Bayesian Network Analysis for Modeling Cervical Cancer Risk Factors

Picascia Sergio
Data Science and Economics
University of Milan

Abstract

Cervical cancer is one of the most commonly occurring cancer among women, with 500'000 new yearly cases worldwide. Despite being one of the easiest to prevent, it kills thousands of women each year, especially in areas with high poverty levels and, therefore, low screening rates. It is commonly associated with human papilloma viruses: even though not every woman with HPV develops cervical cancer, its infection increases by a lot the odds of developing it. Other risk factors may include the age of the patient, her sexual activity, the usage of hormonal contraceptives, a weak immune system. The aim of this analysis is to build a Bayesian network, in order to quantify the impact of the already known risk factors and, possibly, to identify new sources of risk from the available features. The dataset used was collected at 'Hospital Universitario de Caracas', in Venezuela; it is made of variables regarding medical records, habits and demographical information of more than 800 hundred patients.

Introduction

Cervical cancer is a disease that affects the cervix, the lower part of the uterus in the human female reproductive system. In spite of the fact that it can be easily diagnosed, it is still threatening thousands of women each year. It is particularly common among less wealthy individuals, who do not have access to screening or health insurance; for this reason, it usually develops at further stages in african-american or hispanic women rather than in caucasian ones.

The data took into consideration was gathered in Caracas, Venezuela, at the 'Hospital Universitario de Caracas': it collects information about 858 female patients of the hospital, in particular their demographic data, habits and medical history. Apart from these, there are four variables indicating different tests run in order to diagnose cervical cancer: Hinselmann, Schiller, Cytology and Biopsy. For the analysis, a new feature has been created, Cervical.Cancer: it takes value 1 if any of the tests resulted positive and 0 otherwise. In this way, out of the 858 individuals, 102 of them tested positive for cervical cancer.

Unfortunately, there were some women that, for privacy reasons, decided to

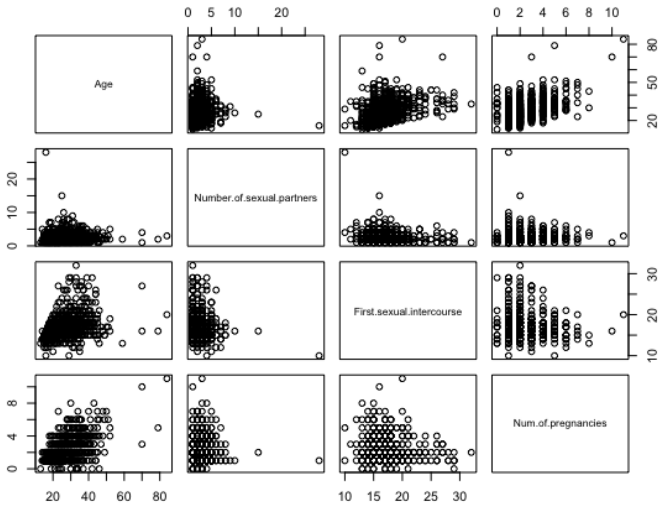

Figure 1: Scatter plots of the four continuous variables.

not answer to all the question asked; therefore more than one hundred observation were discarded due to missing values.

From the initial set of variables, some of them have been discarded since they did not bring useful information (columns of all zeros) or contained redundant data from other features. Among the final 19 attributes considered, only 4 of them were continuous (Figure 1) and, given the methods that have been implemented in the analysis, it was thought to be appropriate to transform them into categorical ones.

Analysis

The aim of the analysis is to observe how the already known risk factors for cervical cancer affect the probability of developing it and, possibly, discover new sources of exposure. The research focused on the development of a Bayesian network which allows to represent probabilistic dependencies among a set of random variables; in particular, causal relationships can be greatly defined thanks to the structure of the directed acyclic graph.

In order to build the Bayesian network, an hybrid method has been used, applying both an expert-driven and a data-driven approach: some relationships have been defined according to the domain knowledge, while others have been discovered using structure learning algorithms. From a prior research, it has been possible to discover which are the already known risk factors of cervical cancer; for example, we know that women are more likely to develop this

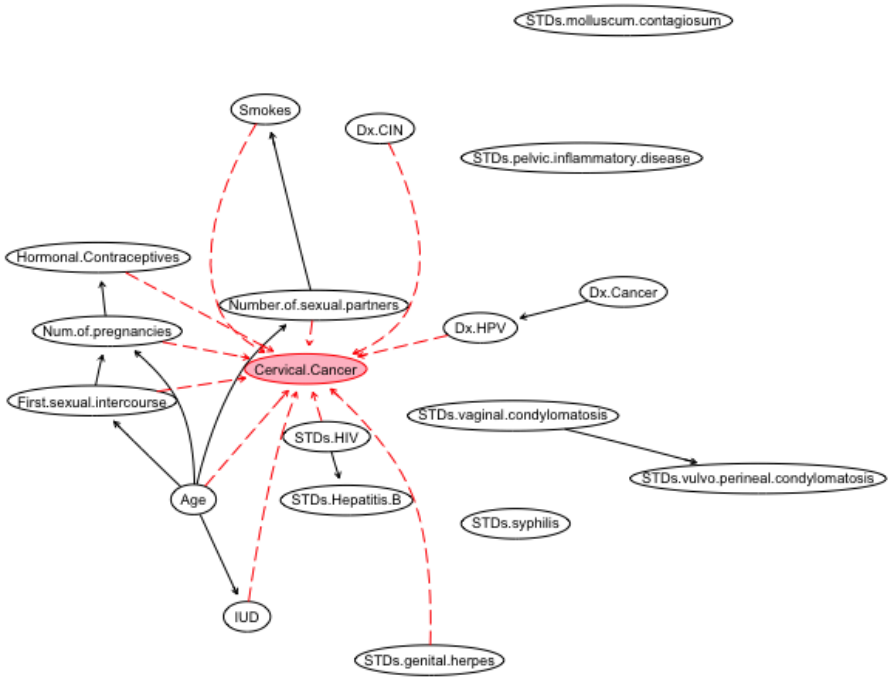

Figure 2: The final structure of the network; highlighted in red the incoming arcs for Cervical.Cancer.

disease in the age range from 20 to 30, or that having an intrauterine device (IUD) decreases this chance. Other known sources of risk are smoking, sexual activity, high number of pregnancies and other diseases, like HIV and HPV, in particular. Given this knowledge, it has been possible to define a whitelist, a list of relationships that were always taken into consideration while developing the models. On the other hand, also a blacklist was necessary, in order to avoid the algorithms to draw edges in directions that would not make sense; in particular, it was necessary to avoid any outgoing arcs from the variable of interest Cervical.Cancer.

Different structure learning algorithm were deployed in order to discover possible new relationships among the variables; despite using different class of algorithms, not only constraint-based (PC, Grow-Shrink and IAMB) but also score-based (Hill-Climbing, Tabu search) and hybrid ones (Max-Min Hill-Climbing, Max-Min Parents & Children), there wasn't any discovery about new risk factors for cervical cancer; on the other hand, interesting relationships among the other feature were discovered, for example how age is related to every attribute regarding sexual activity or other forms of cancer being a parent node for HPV.

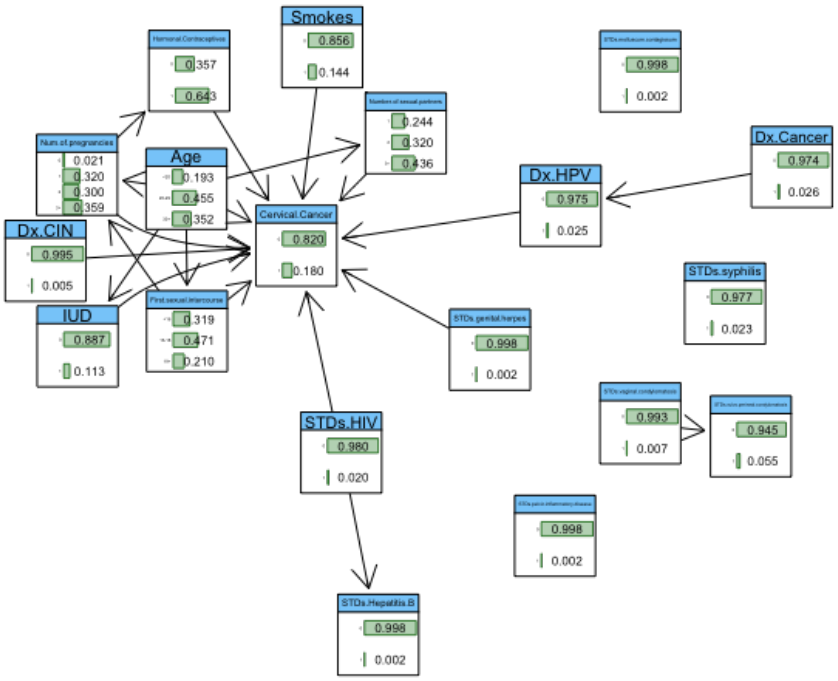

Figure 3: The network with the marginal probabilities for each node.

After defining the structure of the network, it was the turn of estimating the parameters: this task was accomplished applying the Bayesian approach. At this point, it was possible to plot the final graph of the network (Figure 2), where there have been highlighted in red the incoming arcs of the Cervical.Cancer node, and the plot of the marginal probability of each node (Figure 3).

Finally, inference on a Bayesian network can be mainly done performing two types of queries: conditional probability query or maximum a posteriori query. Fixing equal to '1' the state of the node about cervical cancer, thus considering the case of presence of the disease, there have been computed conditional probabilities distributions for the first type of queries (CPQ) and join posterior for the second one (MAP).

Conclusion

In spite of the fact that there wasn't any new factor discovered to increase the risk of developing cervical cancer, it is still possible to make inference and gain knowledge from the performed queries. Thanks to the conditional probability distributions (Figure 4), it can be seen which categories are more likely to de-

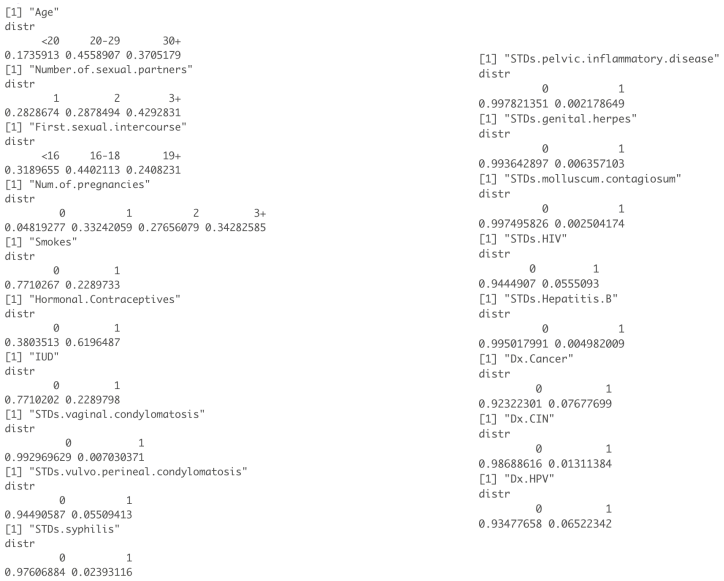

Figure 4: Conditional probability distributions for Cervical.Cancer = '1'.

velop the disease: if, on one hand, the results confirm what is already known in the literature, like an higher probability for women in the age range 20-29 or the ones with more pregnancies or sexual partners, on the other hand, the values for smokers and individuals with HPV contradicts the evidence.

The maximum a posteriori query (Figure 5), which represents the scenario of an individual with the highest chance of getting cervical cancer, suggests similar results to the conditional probability distributions observed before. These conflicting results are probably due to the dataset taken into consideration and, of course, need a deeper research to be confirmed.

Moreover, the same analysis could be conducted on the whole dataset, including the observations with missing values, maybe applying a method for filling those gaps, or also building a Bayesian network for mixed data in such a way to include also other variables that were left out.

```
$state
                            Age          Number.of.sexual.partners           First.sexual.intercourse
                          "<20"                                 "1"                             "<16"
             Num.of.pregnancies                             Smokes            Hormonal.Contraceptives
                            "1"                                 "0"                              "1"
                            IUD               STDs.genital.herpes                           STDs.HIV
                            "0"                                 "0"                              "0"
                         Dx.CIN                            Dx.HPV                  STDs.Hepatitis.B
                            "0"                                 "0"                              "0"
                     Dx.Cancer        STDs.vaginal.condylomatosis STDs.vulvo.perineal.condylomatosis
                            "0"                                 "0"                              "0"
                  STDs.syphilis   STDs.pelvic.inflammatory.disease        STDs.molluscum.contagiosum
                            "0"                                 "0"                              "0"

$prob
[1] 0.03438117
```

Figure 5: Maximum a posteriori for Cervical.Cancer = '1'.

R code

```
library(dplyr)
library(bnlearn)
library(Rgraphviz)
library(gRain)
library(gRbase)

### DATA MANIPULATION ###

# Import data
path <- '/Users/sergiopicascia/Desktop/risk_factors_
    cervical_cancer.csv'
data <- read.csv(path, na='?')
summary(data)

# Remove redundant columns and ones with all zeros
data <- subset(data, select = -c(STDs.cervical.
    condylomatosis, STDs.AIDS, STDs..Number.of.diagnosis,
                                     STDs..Time.since.last.
                                         diagnosis, Smokes..
                                         years., Smokes..packs
                                         .year.,
                                     Hormonal.Contraceptives
                                         ..years., IUD..years
                                         ., STDs..Time.since.
                                         first.diagnosis,
                                     STDs..number., STDs,
                                         STDs.condylomatosis,
                                         Dx, STDs.HPV))
```

```
# New feature 'Cervical.Cancer': 1 if at least one test
    is positive, 0 otherwise
data$Cervical.Cancer <- with(data, ifelse((Hinselmann+
    Schiller+Citology+Biopsy) >= 1, 1, 0))

# Remove rows with NAs
df <- na.omit(data)

# Plotting continuous vars
hist(data$Age)
hist(data$Number.of.sexual.partners)
hist(data$First.sexual.intercourse)
hist(data$Num.of.pregnancies)
pairs(data[, 0:4])

# Discretize continuous vars
df <- df %>% mutate(Age = case_when(Age < 20 ~ '<20',
                                    Age >= 20 & Age < 30
                                        ~ '20-29',
                                    Age >= 30 ~ '30+'),
                    Number.of.sexual.partners = case_when
                        (Number.of.sexual.partners == 1 ~
                        '1',
                                                    Number
                                                        .
                                                    of
                                                        .
                                                    sexual
                                                        .
                                                    partners
                                                        ==
                                                        2
                                                        ~
                                                        ,
                                                        2
                                                        '
                                                        ,
                                                    Number
                                                        .
                                                    of
                                                        .
```

```
                                                    sexual
                                                    .
                                                    partners
                                                    >=
                                                    3
                                                    ~
                                                    ,
                                                    3+
                                                    '
                                                    )
                                                    ,
First.sexual.intercourse = case_when(
    First.sexual.intercourse < 16 ~ '
    <16',
                                                First
                                                    .
                                                    sexual
                                                    .
                                                    intercourse
                                                    >=
                                                    16
                                                    &
                                                 First
                                                    .
                                                    sexual
                                                    .
                                                    intercourse
                                                    <
                                                    19
                                                    ~
                                                    ,
                                                    16-18
                                                    '
```

```
                                                    ,
                                                First
                                                    .
                                                    sexual
                                                    .
                                                    intercourse
                                                   >=
                                                    19
                                                    ~
                                                    ,
                                                    19+
                                                    '
                                                    )
                                                    ,
                    Num. of . pregnancies = case_when(Num. of
                        . pregnancies == 0 ~ '0',
                                                Num. of
                                                    .
                                                    pregnancies
                                                     ==
                                                     1
                                                    ~ ,
                                                    1',
                                                Num. of
                                                    .
                                                    pregnancies
                                                     ==
                                                     2
                                                    ~ ,
                                                    2',
                                                Num. of
                                                    .
                                                    pregnancies
                                                     >=
                                                     3
                                                    ~ ,
                                                    3+'
                                                    ))

# Convert variables to factor
```

```
df[colnames(df)] <- lapply(df[colnames(df)], factor)

### BAYESIAN NETWORK ###
df <- subset(df, select = -c(Hinselmann, Schiller,
    Citology, Biopsy)) # Considering only Cervical.Cancer

# Blacklist and whitelist
cols <- colnames(df1)
bl1 <- data.frame(from = cols[-grep('Age', cols)], #
    Prevent parents of Age
                  to   = c('Age'))
bl2 <- data.frame(from = c('Cervical.Cancer'), # Prevent
    children of Cervical.Cancer
                  to   = cols[-grep('Cervical.Cancer',
                      cols)])

bl <- rbind(bl1, bl2)

wl <- data.frame(from = c('Dx.HPV', 'STDs.HIV', 'Smokes',
     'Hormonal.Contraceptives', 'Num.of.pregnancies',
                           'Number.of.sexual.partners', '
                              STDs.genital.herpes', 'Age',
                               'IUD', 'First.sexual.
                              intercourse',
                           'Dx.CIN'),
                 to   = c('Cervical.Cancer'))

# Run structure learning algorithms and sum the adjacency
     matrices
sl_algos <- c(pc.stable, gs, iamb, hc, tabu, rsmax2, mmhc
    )
models <- list()
adj_mat <- matrix(0L, nrow = 19, ncol = 19)

for (a in sl_algos) {
 model = a(df1, blacklist = bl, whitelist = wl)
 models <- append(models, list(model))
 adj_mat <- adj_mat + amat(model)
 graphviz.plot(model, shape = 'ellipse', layout = 'fdp')
}

# Retrieve the most frequent edges
adj_mat[adj_mat < 2] <- 0L
adj_mat[adj_mat >= 2] <- 1L

# Build the BN
```

```
model <- empty.graph(cols)
amat(model) <- adj_mat
model <- pdag2dag(model, ordering = cols)
model

# Parameter learning
fitted_model <- bn.fit(model, df1, method = 'bayes')

# Plot of the graph
graphviz.plot(model, shape = 'ellipse', layout = 'fdp',
              highlight = list(nodes = 'Cervical.Cancer',
                  arcs = incoming.arcs(model, 'Cervical.
                 Cancer'),
                               lty = 5, fill = 'pink',
                                   col = 'red'))

# Plotting marginal probabilities
graphviz.chart(fitted_model, layout = 'fdp', type = '
    barprob', scale = c(2, 2), bar.col = "darkgreen",
              strip.bg = "lightskyblue")

# Conditional probability distributions
for (col in cols) {
 distr <- cpdist(fitted_model, col, (Cervical.Cancer == '
     1'))
 n <- nrow(distr)
 print(col)
 print(table(distr)/n)
}

# Maximum a posteriori query
gr.fit <- as.grain(fitted_model)
cerv.canc <- setEvidence(object = gr.fit, nodes = '
    Cervical.Cancer', states = '1')
joint.post <- querygrain(cerv.canc, type = 'joint')

map <- function(joint) {
 ind_max <- which(sapply(joint, function(v) isTRUE(all.
     equal(max(joint), v))))
 ind <- arrayInd(ind_max, .dim = dim(joint))
 state <- mapply('[', dimnames(joint), ind)
 prob <- joint[ind_max]
 list(state=state, prob=prob)
}

map(joint.post)
```