# Application of Bayesian Networks to the survival analysis of breast cancer
# Probabilistic Modelling project - DSE

Nicole Maria Formenti

March 2021

# Application of Bayesian Networks to the survival analysis of breast cancer

## Introduction and aims

This project will focus on the survival analysis of patients affected by breast cancer, which is the prediction of their life span, after they have undergone breast surgery. The analysis will try to understand which are the specific factors affecting the survival expectation. This problem has been often tackled in literature by means of Bayesian network models, which are particularly suited for biomedical use. One of the strengths of Bayesian networks is that they can either be built only based on data or by incorporating experts' knowledge and evidence from previous studies or a mix of the two. They have been widely used to support medical decision concerning the treatment of patients affected by cancer and to provide better psychological support in case of bad prognosis. In this way doctors are able to estimate more accurately the probability of the outcome in different situations and classify patients' prognosis with a certain level of confidence.
Some reference papers to such studies are [**5**], which is an overview on the use of Bayesian network in medicine, [**1**], [**6**], [**7**], which performs survival analysis on patients having respectively gallbladder carcinoma and bone sarcomas. Several articles refer to the specific case of survival analysis in the context of breast cancer [**2**], [**3**], [**4**], [**8**], [**10**], [**9**], where the latter is a meta-analysis about the use of this type of models for this problem. Some of the studies compare the Bayesian models with other models, such as the support vector machines [**2**] and [**8**], the Cox regression [**3**] and other several algorithms [**10**]. Their findings is that the performance of the Bayesian networks is in line with that of others state-of-the-art algorithms, in some cases they work even better.
An interesting point highlighted in some of the papers, namely [**8**] and [**10**], is that these probabilistic models are able to deal in a robust way with missing data, meaning that they can infer the correct probabilities when such data are present. This is a very interesting finding, in particular in biomedical application, since medical data are often incomplete, due to lack of analysis carried out on the patient, lack of information, mistakes in recording or difficult integration of different data sources. Moreover, medical data are often scarce, so having a method able to deal with missing variables means being able to exploit all the available data.

The aims of this analysis are multiple. The first goal is to assess whether the above findings applies in reality, that is whether Bayesian networks are robust when missing data are present in the dataset. Two different Bayesian networks will be fitted, one using a version of the dataset containing only complete observations (420 observations) and another one using the full dataset including the missing variables (1483 observations).
A further aim is to quantify the ability of the network in classifying the different patients according to their expected survival. Different models will be built and validated for each type of dataset, containing and not containing the missing data, and a final model will be chosen according to the Accuracy metrics, that indicates the percentage of correctly classified instances, and the F1 score, which balances recall and precision. Two models will be obtained and their performance will be assessed by using the ROC (Receiver Operating Characteristic) curve, where the true positive rate is plotted against the false positive rate at different thresholds. The AUC (Area Under the Curve) is the metric used, which indicates the

probability with which the model is able to distinguish a negative from a positive class. Usually, values greater than 70 are considered sufficiently good. In this way, it can be evaluated whether the model built on top of missing data has a better, worse or equal performance.
The final goal is to assess how the change in some variables affects the probability of the values assumed by the target variable, in particular the focus will be on the variables directly affecting the target. In order to do it, some evidence is set and the new posterior probabilities are estimated according to it.

# Data-set description

The dataset chosen for this analysis is the *Breast Cancer (METABRIC)*, which can be found on the *Kaggle* dataset repository [**11**]. The dataset contains data about 2.509 patients having breast cancer which have undergone breast surgery and it includes 35 attributes, namely clinical and gene expression attributes. All the variables have been discretised in order to use a single data type. They are as follows:

1. *Patient ID*: unique identifier of the patient. It is removed.
2. *Age at diagnosis*: float number indicating the age of the patient. It is discretised, based on the values of the quartiles, in the following classes:
   - $\leq 50$ years
   - $> 50$ and $\leq 60$ years
   - $> 60$ and $\leq 70$ years
   - $\geq 50$ years
3. *Type of breast surgery*: it contains two levels:
   - Mastectomy: all the breast tissue is removed
   - Breast conserving: only the part of the breast affected by the cancer is removed
4. *Cancer type*: it is removed since the variable Cancer type detailed contains the same information but more in detail.
5. *Cancer type detailed*: it refers to the type of breast cancer. It contains five levels: *Breast invasive ductal carcinoma, Breast mixed ductal and lobular carcinoma, Breast invasive lobular carcinoma, Breast invasive mixed mucinous carcinoma, Metaplastic breast cancer.*
6. *Cellularity*: it is the cellularity post chemotherapy, which is the amount of tumor cells and their arrangement. It can take three values: *Low, Moderate, High.*
7. *Chemotherapy*: it indicates whether the patient has been treated with chemotherapy. It is a binary variable
8. *Pam 50 and claudin-low subtype*: it is a test used to identify the profile of a tumour in order to understand whether some cancers (ER positive and HER-2 negative) are likely or not to metastasize. It is defined based on some gene expression characteristics. It takes values:
   - Basal: basal-like
   - Claudin-low Her2: Her2 enriched
   - LumA: luminal A
   - LumB: luminal B
   - NC
   - Normal: normal-like
9. *Cohort*: it is a group of subjects sharing similar characteristics. Since it is not clearly explained which are the characteristics the classification refers to, the variable is removed.

10. *ER status measured by IHC*: it indicates if the cancer cells are positive or negative for estrogen receptors. If they are positive, the cancer cells use the hormone estrogen to grow faster. It takes binary values. It is based on a methodology called immune-histochemistry. Since it contains information similar to the variable ER status and it contains more missing values, it is eliminated. However, the results of the two seem consistent.

11. *ER status*: it indicates if the cancer cells are positive or negative for estrogen receptors.

12. *Neoplasm histologic grade*: it is determined based on the aggressiveness of the cancer cells. It can take values 1,2 or 3.

13. *HER2 status measured by SNP6*: it indicates if the cancer cells are positive or negative for HER2, which is a growth promoting protein on the outside of the breast cancer cells. High levels of HER2 means that the cancer spread at a faster rate, however it also responds better to drugs targeting the HER2 protein.It is assessed through a type of next generation sequencing methodology.

    - Gain: the gene is over-expressed, the mutation can cause an abnormal expression
    - Loss: the gene is under-expressed, the mutation can suppress the normal expression of the gene
    - Neutral
    - Undefined

14. *HER2 status*: it indicates if the cancer cells are positive or negative for HER2. It is eliminated since it contains similar information to the variable *HER2 status measured by SNP6*, but the latter provides more detailed data.

15. *Tumour other histologic subtype*: it is the type of cancer based on microscopic examination of the tissues. It takes values: *Ductal/NST, Mixed, Lobular, Tubular/Cribriform, Mucinous, Medullary, Other, Metaplastic.*

16. *Hormone therapy*: it indicates whether the patient has been treated with hormonal therapy. It is used in the case of ER and PR (progesterone receptor) positive cancers to reduce the amount of these hormones. It takes binary values.

17. *Inferred menopausal state*: it indicates whether the patient is post or pre menopausal.

18. *Integrative cluster*: it is the molecular subtype of the cancer based on gene expression. It takes discrete values from 1 to 9, expect for:

    - 4ER+
    - 4ER-

19. *Tumour laterality*: it indicates whether it is involved the right or the left breast.

20. *Lymph nodes examined positive*: it is the number of lymph nodes found involved with cancer sampled during the breast surgery. It takes discrete values from 1 to 10 and then two other classes for higher numbers has been created to reduce the total number of variables:

    - $> 10$ and $\leq 20$
    - $> 20$

21. *Mutations count*: it is the number of genes with mutations relevant for the cancer. It takes discrete values from 1 to 15 and then a class $> 15$ has been created to include the rest of the observations.

22. *Nottingham prognostic index*: it determines the prognosis after breast cancer surgery. Its value is based on the size of the tumour, the number of positive lymph nodes and the grade of the tumour. It takes continuous values and it has been discretized by rounding the result. It takes discrete values from 1 to 6.

23. *Oncotree code*: it is an open-source ontology for cancer type diagnosis from a clinical perspective. It takes values: *BRCA, BREAST IDC, ILC, IMMC, MBC, MDLC, PBS*.

24. *Overall survival months*: it is the period between the breast surgery and the death of the patient expressed in months. It is the **target variable**. All the observations having missing values for this variable have been eliminated since they cannot be evaluated for classification. It is transformed into a binary variable taking values:

    - $<$ 60 months ($<$ 5 years): 60 months correspond to the first quartile, so the resulting 2 classes are not too unbalanced and they contain enough observations. This threshold has been chosen since it is useful from a clinical point of view, in order to understand the life expectancy of the patient and identify the best possible treatment accordingly.
    - $\geq$ 60 months ($\geq$ 5 years)

25. *Overall survival*: it indicates whether the patient is still alive or not. It is eliminated since it encodes the same information of the target variable.

26. *PR status*: it indicates whether the cancer cells are positive or negative to progesterone receptors. It takes binary values.

27. *Radio therapy*: it indicates whether or not the patient as been treated with radio therapy. It takes binary values.

28. *Relapse free status*: it indicates whether the patient has experienced a relapse of breast cancer or not. It takes binary values.

29. *Relapse free status in months*: it indicates whether the patient has experienced a relapse of breast cancer or not and it is expressed in months. The variable has been deleted and only the binary feature *Relapse free status* has been kept for simplicity.

30. *Sex*: it indicates the sex of the patient. Since all the observations refer to female patients, the variable is not informative so it has been deleted.

31. *Three gene classifier subtype*: it is a type of classification based on the gene expression. It can take values:

    - ER-/HER2- : basal-like tumour
    - ER+/HER2- High Prolif: combined luminal A and B with high proliferation
    - ER+/HER2- Low Prolif: combined luminal A and B with low proliferation
    - HER2+ : HER2 enriched

32. *Tumour size*: it is the size of the tumour measured by imaging techniques. The size has been discretised in the following bins: $0-10$, $10-20$, $20-30$, $30-40$, $40-50$, $50-60$, $60-70$, $70-80$, $80-90$, $90-100$, $>100$

33. *Tumour stage*: it is the classification of the stage of the tumour, which is based on the surrounding structure, the lymph nodes positive and the distance of the spread of the cancer. It is expressed as a discrete variable from 0 to 5

34. *Death from cancer*: whether the death was due to cancer or not. All the observations of patients dead for causes other than cancer have been eliminated, as well as the variable itself which became non-informative.

After the data cleansing process, the number of features is equal to 25 and the dataset is present in two version, one including missing data with 1483 observations and another one including only complete records with 420 observations.

# Methodology

The model applied to the dataset is the Bayesian network model, which is represented as a directed acyclic graph, so a graph that encodes cause-effect relationships by means of directed arcs. It is particularly suited for biomedical applications, where different factors interact with each other in specific ways. There are three families of structure learning algorithms, which are the constraint-based, score-based and hybrid. It is used an algorithm of each type and since the dataset contains many variables, the chosen algorithms are those best suited for high dimensional modelling. Moreover, they are specific for categorical variables. They are the *PC-algorithm*, the *hill-climbing* and the *max-min hill-climbing*. The packages used for the complete version of the dataset are the ***pcalg*** for the *PC-algorithm* and the ***bnlearn*** for the other two. As for the dataset containing missing values, the package used is ***bnstruct***, which allows the application of Bayesian models in the presence of missing data, differently from the previous packages. This package doesn't implement the *PC-algorithm*, but only the *hill-climbing* and *max-min hill-climbing*.
Based on each algorithm, two different structures have been created, one without any prior knowledge, so entirely learnt from data, and another one incorporating experts' knowledge and well-known findings from biomedical literature. In order to implement the second type of Bayesian network, some interactions between particular nodes have been forced and other have been forbidden.
The connections that have been fixed a priori are the following:

- **Relapse free → Survival**: this connection has been included since it were present in almost all the networks whose structure was entirely data-driven. Furthermore, it is a logical relation since a patient experiencing a recurrence in breast cancer has a higher probability of not surviving compared to a relapse-free patient, which is subject to cancer only once during his lifetime.

- **Stage → Survival**: according to the *cancer.net* website [**12**], which is aimed at informing cancer patients and their family and approved by the ASCO (American Society of Clinical Oncology), the level of spread of the breast cancer across the body greatly impacts the survival of the patient. Cancers which have metastasised lead more frequently to the death of the patient with respect to locally located cancer. More in details, the survival rate estimated for a cancer located only in the breast is 99%, if the cancer spreads to the lymph nodes, the rate becomes 86% and when the cancer metastasises in a far region, the percentage dramatically drops to 28%.

- **Neoplasm histological grade, Size, Lymph nodes examined positive → Stage**: according to *cancer.net* [**13**], the stage of the cancer is commonly defined through the TNM system, which stands for the size of the tumour and where it is located (Tumour), the number of lymph nodes which have been affected by the cancer (Node) and whether the cancer has spread in other parts of the body (Metastasis). Also the histological grade of the cancer cells, the biomarkers and the genetic characteristics of the tumour are considered.

- **PR status, ER status, HER2 status → Survival**: from WebMD website [**14**], which is an online publisher providing health information, the 4-years survival rate of breast cancer patients depends on their positivity to estrogen or progesterone and on HER2 status. The better survival chances are for patients which are hormone positives (HR) and HER2 negative (92.5%), then for patients which are HR positive and HER2 positive (90.3%). If patients are HR negative, their survival chances reduce depending whether they are HER2 positive (82.7%) or HER2 negative (77%). The latter case is called triple negative and it has the worst outcomes.

- **Pam 50 and claudin-low subtype → Survival**: it is a tumor profiling test [**15**] which helps to understand how likely a breast cancer is to metastasise, thus impacting on the survival of the patient.

- **Cancer type detailed → Oncotree code**: the Oncotree code [**16**] is an open-source ontology which has been developed to standardise cancer type diagnosis, so it depends on it.

- **Hormone therapy, Chemotherapy and Radiotherapy → Relapse free**: these types of therapies are also called adjuvant therapies and they are usually given in addition to the primary therapy in order to reduce the risk of recurrence of the cancer. According to *cancer.net* [**17**],

radiotherapy and chemotherapy tends to reduce the recurrence of breast cancer, especially when combined with breast surgery. The same applies for hormonal therapy, which can be given either before the surgery of afterwards, to both shrink the cancer and reduce the risk of recurrence.

- **PR and ER status → Hormone therapy**: from the *cancer.net* website [**17**], hormonal therapy is used only in the cases of tumour cells which are either PR-positive or ER-positive, meaning they respectively make use of progesterone and estrogen hormone to grow faster. Hormonal therapy suppresses or reduces the production of such hormones within the body, starving the cancer cells. As a consequence, this kind of therapy is useless for PR-negative and ER-negative tumours.

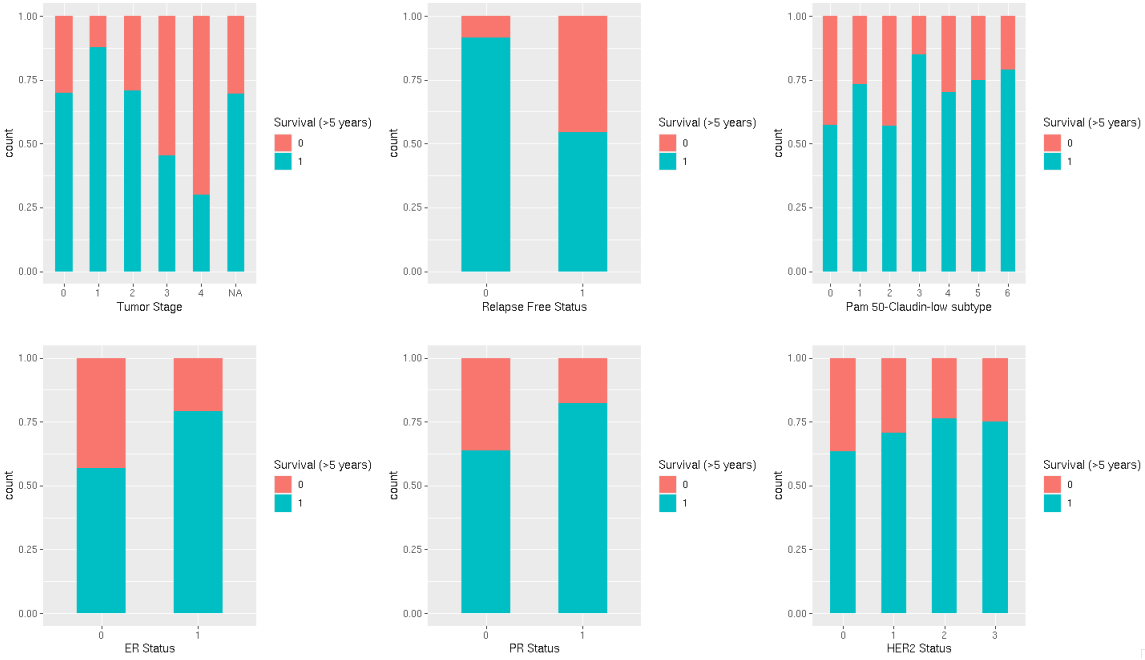


Figure 1: *Bar charts of the variables of the dataset influencing the survival according to literature. It can be seen that all of them have an impact on the survival rate. The impact of the HER2 status on survival is less marked, while the strongest differences in the survival rate are due to the relapse free status and the tumour stage.*

On the other hand, some relationships are not allowed to be inserted in the final network. The causal relationships starting from an encoding system based on physiological characteristics or a therapy and ending to a physiological characteristic or state of the patient have no logical sense. The same applies for arcs going from the age at diagnosis of the patient to other physiological characteristics.

After the structure has been learnt, the parameters of the model are fitted through the Bayesian parameters estimation, provided by the package ***bnlearn***. In order to select the best model for each version of the dataset, the models are tested with a 10-fold inner cross validation by means of a validation and training set, and then their performance is evaluated with a 10-fold outer cross validation, by retraining the model on both training and validation set and testing its performance on a held-out portion of the dataset. The metrics used to compare the models are the Accuracy and F1 score, in addition, the best model is evaluated through the AUC of the ROC curve, which allows to estimate how good is the final classification. The curve shows the proportion of misclassified positive or negaitve instances for different classification thresholds, so that one can choose the more appropriate threshold depending on the problem at hand. This metric has been used in almost all of the reference papers to validate the models and, in general, it is popular to assess biomedical classification problems.

# Results

As a result of the structure learning phase, the total number of models found are six for the dataset version containing only complete data and eight for the dataset in its original version.

Regarding the first group of models, the algorithms used are the *PC*, *hill-climbing* and *max-min hill-climbing* algorithms. For each of them, a version of the dataset entirely data-driven and another one based on prior knowledge have been fitted. The models built using PC have been excluded since the implementation in the package *pcalg* doesn't allow to set the direction of arcs, but only the connections, so some results don't make sense in terms of causal-effect relationship. As for the other two type of algorithms used, the version built with no prior knowledge has too few connections affecting the target variable and it is quite sparse. They have been both fitted by using a modified version of the BIC score. The parameters of the models are fitted by means of the Bayesian parameters' estimation. Then, the classification performance of the models is compared by first dividing the dataset in two parts and then performing an inner 10-folds cross-validation on the first part of the dataset *(Fig.2)*.

| | avg_balaccuracy | avg_precision | avg_recall | avg_F1 |
|---|---|---|---|---|
| pc.1 | 0.7105632 | *NA* | 0.7889036 | *NA* |
| pc.2 | 0.6011995 | 0.5161129 | 0.3996810 | 0.4259050 |
| hc.1 | 0.5846777 | 0.7007937 | 0.2577477 | 0.3512687 |
| hc.2 | 0.6621216 | 0.5439765 | 0.5566854 | 0.5290892 |
| mmhc.1 | 0.5000000 | *NA* | 0.0000000 | *NA* |
| mmhc.2 | 0.6736701 | 0.5697241 | 0.5613307 | 0.5485253 |

Figure 2: *10-folds inner cross validation results for the models built on the complete version of the dataset*

The models built using the *hill-climbing* and *max-min hill-climbing* plus the prior knowledge have a similar performance, so the chosen model is the one built with the *MMHC* algorithm. The network structure is refitted by eliminating the variables corresponding to the nodes not connected to the main bulk of the graph. In *(Fig.3)* the resulting final network is showed.

From the network, conditional independences can be derived. They indicate which variables provide a redundant information given other variables. The independences considered are only the main ones related to the target variable *Survival*.
First, it can be easily seen that

$$Survival \perp\!\!\!\perp (Positive\ lymph\ nodes, Tumour\ size, Neoplasm\ histological\ grade, Nottingham\ prognostic\ index) \mid Tumour\ stage$$

The remaining conditional independences are:

$$Survival \perp\!\!\!\perp Hormone\ therapy \mid (PR\ status, ER\ status, Relapse\ free\ status)$$

$$Survival \perp\!\!\!\perp (Chemotherapy, Age\ at\ diagnosis, Radiotherapy, Type\ of\ breast\ surgery, Inferred\ menopausal\ state) \mid Relapse\ free\ status$$

In addition, the target variable is marginally independent from the variables that have been eliminated since they weren't linked to the main body of the network, which are: *Cancer type detailed, Cellularity, Tumour other histologic subtype, Integrative cluster, Primary tumor laterality, Mutation count, Oncotree code, 3 gene classifier subtype.*
The variables on which the target variable directly depends on are the following: *HER2 status, PR status, ER status, PAM 50 and Claudin-low subtype, Relapse free status, Tumour stage.* Those are all the variables included in the model based on prior knowledge. No other variables having a direct effect on the target have been included by the algorithm.

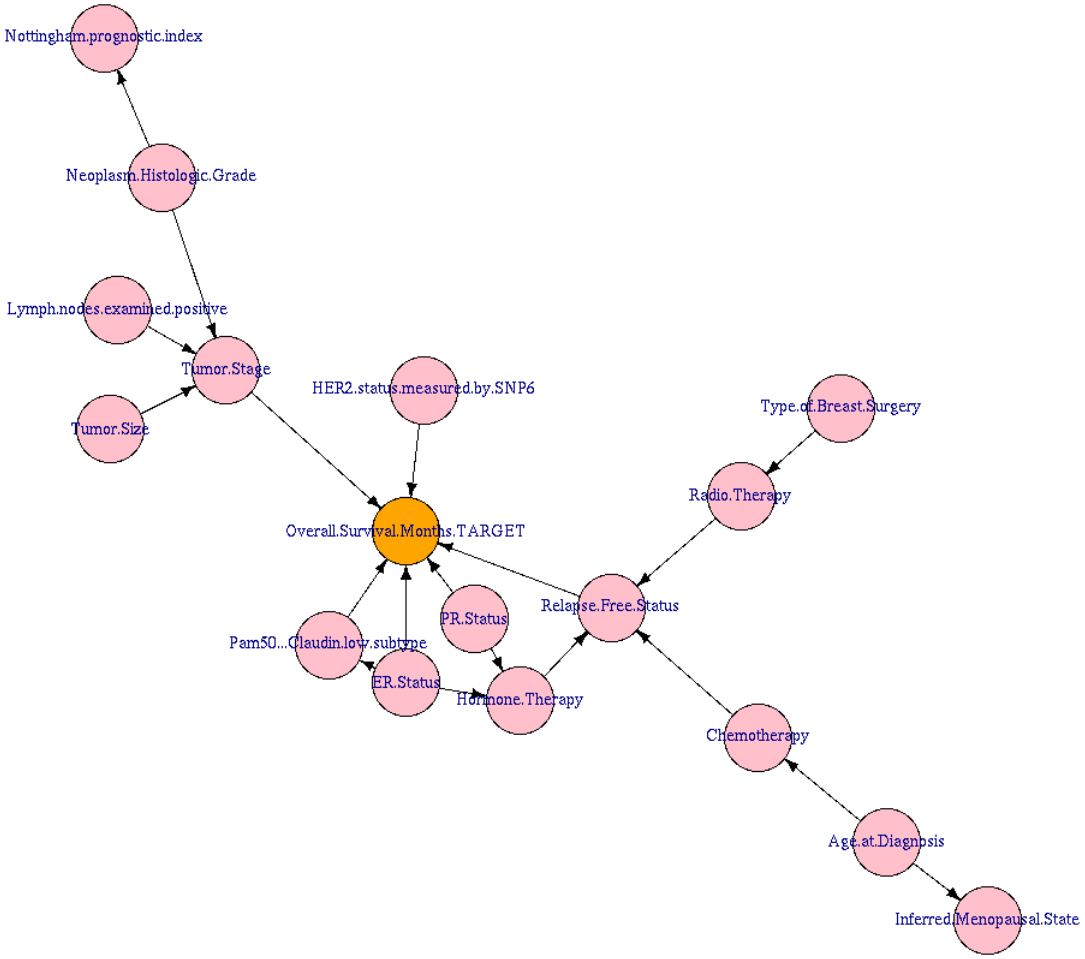


Figure 3: *Final structure of Bayesian network found by using the MMHC algorithm on the complete verison of the dataset*

Afterwards, an outer cross validation is performed by refitting the parameters of the network on the union of the training and validation set, and then tested on the held-out set. The final performance of the model is 70.5% Accuracy and 62.1% F1 score.
It is possible that by changing the probability threshold used to assign the classes the F1 score improves, or at least, a better proportion between the recall and the precision could be achieved. Setting a threshold means to classify an observation as $Survival > 5\, years$ if the corresponding probability of belonging to this class is greater than the thresholds. Thus, a further 10-folds cross-validation is applied for different thresholds levels, which are: 0.35, 0.40 , 0.45, 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.90. The best threshold is around 0.70, with an Accuracy of 72.3% and a F1 score of 66%, and with a fairly good balance between precision and recall.
In order to assess the goodness of the classifier, the ROC curve is plotted and the AUC calculated *(Fig.4)*. To have a more complete assessment, the ROC curve is built both by using the version of the dataset with no missing variable and the original dataset, whose missing values have been imputed by using the package *bnlearn*.

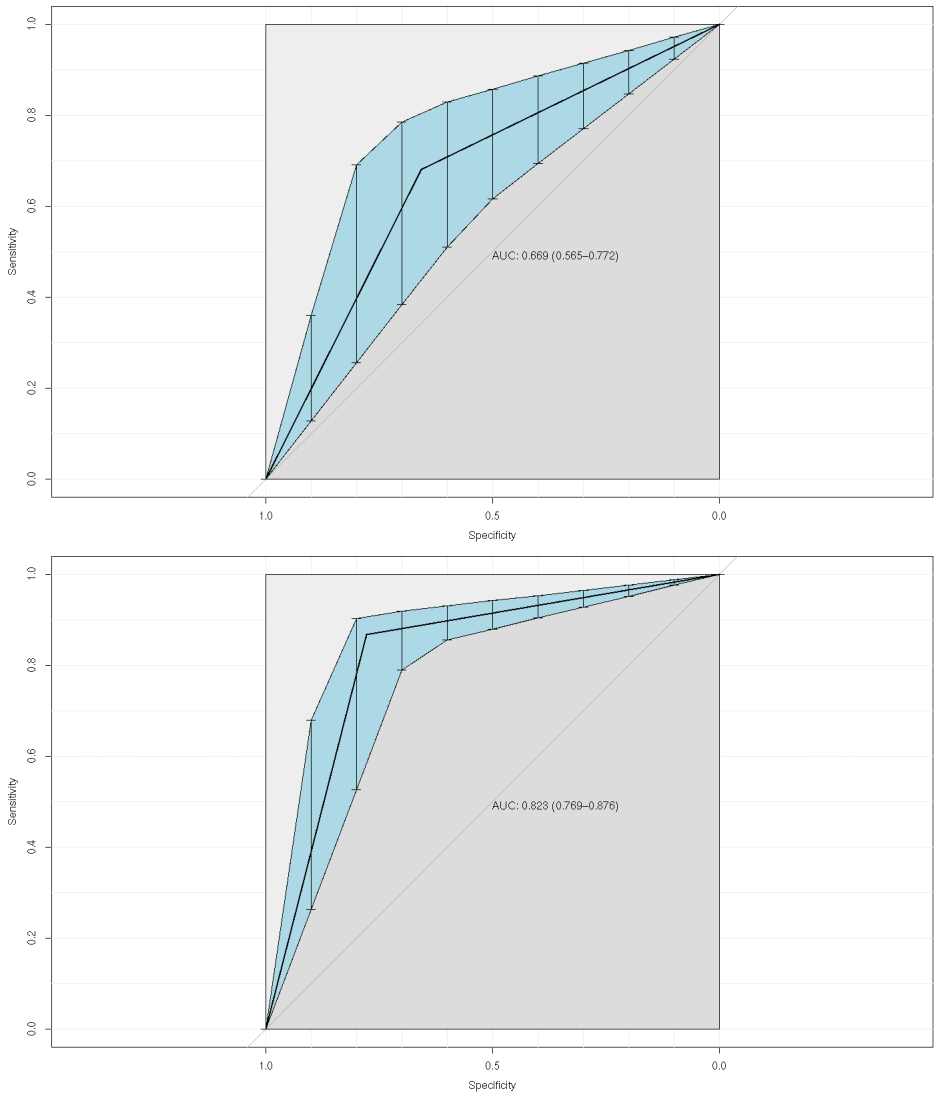


Figure 4: *At the top, the ROC curve for the complete version of the dataset, at the bottom, the ROC curve for the original imputed dataset*

The ROC resulting from the complete version of the dataset has an AUC of 66.9%, which is discrete, however it has a high variance, so it is quite unstable. The second ROC curve, built by using the imputed dataset, has a quite good AUC of 82.3%, which is also more stable. This AUC tells us that the model is, on average, able to correctly distinguish between positive and negative classes 82.3% of the time.

As for the second group of models, namely those built by using the dataset containing missing values, a total of eight possible network structures is fitted. Only the *hill-climbing* and *max-min hill-climbing* are available for this type of data. The procedure is the same as before but, in addition, for each type of algorithm both AIC and BIC scores are used. The results from the inner 10-folds cross-validation *(Fig.5)*, used to compare all the models, show that only the models incorporating prior knowledge have a good enough performance. Moreover, there isn't much difference between the models using AIC and BIC, neither between the structures built with the two different algorithms.

| | avg_balaccuracy | avg_precision | avg_recall | avg_F1 |
|---|---|---|---|---|
| hc_AIC1 | 0.5000000 | NA | 0.0000000 | NA |
| hc_BIC1 | 0.5000000 | NA | 0.0000000 | NA |
| hc_AIC2 | 0.6821478 | 0.5994182 | 0.4807653 | 0.5309173 |
| hc_BIC2 | 0.6748501 | 0.5725197 | 0.4810311 | 0.5193380 |
| mmhc_AIC1 | 0.5000000 | NA | 0.0000000 | NA |
| mmhc_BIC1 | 0.5000000 | NA | 0.0000000 | NA |
| mmhc_AIC2 | 0.6709087 | 0.5849270 | 0.4646297 | 0.5136004 |
| mmhc_BIC2 | 0.6718545 | 0.5913082 | 0.4605549 | 0.5138809 |

Figure 5: *10-folds inner cross validation results for the models built on the original dataset containing missing values*

The model chosen is the one using the *HC* and fitted with the AIC score. The structure of the model is refitted again by removing the unconnected nodes. It is worth noting that the variables removed from this model, so the ones that has no effect on the main body of the network, have also been removed from the previous model. They are: *Cellularity, Inferred menopausal state, Tumour other histologic subtype, Type of breast surgery, Primary tumour laterality.* The final structure is represented in *(Fig.6)*.
As before, the conditional independences in terms of the target variable will be listed. Some conditional independences are the same with respect to the previous graph:

$$Survival \perp\!\!\!\perp Hormone\ therapy \mid (PR\ status, ER\ status, Relapse\ free\ status)$$

However, there is another conditional independences which is similar:

$$Survival \perp\!\!\!\perp (Chemotherapy, Radiotherapy) \mid Relapse\ free\ status$$

The remaining conditional independence is:

$$\begin{gathered} Survival \perp\!\!\!\perp (Positive\ lymph\ nodes, Tumour\ size, Neoplasm\ histological\ grade, Primary\ tumour\ laterality, \\ Mutation\ count, Integrative\ cluster, Cancer\ type\ detailed, Oncotree\ code, 3\ gene\ classifier \\ Nottingham\ prognostic\ index) \\ \mid (Pam50\ claudin-low\ subtype, Tumour\ stage, HER2\ status) \end{gathered}$$

The target variable is marginally independent from the features that have been eliminated since they weren't connected to the network. They are: *Age at diagnosis, Type of breast surgery, Cellularity, Tumour other histologic subtype, Inferred menopausal state.*
It is interesting the fact that the variable *Lymph nodes examined positive* has been shaped by the model to work as a hub, influencing many other variables.
The target variable is dependent only on the variables selected a priori, namely *Pam50 claudin-low subtype, HER2 status, PR status, ER status, Relapse free status, Tumour stage.*

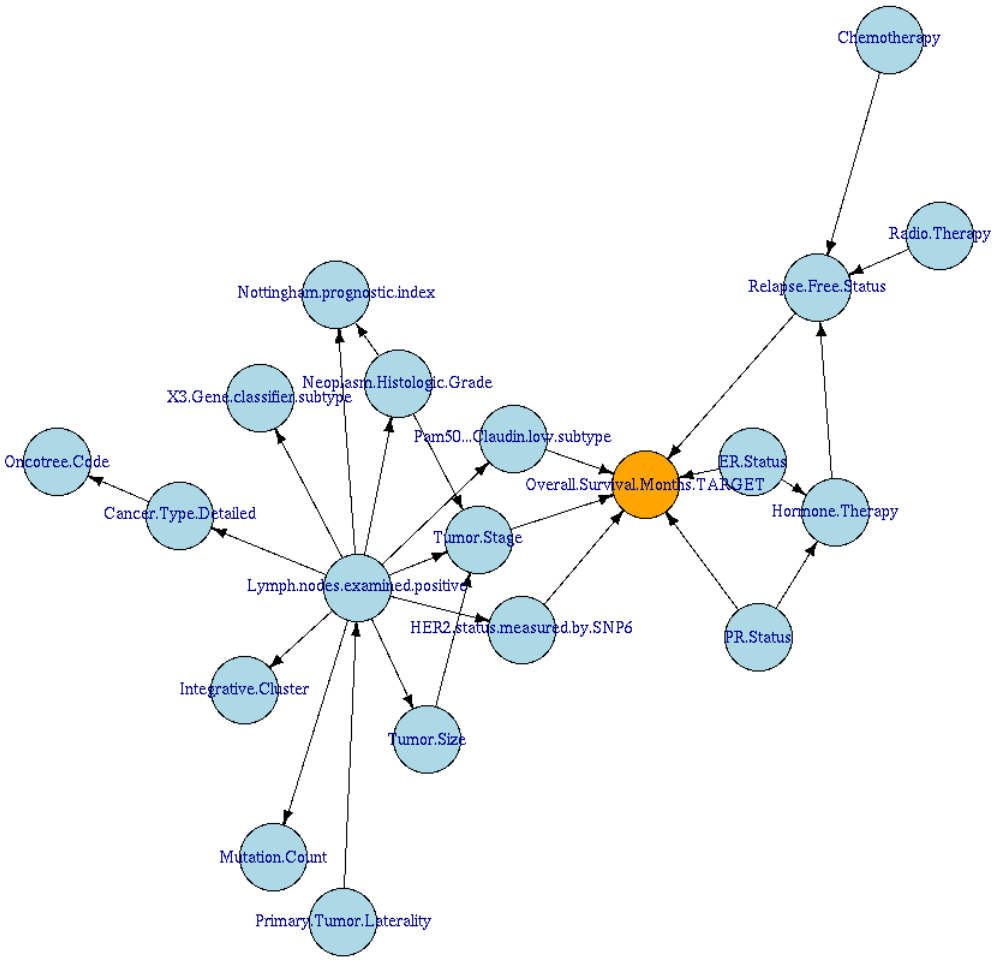


Figure 6: *Final structure of Bayesian network found by using the HC algorithm and AIC score on the version of the dataset containing missing values*

The parameters of the Bayesian network are estimated through the Bayesian parameters' estimation. The fitting is performed on the whole dataset, thus the missing values are imputed as before by exploiting the structure of the model learnt on the training set. The resulting Accuracy is 70.18% and the F1 score is 55.72% when assessed on the original dataset and the same parameters are respectively 68.51% and 59.08% when assessed on the complete dataset. This means that the model performance is consistent.
Different thresholds are tested by using a 10-folds cross-validation in order to improve the metrics. The thresholds tested are the same as before. As for the imputed dataset, the best threshold to use for classification seems to be 0.6, and the corresponding metrics are 73.7% Accuracy and 60.7% F1 score. For the complete version of the dataset, the chosen threshold is 0.7 with an Accuracy of 72.3% and a F1 score of 66%.

As previously done, the ROC curve is plotted by using both the imputed dataset and the original one *(Fig.7)*.
The result is in line with the previous model. The AUC for the version without missing data is 69.3%, which is acceptable, however it is still highly unstable, so it cannot be considered good enough. On the other hand, the AUC for the imputed dataset is rather good, with a value of 82.7%, moreover its variance is much lower.

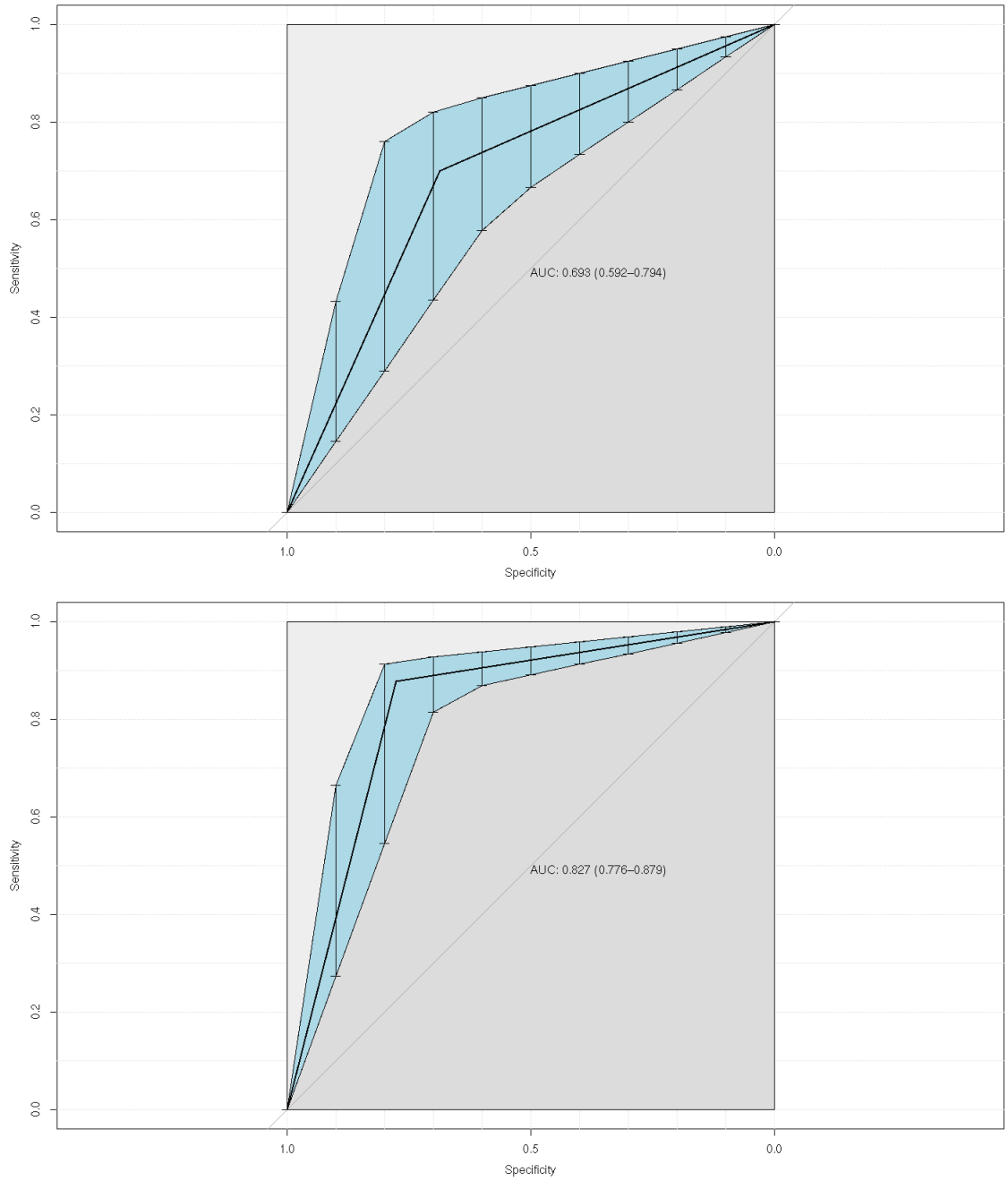


Figure 7: *On the left, the ROC curve for the complete version of the dataset, on the right, the ROC curve for the original imputed dataset*

The two Bayesian networks have two conditional independences in common, namely:

$$Survival \perp\!\!\!\perp Hormone\,therapy \mid (PR\,status, ER\,status, Relapse\,free\,status)$$

$$Survival \perp\!\!\!\perp (Chemotherapy, Radiotherapy) \mid Relapse\,free\,status$$

Moreover, the target variable resulted in being marginally independent from several common variables, which are: *Nottingham prognostic index, Inferred menopausal state, Cancer type detailed, Cellularity, Tumour other histologic subtype, Integrative cluster, Mutation count, Oncotree code, 3 gene classifier subtype.* In total, they are 9 variables.

The variables directly affecting the Survival are the same in both networks.

By comparing the performance of the two models in terms of the AUC of the ROC curve, their results are consistent. They both have a fairly good predictive power on the imputed version of the dataset, while their performance reduces and becomes unstable on the complete dataset. The difference in performance on the first and second dataset probably depends on the fact that the complete version of the dataset contains too few observations for such a large number of variables, so it is not possible to make a robust estimation of the predictive power of the model. Moreover, it is possible that the parameters have not been properly fitted. In fact, the original version of the dataset has almost four times the observations of the complete dataset.
In conclusion, the Bayesian network built on top of the dataset containing also missing values is consistent, since it provides results in line with the other model. Its performance is neither better nor worse. The fact that they don't differ that much is probably due to the fact that the structure of the network has been constrained in advance, since the structure learnt with a total data-driven approach wasn't satisfactory. Actually, the target variable was connected only to one or very few variables, thus the resulting network wouldn't have been useful for clinical support.

Finally, some inference is done in order to retrieve the new conditional probabilities of the target variable *Survival* given the variables which directly affect it. These are the only variables considered for the sake of brevity. The parameters have been fitted by using the whole dataset. As a reference, the marginal probability of $Survival > 5\,years$ is 61% in the first model and 64.2% in the second model.

- *Relapse free status*: it is one of the variables affecting the most the target. The magnitude of the effect is nearly the same in both models. In the first model, the *Survival* rate $> 5\,years$ for people with recurrences is 51.1% and it is 77% for people who don't experience recurrences. Regarding the second model, the values are respectively 53.3% and 75%.

- *Tumour stage*: its effect is consistent in the two models. The early mortality for tumours of stage 1 is unexpectedly high with respect to other stages, even though this should be the less advanced stage of the disease. $Survival > 5\,years$ for stage 1 has a rate of 53.2% in the first model and 51% in the second one, for stage 2 it increases respectively to 62.4% and 74.9%, then it begins decreasing again for stage 3 to 58.5% and 66.3%, and finally for stage 4 it is 50% and 56%. Surprisingly, according to the model, the survival of people with stage 1 cancer is nearly the same as for stage 4 cancer.

- *PR status*: this variable doesn't influence much the target. In fact, $Survival > 5\,years$ is around 61% in the first model and around 63-65% in the second one for both scenarios.

- *ER status*: the variable *ER status* has a considerable impact on the survival on the patient. In fact, $Survival > 5\,years$ is 66.9% and 68% in the two different models when the patient is ER positive, while it is 46% and 54% when the patient is ER negative.

- *HER2 status*: the *HER2 status* has a quite moderate impact on the survival in the second model. When the levels of the HER2 protein are not abnormal, the survival rate is in line with the marginal probabilities of the variable. When there is an over expression of the protein, the rate of people with $Survival > 5\,years$ is 57.6% and 64% in the two models, while when the protein is under-expressed the percentages are respectively 62% and 57%. The effect is the opposite in the two models, but their magnitude is low.

- *PAM 50 and claudin-low subtype*: some of the *PAM 50* classifications have a significant effect on the survival rate, while others such as Normal and Luminal B have a less relevant impact. The variable which increases the most the $Survival > 5\,years$ is the Luminal A, with values 76.6% and 70.7%. The effect of the other classifications goes in the opposite directions in the two models. Regarding the Claudin-low, the rates are 56.1% and 70%, while for the HER2 the values are respectively 42.4% and 61%.

In general, the effect of the variables upon which the *Survival* variable depends is consistent in the two models, with only some differences.

In conclusion, both models have reached a fairly good performance for the survival classification task, however they still have some limits. The limits in their performance and usability for medical support could be possibly overcome by defining a more precise initial structure, built upon the knowledge of real experts in the breast cancer field.

# Bibliography


[1] Zhi-Min Geng et al.
*Estimating survival benefit of adjuvant therapy based on a Bayesian network prediction model in curatively resected advanced gallbladder carcinoma.* World J Gastroenterol, 2019 October 7

[2] Dania Abed Aljawad et al.
*Breast cancer surgery survivability prediction using Bayesian network and support vector machines.* Computer science department, University of Damman, kingdom of Saudi Arabia

[3] Imran Kurt Omurlu et al.
*Comparison of Bayesian survival analysis and Cox regression analysis in simulated and breast cancer data sets.* Trakya university medical faculty, Department of biostatistics, Turkey, 2009

[4] Esin Avci
*Bayesian survival analysis: comparison of survival probability of hormone receptor status for breast cancer data.* Int. J. Data analysis techniques and strategies, Vol.9, No.1, 2017

[5] Peter Lucas
*Bayesian networks in medicine: a model-based approach to medical decision making.* Department of computing science, university of Aberdeen, Scotland UK

[6] Zhi-qiang Cai et al.
*Analysis of prognostic factors for survival after surgery for gallbladder cancer based on a Bayesian network.* Nature scientific reports

[7] Rajpal Nandra et al.
*Can a Bayesian belief network be used to estimate 1-year survival in patients with bone sarcomas?.* Clinical Orthopaedics and Related Research, 2017

[8] K.Jayasurya et al.
*Comparison of Bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy.* Medical Physics, volume 37, issue 4, 2010 for missdata

[9] P.W. Simoes et al.
*Meta analysis of the use of Bayesian networks in breast cancer diagnosis.* Cad. Saude Publuca, Rio de Janeiro, Brasil, jan 2015

[10] A.Endo et al.
*Comparison of seven algorithms to predict breast cancer survival.* Biomedical soft computing and human sciences, vol.13, no.2, 2008

[11] Breast cancer (METABRIC) dataset on Kaggle
`https://www.kaggle.com/gunesevitan/breast-cancer-metabric`.

[12] cancer.net survival rate by stage
`https://www.cancer.net/cancer-types/breast-cancer/statistics`

[13] cancer.net TNM staging system
https://www.cancer.net/navigating-cancer-care/diagnosing-cancer/stages-cancer

[14] WebMD HR and HER2 status on survival
https://www.webmd.com/breast-cancer/guide/her2-positive-breast-cancer-survival-rates

[15] Susan G Komen Pam50
https://www.komen.org/breast-cancer/diagnosis/factors-that-affect-prognosis/pam50-prosigna/

[16] github Oncotree system
https://github.com/cBioPortal/oncotree

[17] cancer.net breast cancer therapy
https://www.cancer.net/cancer-types/breast-cancer/types-treatment