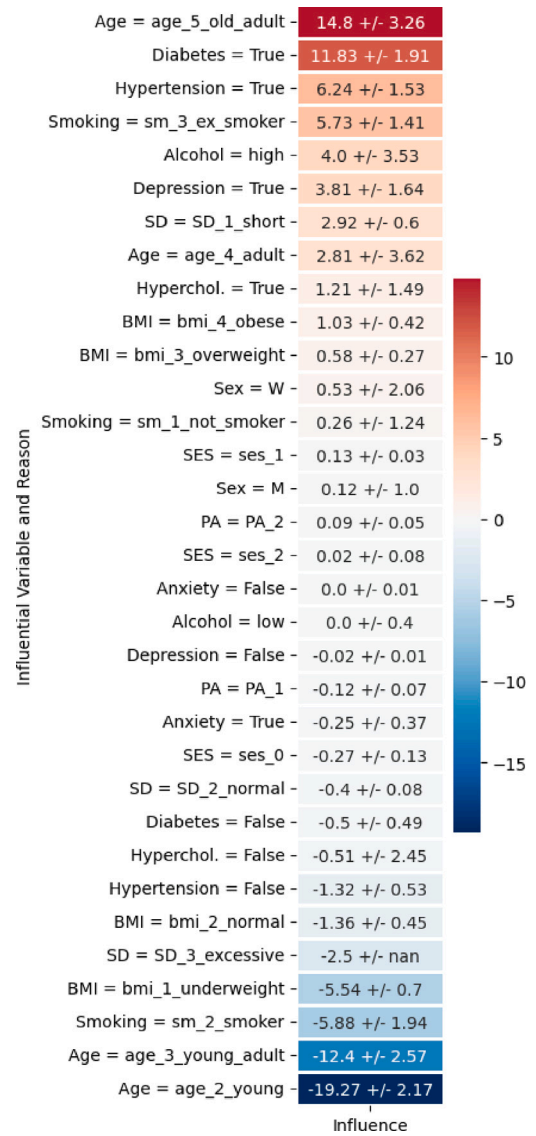


Fig. 7. Ranking of influential variables.

structure and an extensive database to update and complement it, from which we also built its conditional probability tables, with uncertainty in the beliefs acknowledged through posterior distributions.

We illustrated its use to provide risk maps and uncover CRC influential variables. But there are other relevant medical use cases which we briefly sketch:

- As mentioned, we had access to individuals' postcodes. This enables displaying geographical risk maps similar to those of Section 3.1 with the whole country as baseline and cells representing, say, provinces and their population size.
- Another important use is the classification of individuals, which we sketched in Section 2.2.3 for validation purposes, facilitating classifying an individual as more likely than not to have CRC. Should a different utility function be available, we would assign individuals to the class with maximum expected utility.
- In turn, and similarly, we could use the BN to segment a population based on posterior CRC probabilities or posterior expected utilities, given certain features, say for screening purposes, as we shall do in future work.
- A further important application of the network is for synthetic data generation purposes when available data are proprietary and

we need to share the data with a related organization [43]; this is easily achieved by sampling from the model defined in (1).

- A collateral use of our BN would be to generate interesting medical hypothesis. As an example, Tables 2 and 3 show how sleep duration is affected by age, as older people seem to sleep for shorter periods than younger people. There also seems to be a significant gap between men and women in terms of sleep duration being women the ones that sleep less, with this gap accentuated with age.

Our discussion in Section 2.2.2 about the prior chosen reflected the important dynamical aspect of updating the initial prior through the data over various years. This is of interest as the model can be easily updated to consider the most recent data acquired by the health insurance provider in order to be used again for risk assessment purposes with up-to-date information.

In future work, we shall incorporate this predictive model into the larger decision-support picture related to coherently advising screening methods. For this, we would need to consider the possible overall impact of the medical conditions using decision variables and utility functions. A decision-making problem will be defined for which the goal would be to find the portfolio of screening recommendations with maximum expected utility in line with precision vs current one-size-fits-all based on age approaches to screening [2]. Such model would facilitate the design of incentives to promote the adoption of CRC screening mechanisms and overcome current low adoption rates.

We conclude by pointing out several limitations of this study. First, the exploratory analysis described in Table 7 suggests a labor structure most probably different to that in other countries meaning that this model would either have to be adapted to the population structure in those countries or be used with some care taking into account this fact; yet the broad pipeline described would be reproducible. Second, some of the data were self-reported; however any possible fault was mitigated by several quality control strategies as described in [44]. Third, we had no data available concerning diet, genetics, and gut microbiome data; BMI, diabetes, and hypercholesterolemia might partly account for diet information, but this would be a confounding variable; concerning genetics, Marley and Nan [3] claim that about 35% of the CRC development risk is due to genes positively or negatively influencing patients. Very importantly, as mentioned above, the absence of the above three factors would prevent from causality claims in this study. Note though, again as discussed above, that we could anyway conclude predictive claims in the sense of Hernan and Robins [15], much as we did above in relation to sleeping duration. Finally, also hinted above, although we have updated the model over the years, it would also be of interest to consider the case of a dynamic BN framework to model disease evolution over time. This approach would aid also in extricating some cause–effect relationships between the variables.

Funding

This work was supported by the AXA-ICMAT Chair in Adversarial Risk Analysis; the Spanish Ministry of Science project PID2021-124662OB-I00; and the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement N. 101097036 (ONCO-SCREEN).

CRediT authorship contribution statement

D. Corrales: Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis, Conceptualization. **A. Santos-Lozano:** Writing – review & editing, Validation, Resources. **S. López-Ortiz:** Writing – review & editing, Validation, Data curation. **A. Lucia:** Writing – review & editing, Validation, Resources. **D. Ríos Insua:** Writing – review & editing, Writing – original draft, Supervision, Funding acquisition, Formal analysis, Conceptualization.

Table 6

Fourteen variables in the model.

| Variable | Definition | Levels |
|---------------|----------------------|--|
| v_{sex} | Sex | {female, male} |
| v_{age} | Age | (24,34], (34,44], (44,54], (54,64] |
| v_{SES} | Socioeconomic status | {1,2,3} |
| v_{BMI} | Body mass index | {underw., normal, overw., obese} |
| v_{PA} | Physical activity | {insufficiently active (1), sufficiently active (2)} |
| v_{SD} | Sleep duration | {short, normal, excessive} |
| v_{alc} | Alcohol consumption | {low, high} |
| v_{smok} | Smoker profile | {non-smoker, ex-smoker, smoker} |
| v_{anx} | Anxiety | {yes, no} |
| v_{dep} | Depression | {yes, no} |
| v_{hypert} | Hypertension | {yes, no} |
| $v_{hypchol}$ | Hypercholesterolemia | {yes, no} |
| v_{diab} | Diabetes | {yes, no} |
| v_{CRC} | Colorectal cancer | {yes, no} |

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.

Acknowledgments

We are grateful to Quirónprevención for the provision of data. Discussions with Victoria Ley, Víctor Lopez, and Isabela Rios were very useful.

Appendix. Data used

Table 6 provides the states of the fourteen variables used and how they are coded.

We briefly discuss how key variables were categorized. Age was divided into four groups ((24,34], (34,44], (44,54], and (54,64]), using the INE National Sport Habits survey coding, as in [44]. The socioeconomic status, originally a continuous variable, was discretized in three levels by binning its values using specified quantiles based on the variable's mean and standard deviation, with a larger index indicating a higher socioeconomic level.

Concerning BMI, we used the four WHO classes: *underweight* (< 18.5 kg/m²), *normal weight* ([18.5, 25) kg/m²), *overweight* ([25, 30) kg/m²), and *obese* (≥ 30 kg/m²). Participants' leisure-time PA levels were assessed as in [16], distinguishing between patients not meeting WHO minimum recommendations for aerobic PA in adults (*insufficiently active*) and meeting them (*regularly active*). SD was categorized as *short* (less than 6 h), *normal* (6–9 h), and *excessive* (> 9 h). The smoker profile reflected whether the patient was an active smoker, had never smoked, or was an ex-smoker. We also extracted whether the patient had anxiety or depression.

Concerning medical conditions, we used the following criteria: *diabetes*, medicated for it or glycemia ≥125 mg/dL; *hypercholesterolemia*, medicated for it or LDL ≥130 mg/dL, HDL ≤ 40 mg/dL, triglycerides ≥150 mg/dL or total cholesterol ≥200 mg/dL; *hypertension*, medicated for it or systolic/diastolic blood pressure ≥139/90 mm Hg.

Table 7 describes the full dataset distribution over all the years. With the exception of the lower presence of females, due to the labor sectors served by the incumbent health insurance provider, the structure and its health status seem by and large representative of the Spanish labor market. A *healthy worker effect* [45] might explain some of the somewhat lower estimates (anxiety, depression, diabetes).

Table 7
Percentage of observations at each class for variables in the model.

| Variable | States | Marginal | Variable | States | Marginal |
|----------------------|-------------|----------|---------------|-----------|----------|
| Sex | Female | 30.68% | Physical Act. | 1 | 47.21% |
| | Male | 69.32% | | 2 | 52.79% |
| Age(y) | (24, 34] | 21.21% | Anxiety | Yes | 2.70% |
| | (34, 44] | 38.02% | | No | 97.30% |
| | (44, 54] | 29.03% | Sleep Dur. | <6h | 10.88% |
| | (54, 64] | 11.73% | | (6 h–9 h) | 89.01% |
| Socioeconomic status | 1 | 23.93% | Depression | >9 h | 0.11% |
| | 2 | 61.97% | | Yes | 0.47% |
| | 3 | 14.10% | | No | 99.53% |
| BMI | Underweight | 1.10% | Diabetes | Yes | 3.63% |
| | Normal | 41.27% | | No | 96.37% |
| | Overweight | 40.67% | Hypertension | Yes | 15.05% |
| | Obese | 16.96% | | No | 84.95% |
| Smoker profile | Non-Smoker | 49.90% | Hypercholest. | Yes | 51.32% |
| | Ex-Smoker | 30.16% | | No | 48.68% |
| | Smoker | 19.94% | | CRC | Yes |
| Alcohol | Low | 95.05% | No | | 99.93% |
| | High | 4.95% | | | |

References

- [1] WHO, Colorectal cancer, 2023, <https://www.who.int/news-room/fact-sheets/detail/colorectal-cancer>.
- [2] F. Kastrinos, S.S. Kupfer, S. Gupta, Colorectal cancer risk assessment and precision approaches to screening: Brave new world or worlds apart? *Gastroenterology* 164 (5) (2023) 812–827, <http://dx.doi.org/10.1053/j.gastro.2023.02.021>.
- [3] A.R. Marley, H. Nan, Epidemiology of colorectal cancer, *Int. J. Mol. Epidemiol. Genet.* 7 (3) (2016) 105–114, URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5069274/>.
- [4] T. Sawicki, M. Ruskowska, A. Danielewicz, E. Niedźwiedzka, T. Arłukowicz, K.E. Przybyłowicz, A review of colorectal cancer in terms of epidemiology, risk factors, development, symptoms and diagnosis, *Cancers* 13 (9) (2021) 2025, <http://dx.doi.org/10.3390/cancers13092025>.
- [5] F. Jensen, T. Nielsen, *Bayesian Networks and Decision Graphs*, second ed., Springer, New York, 2007, <http://dx.doi.org/10.1007/978-0-387-68282-2>.
- [6] M. Scutari, J.-B. Denis, *Bayesian Networks: with Examples in R*, Chapman and Hall/CRC, 2021, <http://dx.doi.org/10.1201/9780429347436>.
- [7] S. McLachlan, K. Dube, G.A. Hitman, N.E. Fenton, E. Kyrimi, Bayesian networks in healthcare: Distribution by medical condition, *Artif. Intell. Med.* 107 (2020) 101912, <http://dx.doi.org/10.1016/j.artmed.2020.101912>.
- [8] K.-M. Wang, K.-J. Wang, B. Makond, Survivability modelling using Bayesian network for patients with first and secondary primary cancers, *Comput. Methods Programs Biomed.* 196 (2020) 105686, <http://dx.doi.org/10.1016/j.cmpb.2020.105686>.
- [9] B.-S. Jang, S.-J. Chun, H.S. Choi, J.H. Chang, K.H. Shin, et al., Estimating the risk and benefit of radiation therapy in (y) pN1 stage breast cancer patients: A Bayesian network model incorporating expert knowledge (KROG 22–13), *Comput. Methods Programs Biomed.* 245 (2024) 108049, <http://dx.doi.org/10.1016/j.cmpb.2024.108049>.
- [10] S. Liu, J. Zeng, H. Gong, H. Yang, J. Zhai, Y. Cao, J. Liu, Y. Luo, Y. Li, L. Maguire, et al., Quantitative analysis of breast cancer diagnosis using a probabilistic modelling approach, *Comput. Biol. Med.* 92 (2018) 168–175, <http://dx.doi.org/10.1016/j.combiomed.2017.11.014>.
- [11] R. Myte, B. Gylling, J. Häggström, J. Schneede, P. Magne Ueland, G. Hallmans, I. Johansson, R. Palmqvist, B. Van Guelpen, Untangling the role of one-carbon metabolism in colorectal cancer risk: a comprehensive Bayesian network analysis, *Sci. Rep.* 7 (1) (2017) 43434, <http://dx.doi.org/10.1038/srep43434>.
- [12] M. Sieswerda, R. van Rossum, I. Bermejo, G. Geleijnse, K. Aben, F. van Erning, I. de Hingh, V. Lemmens, A. Dekker, X. Verbeek, Estimating treatment effect of adjuvant chemotherapy in elderly patients with stage III colon cancer using Bayesian networks, *JCO Clin. Cancer Inform.* 7 (2023) e2300080, <http://dx.doi.org/10.1200/CCI.23.00080>.
- [13] B. Osong, C. Masciocchi, A. Damiani, I. Bermejo, E. Meldolesi, G. Chiloiro, M. Berbee, S.H. Lee, A. Dekker, V. Valentini, 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>.
- [14] E. Ferlizza, R. Solmi, M. Sgarzi, L. Ricciardiello, M. Lauriola, The roadmap of colorectal cancer screening, *Cancers* 13 (5) (2021) 1101, <http://dx.doi.org/10.3390/cancers13051101>.
- [15] M. Hernan, J. Robins, *Causal Inference: What If*, in: Chapman & Hall/CRC Monographs on Statistics & Applied Probab, CRC Press, 2023, URL <https://books.google.es/books?id=KnHIAAACAAJ>.
- [16] J. Ordovas, D. Rios-Insua, A. Santos-Lozano, A. Lucia, A. Torres, A. Kosgodagan, J. Camacho, A Bayesian network model for predicting cardiovascular risk, *Comput. Methods Programs Biomed.* 231 (2023) 107405, <http://dx.doi.org/10.1016/j.cmpb.2023.107405>.
- [17] K. Wada, Outliers in official statistics, *Jpn. J. Stat. Data Sci.* 3 (2) (2020) 669–691, <http://dx.doi.org/10.1007/s42081-020-00091-y>.
- [18] M. Scutari, C.E. Graafland, J.M. Gutiérrez, Who learns better Bayesian network structures: Accuracy and speed of structure learning algorithms, *Internat. J. Approx. Reason.* 115 (2019) 235–253, <http://dx.doi.org/10.1016/j.ijar.2019.10.003>.
- [19] M. Scanagatta, A. Salmerón, F. Stella, A survey on Bayesian network structure learning from data, *Prog. Artif. Intell.* 8 (4) (2019) 425–439, <http://dx.doi.org/10.1007/s13748-019-00194-y>.
- [20] L. BayesFusion, Genie modeler: Complete modeling freedom, 2023, <https://support.bayesfusion.com/docs/GeNie.pdf>, (Accessed 1 March 2024).
- [21] G. Ducamp, C. Gonzales, P.-H. Wuillemin, aGRUM/pyAgrum: a toolbox to build models and algorithms for probabilistic graphical models in python, in: 10th International Conference on Probabilistic Graphical Models, in: Proceedings of Machine Learning Research, vol. 138, Skørping, Denmark, 2020, pp. 609–612, URL <https://hal.archives-ouvertes.fr/hal-03135721>.
- [22] A. Ankan, A. Panda, pgmpy: Probabilistic graphical models using python, in: Proceedings of the 14th Python in Science Conference, SCIPY 2015, Citeseer, 2015, URL <https://pgmpy.org/>.
- [23] D. Koller, N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*, in: Adaptive Computation and Machine Learning, MIT Press, 2009, URL <https://books.google.co.in/books?id=7dzpHCHzNQ4C>.
- [24] M. Scutari, An empirical-Bayes score for discrete Bayesian networks, in: Conference on Probabilistic Graphical Models, PMLR, 2016, pp. 438–448, URL <https://proceedings.mlr.press/v52/scutari16.html>.
- [25] M. Scutari, Dirichlet Bayesian network scores and the maximum relative entropy principle, *Behaviormetrika* 45 (2018) 337–362, <http://dx.doi.org/10.1007/s41237-018-0048-x>.
- [26] S. French, D. Rios Insua, *Statistical Decision Theory*, first ed., Wiley, New York, 2000.
- [27] W. Buntine, Theory refinement on Bayesian networks, in: Uncertainty Proceedings 1991, Elsevier, 1991, pp. 52–60, <http://dx.doi.org/10.1016/B978-1-55860-203-8.50010-3>.
- [28] R. Castelo, A. Siebes, Priors on network structures. Biasing the search for Bayesian networks, *Internat. J. Approx. Reason.* 24 (1) (2000) 39–57, [http://dx.doi.org/10.1016/S0888-613X\(99\)00041-9](http://dx.doi.org/10.1016/S0888-613X(99)00041-9).
- [29] M. Ueno, Learning networks determined by the ratio of prior and data, 2012, <http://dx.doi.org/10.5555/3023549.3023620>, arXiv preprint [arXiv:1203.3521](https://arxiv.org/abs/1203.3521).
- [30] T.V. Allen, R. Greiner, Model selection criteria for learning belief nets: An empirical comparison, in: Proceedings of the Seventeenth International Conference on Machine Learning, 2000, pp. 1047–1054, <http://dx.doi.org/10.5555/645529.657974>.
- [31] M. Scutari, C. Vitolo, A. Tucker, Learning Bayesian networks from big data with greedy search: computational complexity and efficient implementation, *Stat. Comput.* 29 (2019) 1095–1108, <http://dx.doi.org/10.1007/s11222-019-09857-1>.
- [32] M. Ueno, Robust learning Bayesian networks for prior belief, in: Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, 2011, pp. 698–707, <http://dx.doi.org/10.5555/3020548.3020629>.
- [33] T. Silander, P. Kontkanen, P. Myllymäki, On sensitivity of the MAP Bayesian network structure to the equivalent sample size parameter, in: Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence, 2007, pp. 360–367, <http://dx.doi.org/10.5555/3020488.3020532>.
- [34] C. Bielza, P. Larranaga, Discrete Bayesian network classifiers: A survey, *ACM Comput. Surv.* 47 (1) (2014) 1–43, <http://dx.doi.org/10.1145/2576868>.
- [35] S. González, S. García, M. Lázaro, A.R. Figueiras-Vidal, F. Herrera, Class switching according to nearest enemy distance for learning from highly imbalanced data-sets, *Pattern Recognit.* 70 (2017) 12–24, <http://dx.doi.org/10.1016/j.patcog.2017.04.028>.
- [36] J. Ri, H. Kim, G-mean based extreme learning machine for imbalance learning, *Digit. Signal Process.* 98 (2020) 102637, <http://dx.doi.org/10.1016/j.dsp.2019.102637>.
- [37] T. Smith, M.J. Gunter, I. Tzoulaki, D.C. Muller, The added value of genetic information in colorectal cancer risk prediction models: development and evaluation in the UK biobank prospective cohort study, *Br. J. Cancer* 119 (8) (2018) 1036–1039, <http://dx.doi.org/10.1038/s41416-018-0282-8>.
- [38] B. Van Calster, D.J. McLernon, M. Van Smeden, L. Wynants, E.W. Steyerberg, Calibration: the achilles heel of predictive analytics, *BMC Med.* 17 (2019) 1–7, <http://dx.doi.org/10.1186/s12916-019-1466-7>.

- [39] M.P. Naeini, G.F. Cooper, Binary classifier calibration using an ensemble of near isotonic regression models, in: 2016 IEEE 16th International Conference on Data Mining, ICDM, IEEE, 2016, pp. 360–369, <http://dx.doi.org/10.1109/ICDM.2016.0047>.
- [40] L. Cox, What's wrong with risk matrices? *Risk Anal.: Int. J.* 28 (2) (2008) 497–512, <http://dx.doi.org/10.1111/j.1539-6924.2008.01030.x>.
- [41] S.L. Lauritzen, D.J. Spiegelhalter, Local computations with probabilities on graphical structures and their application to expert systems, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 50 (2) (1988) 157–194, <http://dx.doi.org/10.1111/j.2517-6161.1988.tb01721.x>, arXiv:<https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1988.tb01721.x>.
- [42] GBD 2019 Colorectal Cancer Collaborators, Global, regional, and national burden of colorectal cancer and its risk factors, 1990–2019: a systematic analysis for the global burden of disease study 2019, *Lancet Gastroenterol. Hepatol.* 7 (7) (2022) 627–647, [http://dx.doi.org/10.1016/S2468-1253\(22\)00044-9](http://dx.doi.org/10.1016/S2468-1253(22)00044-9).
- [43] D. Kaur, M. Sobieski, S. Patil, J. Liu, P. Bhagat, A. Gupta, N. Markuzon, Application of Bayesian networks to generate synthetic health data, *J. Am. Med. Inform. Assoc.* 28 (4) (2021) 801–811, <http://dx.doi.org/10.1093/jamia/ocaa303>.
- [44] P. Fernandez-Navarro, M. Aragoes, V. Ley, Leisure-time physical activity and prevalence of non-communicable pathologies and prescription medication in Spain, *PLoS One* 13 (2018) e0191542, <http://dx.doi.org/10.1371/journal.pone.0191542>.
- [45] D.M. Brown, S. Picciotto, S. Costello, A.M. Neophytou, M.A. Izano, J.M. Ferguson, E.A. Eisen, The healthy worker survivor effect: target parameters and target populations, *Curr. Environ. Health Rep.* 4 (2017) 364–372, <http://dx.doi.org/10.1007/s40572-017-0156-x>.