Bayesian-Network updating using novel clinical data

Jeongyeon Huh
University of Twente
Enschede, The Netherlands
j.huh@student.utwente.nl

ABSTRACT

Endometrial cancer is one of the most common cancers affecting women worldwide and exhibits a complex nature with varying patient responses to treatment. The current growing interest in using advanced computational techniques offers promising opportunities to improve prognostic predictions in these complex cases. This study aims to investigate the relationship between endometrial cancer and various biomarkers, analyze possible treatment options, and determine patient-specific probabilities for treatment outcomes by looking into the presence of lymph node metastasis and survival rates. The study utilizes Bayesian networks which can potentially contribute to the development of more accurate and clinically relevant prognostic tools for endometrial cancer patients, improving clinical management and treatment outcomes. The performance of a Bayesian network model by leveraging score-based structure learning and local parameter learning was demonstrated. In the model-building process, the insignificant biomarkers were removed, and new variables were added to more accurately represent endometrial cancer prognosis. The results demonstrate the potential of Bayesian networks to provide personalized prognostic predictions, ultimately enabling clinicians to make better-informed decisions and improve patient outcomes in endometrial cancer treatment.

keywords - Bayesian network, Endometrial cancer, Biomarkers, lymph node metastasis, survival, goodness-of-fit, accuracy, ROC curve, Brier score

1 INTRODUCTION

Endometrial cancer is a significant health concern worldwide, as in 2020, it was ranked as the sixth most common cancer in women with 417,000 newly diagnosed cases globally [1]. With the increasing proportion of the aging population, the number of endometrial cancer patients is expected to increase in the coming decade. This is already noticeable, as the overall incidence of this malignancy has increased by 132% over the past three decades [2].

However, despite the development of new treatment methods, the overall mortality rate for endometrial cancer has not improved significantly. Therefore, researchers from the Department of Gynecology and Obstetrics at Radboud University Medical Centre (RadboudUMC) collected clinical data from patients with endometrial cancer over the last few years. The resulting dataset was used to develop a Bayesian network, called ENDORISK, to forecast the prognosis of the ailment in patients [3]. The model performance was evaluated using data from patients in other countries, and it was demonstrated that the model performs quite well [4, 5].

Although the present model has demonstrated promising outcomes, it has not integrated the latest biomarkers and clinical management variables representing cutting-edge clinical management of patients with the disease. These variables are absent in the collected data so far on account of non-inclusion in the study yet. The majority of these variables pertain to biomarkers, which concern specific molecules that potentially correlate with certain disease outcomes.

The focus of the researchers is primarily on such biomarkers that:

- foresee the presence of lymph node metastasis in the abdomen (mainly in the pelvic region and surrounding the aorta) as a result of the tumor spreading beyond its primary site;
- anticipate the survival of endometrial cancer patients after undergoing surgery and the additional treatment by radiotherapy and chemotherapy.

The study aims to investigate the relationship between endometrial cancer and various biomarkers, then analyze the available treatment options and their results with probability. By updating the Bayesian network and identifying the presence of lymph node metastasis and survival rates, the goal is to contribute to the development of a more accurate and clinically relevant prognostic tool for endometrial cancer patients. The researcher hopes this improves clinical management and outcomes for women with endometrial cancer and finds better treatment options per patient.

To achieve the overall goals, the objectives of the study are:

(1) To update the Bayesian network model for endometrial cancer prognosis prediction that incorporates significant biomarkers predictive of lymph node metastasis, and survival after treatment.
(2) To compare the performance of the updated model against the existing model.
(3) To investigate the new variables for improving the performance of the updated model.

With this, the following research questions will serve as the foundation for investigation and shaping the structure of the research to accomplish the stated goals:

(1) What particular biomarkers from the existing models have no significant impact that needs to be replaced with alternative biomarkers?
(2) How does the updated Bayesian network model demonstrate improved performance in predicting the prognosis of endometrial cancer patients?
(3) How do missing values and the small dataset size affect the performance of the Bayesian network model?

2 BACKGROUND

This subsection provided background on the necessary concepts of biomarkers, and Bayesian networks to be used in the research.

2.1 Endometrial Cancer and Biomarkers

Endometrial cancer is a common and often aggressive malignancy that affects the female genital organs, specifically the inner lining of the uterus known as the endometrium. This complex disease predominantly occurs in postmenopausal women and presents diverse histological types, varying degrees of tumor differentiation, and variable patient responses to treatment making prognostic prediction particularly challenging. Although traditional clinicopathological factors, such as tumor stage, grade, and the presence of lymphovascular space invasion, play a significant role in determining patients' prognoses, these factors alone may not be sufficient for accurately predicting disease recurrence and overall survival [6]. The limitations of the current clinicopathological-based prognostic models for endometrial cancer emphasize the urgency to identify novel, more reliable biomarkers that can enhance the prognostic accuracy.

Recent studies have focused on molecular and genetic markers, such as microsatellite instability, mutations in POLE and p53 genes, and hormone receptor status, for potential prognostic indicators. A few indicative lymph node metastasis biomarkers of endometrial cancer patients from the basis Bayesian network model are discussed here.

Biomarker is a measurable substance in a biological system that indicates a specific biological condition, disease, or treatment response. This aids in diagnosis, prognosis, and treatment monitoring.

The tumor suppressor protein, p53, has a critical function in cell multiplication and apoptosis control. The link of p53 mutations with unfavorable prognoses, such as lymph node metastasis and a decrease in overall survival in patients with endometrial cancer, are reported [8]. In addition, biomarkers such as serum CA-125, lymphadenopathy, and tumor size are identified to be correlated with a heightened likelihood of lymph node metastasis among individuals with endometrial cancer [9].

The depth of myometrial invasion is predictive of lymph node metastasis [10, 12], along with estrogen receptor, and progesterone receptor, which are commonly assessed in pathology reports. They offer valuable insights into the likelihood of lymph node metastasis estimation and overall survival after treatment [12]. Nevertheless, further research is needed to validate their predictive power and determine their potential clinical utility.

2.2 Bayesian Networks

A *Bayesian network* $\mathcal{B} = (G, P)$ is a joint probability distribution P with (conditional) independence constraints, represented as a directed acyclic graph $G = (V, A)$, with $V = \{1, \ldots, n\}$ a set of nodes and $A \subseteq V \times V$ a set of arcs or directed edges [7]. Nodes $v \in V$ in a Bayesian network correspond $1-1$ to variables X_v in the joint probability distribution P as follows:

$$P(X_1, \ldots, X_n) = \prod_{v \in V} P(X_v \mid X_{\pi(v)})$$

where $\pi(v)$ represents the *parents* of node $v \in V$.

The arcs are often interpreted as causal relationships. Common probabilistic operations on Bayesian hetworks are *marginilization* $P(X_v) = \sum_{x_w : w \in V \setminus \{v\}} P(X_1 = x_1, \ldots, X_n = x_n)$ and *conditioning* $P(X_v \mid E) = P(X_v, E)/P(E)$, where E is a set of instantiated variables, called *evidence*. Both types of probabilistic inference can be done by software such as GeNie, which also offer nice visual graphical representations of the resulting marginal probability distributions (cf. Figure 7). Both Bayesian network graph structure and probability distribution can be learned from data using *score-based structure learning* [11]. Specifically hill-climbing and tabu-search were well-known score-based learning algorithms.

To conclude whether one variable is significant or not, its place in the graph is important to know. In particular one needs to check where a node participates in a diverging, serial, or converging connection as shown in Figure 1.

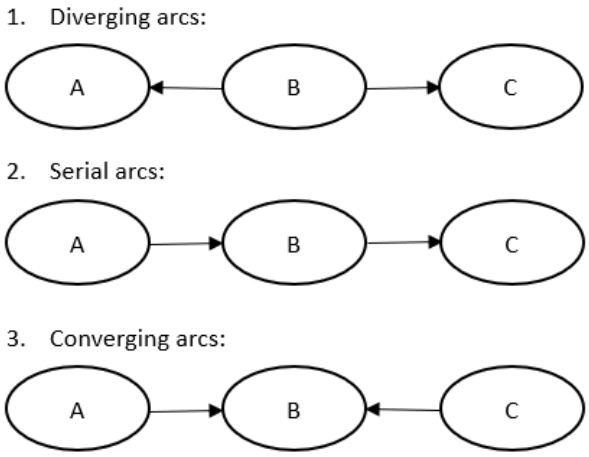


Figure 1: Basic connections of Bayesian network that represent causal relations

The effect of instantiating the central vertex on the flow of probabilistic information was examined starting with converging and serial arcs and then lastly diverging arcs.

Bayesian network modeling techniques have demonstrated remarkable precision in forecasting lymph node metastasis and post-treatment survival for patients with endometrial cancer [4, 5]. A systematic analysis of research spanning the past thirty years revealed that machine learning approaches, incorporating Bayesian network modeling, have been effectively utilized in prognostic estimation for gynecological cancers [13].

A Bayesian network model incorporating different types of data was found to have higher accuracy in predicting cancer subtypes and survival outcomes than models using only one type of data [14].

These studies emphasize the potential of Bayesian network modeling techniques in enhancing cancer prognosis prediction and personalized treatment planning and the integration of multiple levels of patient data to develop more comprehensive and robust prognostic models.

Bayesian network models are widely used to predict disease prognosis because of a probabilistic framework that handles uncertainties in medical data while accurately capturing the complex relationships and dependencies between various variables such as risk factors, symptoms, and disease progression.

3 RELATED WORK

This section deals with studies related to the survival prediction prognosis of endometrial cancer patients.

Many other studies are done on the prognostic survival assessment of endometrial cancer patients. Wan et al. [15] identified several prognostic factors significant to recurrence and survival rates. These factors are age, tumor grade, tumor stage, depth of myometrial invasion, lymph node involvement, and presence of lymphovascular space invasion. In their study, a prognostic model was developed by combining multiple clinical, histological, or molecular variables. AlHilli et al. [16] conducted a similar study. They collected the data in two different groups, the development cohort, and the validation cohort. The identified factors include age, race, tumor grade, FIGO (International Federation of Gynecology and Obstetrics) stage, lymphovascular space invasion, and lymph node status factors. They used a risk-scoring model in which risk scores are obtained from the assigned points based on hazard ratios for risk factors. They further explored the relationship between survival, risk scores, and which treatments the patients can benefit from. Both of the studies were done with the same goal as this study and have identified some factors covered in this study. They were all found to be valuable to be used in the clinical field with some modifications needed just like the basis Bayesian network model for this study. The main difference between these studies is the model used. The Bayesian network model is used in this study.

4 METHODOLOGY

In this section, the detailed steps taken to conduct this study are discussed. Throughout the project, R programming language (4.3.0) with bnlearn and mice packages was used to write programs and develop networks, and GeNie was used to have the graphical representations.

4.1 Data

The researchers from RadboudUMC provided the complete cohort with a total of 952 patients including the training cohort with additional Cancer Genome Atlas (TCGA) data. In the data, only patients diagnosed by an expert gynecological pathologist, with complete clinical and pathological data and follow-up of at least 36 months were included. Since the data was obtained from patients with endometrial cancer anonymously, neither ethical consideration nor informed consent was needed. The data include variables from immunohistochemical analysis of endometrial biopsies. Among those, only selective variables are included in the Bayesian network and a few more variables are to be added.

The data was cleaned by creating a subset with selective variables and replacing blank cells with not applicable 'NA'. The values were changed to more meaningful ones for easy and straightforward interpretation such as 0 and 1 to positive and negative depending on the variable explanation. The subset was used to create two separate data which are 1) data with variables included in the given Bayesian network and 2) another with variables to be newly added (FIGO, MRI_MI, MSI, and POLE) to the Bayesian network which was later combined with its parent nodes. Then, imputation was done using a bnlearn package to the first data set to replace the missing data. For the extra variables to be added, due to many missing values, only cells with records were considered separately per variable. Instead of imputation, local parameter learning was done. Using the data of the node and its parent node(s), the conditional probability table (CPT) was computed with the code as shown in Figure 1.

```
library(bnlearn)

FIGO = read.csv2("Results/combined_FIGO.csv")
MRI = read.csv2("Results/combined_MRI_MI.csv")
MSI_POLE = read.csv2("Results/combined_MSI_POLE.csv")

# Compute new FIGO CPT
count_FIGO <- table(FIGO$PostoperativeGrade, FIGO$FIGO)
new_cpt_FIGO <- prop.table(count_FIGO, 1)

names(dimnames(new_cpt_FIGO)) <- c("PostoperativeGrade", "FIGO")

# Compute new MSI CPT
count_MSI <- xtabs(~ MSI + PostoperativeGrade + LNM, MSI_POLE)
laplace_constant = 1
smoothed_count_MSI <- count_MSI + laplace_constant
new_cpt_MSI_smoothed <- prop.table(smoothed_count_MSI, c(2, 3))

names(dimnames(new_cpt_MSI_smoothed)) <- c("MSI", "PostoperativeGrade", "LNM")
```

Figure 2: Code to compute conditional probability table (CPT)

Depending on the number of the parent nodes, either xtabs() or table() function was used to count each combination of values. Then, prop.table() was used to estimate the new CPT as the conditional probabilities, dividing the count table by the sum of each row. NaN values were encountered for MSI and POLE nodes that the Laplace smoothing technique was used to add a constant of 1 to each count to avoid division by zero.

4.2 Inspection of the Provided ENDORISK Bayesian Network

The Bayesian network was provided to use as a basis. With the additional variable TP53, replacing p53 or keeping both p53 and TP53 variables was considered. Then, chance nodes and arcs were added for additional variables, and their probability was updated with the result of CPT calculations.

The first research question is all about identifying insignificant biomarkers from the basis Bayesian network. To conclude whether one variable is significant or not, a few things were checked. This includes identifying interested variables and classifying the arcs of each node connected to other nodes by either diverging, serial, or converging. Depending on the connection type, either inclusion of the target variable or probability change by giving values to the node was analyzed to deduce the insignificant variables. Then finally, the variable was once again classified between pathological variables and biomarkers. Only the variable identified as a biomarker was removed from the network.

4.3 Model Performance

This research question demonstrates the improvement in the performance of the updated Bayesian network model. The performance of the Bayesian network model was evaluated using goodness-of-fit,

Brier score, and area under the Receiver operating characteristic (ROC) curves. Since these measures were also used to evaluate the provided Bayesian network, the performance of only the updated model was evaluated and then compared. The log-likelihood, sensitivity, and specificity with different thresholds were calculated to establish accuracy and ROC curve plots with the confusion matrix.

Table 1: Confusion matrix.

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

The formulas used with confusion metrics are below:
Sensitivity = $\frac{TP}{TP+FN}$
Specificity = $\frac{TN}{FP+TN}$
Positive predictive value = $\frac{TP}{TP+FP}$
Negative predictive value = $\frac{TN}{TN+FN}$
False negative rate = $\frac{FN}{FN+TP}$
ROC curve plot = $\frac{sensitivity}{1-specificity}$

4.4 Possible Influences on Model Performance

The last research question is to see how the missing values and small dataset size affect the performance of the model. Multiple imputation and performance metrics were used to evaluate the influence on performance.

For the missing data, multiple imputation, creating several complete datasets with imputed values, was done using a mice package. This is to account for the uncertainty introduced by the imputation process. This is because resorting to single imputation can lead to underestimation of the variance and biased parameter estimates assuming the imputed values are correct. Bayesian networks were constructed as described in section 4.2. Then compare the model performances across the multiple imputed datasets by calculating the performance metrics shown in section 4.3. However, multiple imputation was done to answer this research question only.

The performance metrics were also used for the small dataset. Using the imputed data by bnlearn package, the different subsets of 238, 476, 714, and the whole data (952) were created using random sampling. The Bayesian network models were trained with different subsets (sizes) of data. Then chosen performance metrics, Brier score, sensitivity, and specificity, were calculated for each of the models. The changes in the calculated values were analyzed.

5 RESULTS

Building on the discussion of the methodology, the section 5 provides a comprehensive analysis of the results obtained and presents the results.

5.1 Data

Among the total cohort of 952 patient data, only 202 cases had complete information on all variables. Imputed data was generated with bnlearn package in R (4.3.0). For additional variables, only non-empty records were considered. The summary of additional variable data with the total considered records can be found in Table 2.

Table 2: Additional variable data summary.

MRI_MI	POLE	MSI	FIGO
lt_50 : 65	no : 409	no : 352	IA : 532
ge_50 : 37	yes : 35	yes : 92	IB : 228
			II : 64
			IIIC : 54
			IIIA : 23
			IVB : 23
			Other : 8
Total : 102	Total : 444	Total : 444	Total : 932

5.2 Updated Bayesian Network

The updated Bayesian network can be found in Figure 7 in the Appendix. From the basis Bayesian network, every variable was carried over except p53, and the additional variables (FIGO, MRI_MI, MSI, and POLE) were included. The protein variable, p53, was replaced with the gene variable, TP35, since both of them measure the protein, not the gene. This means they are the same variable. In the cleaned data, p53 has 257 missing values while none for TP53. Due to the completeness of TP53, it was chosen instead of p53.

5.3 Insignificant Variable Identification

The interested or target variables are lymph node metastasis (LNM) and the survival rates, especially after 5 years.

The causal relationships of nodes were classified as diverging, serial, or converging. For diverging and serial arcs, when the central node is not initiated, knowing one edge node can tell information about the other edge node. Therefore, the node is considered significant if at least one of the target variables is connected as the edge node with diverging or serial arc. These nodes are MyometrialInvasion, L1CAM, PR, ER, PostoperativeGrade, CA125, LVSI, Recurrence, Chemotherapy, Radiotherapy, and Survival3yr.

For the converging arcs, the probabilistic information transmits from one edge node to another only if the central node is known. The nodes with converging arcs connected to at least one of the interested variables can become significant only after the value is given to the central node. The possible significant nodes include p53, L1CAM, PR, ER, CTMRI, Recurrence, Survival1yr, and Survival3yr. To decide the significant variable, they were given values and the probability changes were analyzed.

Only variables that made more than 0.1 probability difference to the target node were considered significant. These are p53, L1CAM, ER, CTMRI, Recurrence, Survival1yr, and Survival3yr. For survival rates, the probability affects the survival5yr when the value "no" is entered since there is propagated probability making survival5yr no = 1 and yes = 0 automatically.

Table 3: Probability changes of insignificant variable candidates.

		Target variables	
		LNM	Survival5yr
p53	wildtype	0.022	0.01
	mutant	0.112	0.052
L1CAM	negative	0.021	0.008
	positive	0.177	0.065
PR	negative	0.068	0.036
	positive	0.015	0.008
ER	negative	0.124	0.058
	positive	0.014	0.007
CTMRI	no	0.014	0.004
	yes	0.266	0.073
Recurrence	no	0.029	0.062
	regional_distant	0.266	0.496
	local	0.031	0.158
Survival1yr	no	0.254	0.932
	yes	0.005	0.018
Survival3yr	no	0.249	0.932
	yes	0.013	0.047

Then, the only variables not identified as significant so far are Platelets, PreoperativeGrade, and Cytology. Finally, these variables are classified as either pathological or biomarker making only Platelets to be insignificant. Platelets only have an incoming arc from LNM which is one of the target variables and no other incoming or outgoing arc. Since LNM is the target variable, no value is given to LNM that Platelets do not affect the Bayesian network model.

5.4 Model Performance

To check how well the model fits the data, three different log-likelihood scores were computed to compare the Bayesian network model. To use as the baseline comparison, the log-likelihood score on the original model with complete cases of initial data was calculated as -1419.797. The score showing the generalization of learned structure from imputed data to the complete cases was -1444.041. Finally, model fit with the learned structure from imputed data on the imputed data was found to be the lowest with -5923.701. The similar scores of the original model and learned structure from imputed data tested with complete cases of initial data show that the learned structure performs well to capture the underlying relations in the not-imputed data. However, having a much lower score on imputed data learned structure tested on imputed data compared to the others indicates the issue with the imputation process or possible discrepancy in the complete cases of initial data and the imputed data.

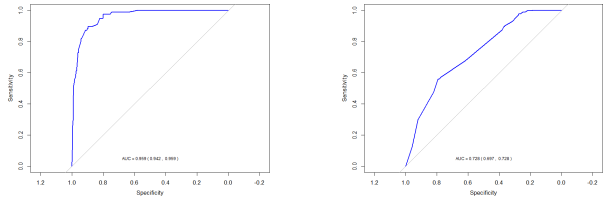
To test the accuracy of probabilistic predictions, Brier scores were computed. The closer the score is to 0, means less difference between the predicted probabilities and the true outcomes, indicating perfect or better predictions. The highest Brier score is 1 which represents the worst possible prediction with the maximum difference. A Brier score near 0.5 implies that the model predicts close to a 50-50 chance for binary outcomes which is similar to random guessing.

The updated BN shows Brier scores of 0.474 for LNM and 0.299 for Survival5yr which were 0.09 and 0.12 in the original model respectively. The updated model has no accurate predictions for the LNM variable and the Survival5yr variable has more accurate predictions than LNM. However, in comparison to the original model, the accuracy is significantly lower.

Table 4: Concordance statistics of the BN.

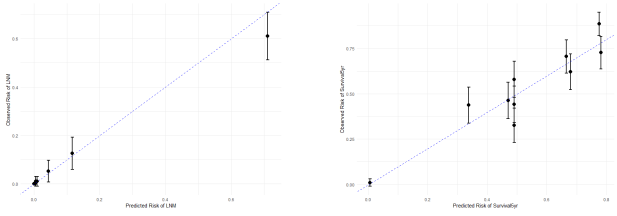
	LNM	Survival5yr
AUC (95%-CI)	0.959 (0.942, 0.959)	0.728 (0.697, 0.728)
Brier score	0.474	0.299
Predicted N of events	94	836
Observed N of events	77	494
Predicted/Observed ratio	1.221	1.692

In addition to the Brier score, the confusion matrix was analyzed to plot the ROC curves. The AUC was 0.959 with 0.942 to 0.959 of 95% confidence interval for LNM which indicates a good discriminatory power in distinguishing the presence of LNM. The 95% confidence interval suggests that the true AUC value falls within the range. On the other hand, the AUC value of 0.728 with a 95% confidence interval of (0.697, 0.728) for Survival5yr indicates a moderate discriminatory power in predicting the survival rate. Although the lower bound of the confidence interval is still relatively high, the wide range suggests that the true AUC value may have some uncertainty and is likely to vary around the point estimate.



(a) LNM **(b) Survival5yr**

Figure 3: ROC curve plot



(a) LNM **(b) Survival5yr**

Figure 4: Scatter plot with Preoperative Grade

The AUC and Brier scores show contradictory results. The AUC for LNM indicates a strong performance, while the Brier score reflects poor performance. Likewise, the AUC for the Survival5yr

suggests a moderate performance, but the Brier score indicates a relatively better performance. This discrepancy may be due to the different aspects emphasized by these two metrics. The AUC focuses on rank prediction and analyzes the ability of a model to discriminate between positive and negative classes by examining the true positive rate and the false positive rate at different thresholds. On the other hand, the Brier score is an appropriate scoring rule to evaluate the accuracy of probabilistic predictions. Measures the mean squared difference between predicted probabilities and actual outcomes. In this research. the Brier score is a more informative metric to consider as the quality of the predicted probabilities and overall model calibration are essential.

Table 5: Diagnostic accuracy values for the prediction of the target variables using various cut-off values.

Cut-off	1%	5%	10%	15%	20%	25%
LNM						
Sensitivity	0.987	0.948	0.883	0.818	0.818	0.740
Specificity	0.737	0.810	0.897	0.941	0.942	0.958
PPV	0.248	0.305	0.430	0.548	0.553	0.606
NPV	0.998	0.994	0.989	0.983	0.983	0.977
Predicted Positive	0.081	0.081	0.081	0.081	0.081	0.081
FNR	0.013	0.052	0.117	0.182	0.182	0.260
Survival5yr						
Sensitivity	1	0.998	0.998	0.998	0.998	0.988
Specificity	0.164	0.197	0.221	0.221	0.221	0.240
PPV	0.563	0.573	0.580	0.580	0.580	0.584
NPV	1	0.989	0.990	0.990	0.990	0.948
Predicted Positive	0.519	0.519	0.519	0.519	0.519	0.519
FNR	0.000	0.002	0.002	0.002	0.002	0.012

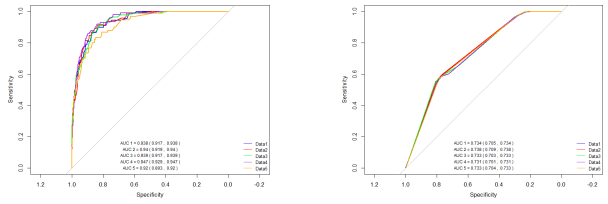
5.5 Possible Influence on Model Performance

To realize the influence of missing values in the data set and small-size data sets, the accuracy and the ROC curve plots were analyzed. With the mice package, 5 different imputed data sets were generated. For LNM, the Brier scores lie between 0.461 and 0.475 while it lies between 0.3 and 0.303 for Survival5yr. The difference in Brier scores of the updated model and models with multiple imputation datasets are relatively consistent.

Table 6: Brier scores for each dataset of multiple imputation.

data set	LNM	Survival5yr
1	0.462	0.303
2	0.461	0.302
3	0.467	0.302
4	0.47	0.301
5	0.475	0.3

The updated model has an AUC of 0.959 for LNM and 0.728 for Survival5yr, while the AUCs for the multiple imputation models are relatively close to these values. For LNM, the AUCs range from 0.92 to 0.947, and for Survival5yr, they range from 0.731 to 0.738. Generally, the AUCs of the multiple imputation models suggest a reasonably consistent performance across the imputed datasets. In terms of 95% confidence intervals (CI), the updated model has a CI of (0.942, 0.959) for LNM and (0.697, 0.728) for Survival5yr. The CI for LNM is narrower compared to Survival5yr, indicating that the prediction performance is more consistent for LNM. The confidence intervals of AUCs for the multiple imputation models overlap with the updated model's CI, suggesting that the performances of these models are comparable.



(a) LNM **(b) Survival5yr**

Figure 5: ROC curve plots with multiple imputation data sets

Based on the Brier scores and AUCs alongside their CIs, the impact of missing data imputation on predictive performance is not substantial, the imputation process has adequately addressed the issue, and the models are relatively stable and less sensitive to variations in the dataset indicating robustness in the performance. While the missing values' influence may not be substantial, employing multiple imputation provides a more accurate representation of the uncertainty in the model's performance due to missing data.

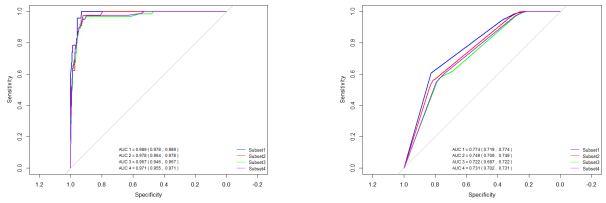
The same approach was done to analyze the influence of small sizes data sets but with a random sampling of one imputed data into the 4 different subsets with different sizes. For LNM, the Brier scores show a general decreasing trend (improving performance) with increasing dataset size. For Survival5yr, the Brier scores consistently decrease (improve) as the dataset size increases. It appears that the performance slightly improves with larger datasets, but it's fairly consistent across different dataset sizes since the largest dataset is still a small dataset.

Table 7: Brier scores for small size datasets.

data set	LNM	Survival5yr
1	0.482	0.325
2	0.462	0.304
3	0.451	0.301
4	0.48	0.3

For LNM, the AUC values suggest stronger performance with smaller-sized datasets. This is an interesting finding as it might indicate potential overfitting, noise, or other confounding factors in the larger dataset which was actually used to update the Bayesian network model. For Survival5yr, the AUC values show a mild decrease (reduced performance) as the dataset size increases, suggesting that smaller datasets surprisingly perform better in this case. The

95% confidence intervals are narrower for the smaller datasets, indicating more certainty in their performance for both LNM and Survival5yr target variables.



(a) LNM **(b) Survival5yr**

Figure 6: ROC curve plots with small size data sets

6 CONCLUSION

The main objective of this study was to update the BN model and evaluate its performance in comparison to the given model. The BN model was updated by removing the insignificant variable and adding new variables. Platelets were identified as insignificant making no changes to the target variables. It had no outgoing arc and only one incoming arc connected to LNM which is one of the target variables. The protein variable (p53) was replaced by the gene variable (TP53) according to the completeness of the data.

Despite the AUC and Brier scores showing contradictory results, it was clear the performance of the updated BN model is fairly low compared to the given BN model. The AUC showed the performance to be strong for LNM and moderate for Survival5yr while the Brier score showed poor and relatively better performance respectively. Given the specific context of prognosis prediction in endometrial cancer, the Brier score is a more important metric taking into account both discrimination and calibration.

The analysis of Brier scores and AUC values along with their 95% confidence intervals indicates that the influence of missing values on model performance is minimal, and the imputation process has effectively addressed the issue. The models derived from multiple imputed datasets display consistent performance with the updated model, demonstrating the models' stability and robustness. The BN model seems to perform fairly well even with small datasets, but smaller datasets show better performance results.

Taking everything into account, this model has stable model performance but still a room for improvements. Since it is proved that the using imputation for missing values and small data sets do not influence the performance heavily but the imputation affect the goodness-of-fit. This model would still need further research to improve the accuracy in order to be used for individualizing decision-making on endometrial cancer in clinical field.

7 FUTURE WORK

For the newly added variables, local parameter learning was done rather than imputation because of the large number of missing values. More data should be collected to improve the prediction and to handle all the data in the same way. As shown in the log-likelihood score comparison, the quality of data imputation needs to be investigated. The first approach would be trying different methods or even doing multiple imputation together as done in research question 3 to not underestimate the variance and no biases to parameters. Refining the imputation technique and adjusting the parameters may improve the overall performance of the model, better capture the relationships between variables, and achieve a higher log-likelihood score for the whole imputed dataset. The possibility of factors such as noise, overfitting, or other confounding factors in the larger datasets introduced in the collected data needs to be examined. Additional investigation, such as exploring different modeling techniques or addressing potential biases in the data, could provide further insights into the model's performance at varying dataset sizes.

REFERENCES

[1] E. J. Crosbie, S. J. Kitson, J. N. McAlpine, A. Mukhopadhyay, M. E. Powell, and N. Singh, "Endometrial cancer," The Lancet, vol. 399, no. 10333, pp. 1412–1428, Apr. 2022, doi: https://doi.org/10.1016/S0140-6736(22)00323-3.

[2] B. Gu, X. Shang, M. Yan, et al (2021). Variations in incidence and mortality rates of endometrial cancer at the global, regional, and national levels, 1990–2019. *Gynecol Oncol* 2021; 161 pp. 573-580.

[3] C. Reijnen, E. Gogou, …, P.J.F. Lucas, J.M.A. Pijnenborg (2020). Preoperative risk stratification in endometrial cancer (ENDORISK) by a Bayesian network model: A development and validation study. *PLOS Medicine* 2020; 17(5): e1003111.

[4] M. Grube, C. Reijnen, P.J.F. Lucas, F. Kommoss, F.K.F. Kommoss, S.Y. Brucker, ... Improved preoperative risk stratification in endometrial carcinoma patients: external validation of the ENDORISK Bayesian network model in a large population-based case series. Journal of Cancer Research and Clinical Oncology (Springer), 1-9, 2022, DOI 10.1007/s00432-022-04218-4.

[5] P. Vinklerova, P. Ovesna, J. Hausnerova, J. Pijnenborg, P.J.F. Lucas, ... External Validation Study of Endometrial Cancer Preoperative Risk Stratification (ENDORISK) Model. Frontiers in oncology, 3895, 2022, DOI 10.3389/fonc.2022.939226.

[6] P.V. Dludla, B.B. Nkambule, B. Jack, Z. Mkandla, T. Mutize, S. Silvestri, P. Orlando, L. Tiano, J. Louw, and S.E. Mazibuko-Mbeje (2019). Inflammation and Oxidative Stress in an Obese State and the Protective Effects of Gallic Acid. Nutrients, 11(1), 23. DOI: 10.3390/nu11010023

[7] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann, 1988.

[8] C. Zhang, and X. Hao (2019). Prognostic significance of CD276 in non-small cell lung cancer. Open Medicine, 14(1), 805-812. https://doi.org/10.1515/med-2019-0076

[9] C. Reijnen, J. IntHout, L. Massuger, F. Strobbe, H. Küsters-Vandevelde, I. Haldorsen, M. Snijders, and J. Pijnenborg (2019). Diagnostic Accuracy of Clinical Biomarkers for Preoperative Prediction of Lymph Node Metastasis in Endometrial Carcinoma: A Systematic Review and Meta-Analysis. The Oncologist, Volume 24, Issue 9, Pages e880–e890. DOI: 10.1634/theoncologist.2019-0117

[10] Y. Zhang, H. Huang, H. Li, Y. Li, and X. Li (2018). Prediction of lymph node metastasis in endometrial cancer using conventional MRI and DWI. Journal of magnetic resonance imaging, 47(5), 1319-1326. DOI: 10.1002/jmri.25823.

[11] Marco Scutari, Learning Bayesian Networks with the bnlearn R Package, Journal of Statistical Software, 35(3), 2010.

[12] L. Shan, J. Wang, Y. Wang, R. Hu, and R. Chen (2021). High expression of CXCL10 is associated with lymph node metastasis in endometrial carcinoma. Aging, 13(1), 465-478. DOI: 10.18632/aging.202230.

[13] J. Sheehy, H. Rutledge, U.R. Acharya, H.W. Loh, R. Gururajan, X. Tao, X. Zhou, Y. Li, T. Gurney, and S. Kondalsamy-Chennakesavan (2023). Gynecological cancer prognosis using machine learning techniques: A systematic review of the last three decades (1990-2022). Artificial Intelligence in Medicine, vol. 139, pp. 102536, 2023. DOI: 10.1016/j.artmed.2023.102536.

[14] M. Zhang, Y. Wang, Y. Wang, L. Jiang, X. Li, H. Gao, M. Wei, and L. Zhao (2020). Integrative Analysis of DNA Methylation and Gene Expression to Determine Specific Diagnostic Biomarkers and Prognostic Biomarkers of Breast Cancer. Frontiers in Cell and Developmental Biology, 8. DOI: 10.3389/fcell.2020.529386

[15] Y. L. Wan et al., "Prognostic models for predicting recurrence and survival in women with endometrial cancer," Cochrane Database of Systematic Reviews, vol. 2021, no. 6, Jun. 2021, DOI: 10.1002/14651858.cd014625.

[16] M. AlHilli, L. Rybicki, C. Carr, M. Yao, S. Amarnath, R. Vargas, R. Debernardo, C.Michener, P.G. Rose (2021). Development and validation of a comprehensive clinical risk-scoring model for prediction of overall survival in patients with endometrioid endometrial carcinoma. Gynecologic Oncology, vol. 163, no. 3, pp. 511-516. DOI: 10.1016/j.ygyno.2021.09.008.

[17] J. M. Ordovas et al., "A Bayesian network model for predicting cardiovascular risk," Computer Methods and Programs in Biomedicine, vol. 231, p. 107405, Apr. 2023, DOI: 10.1016/j.cmpb.2023.107405.

Appendices

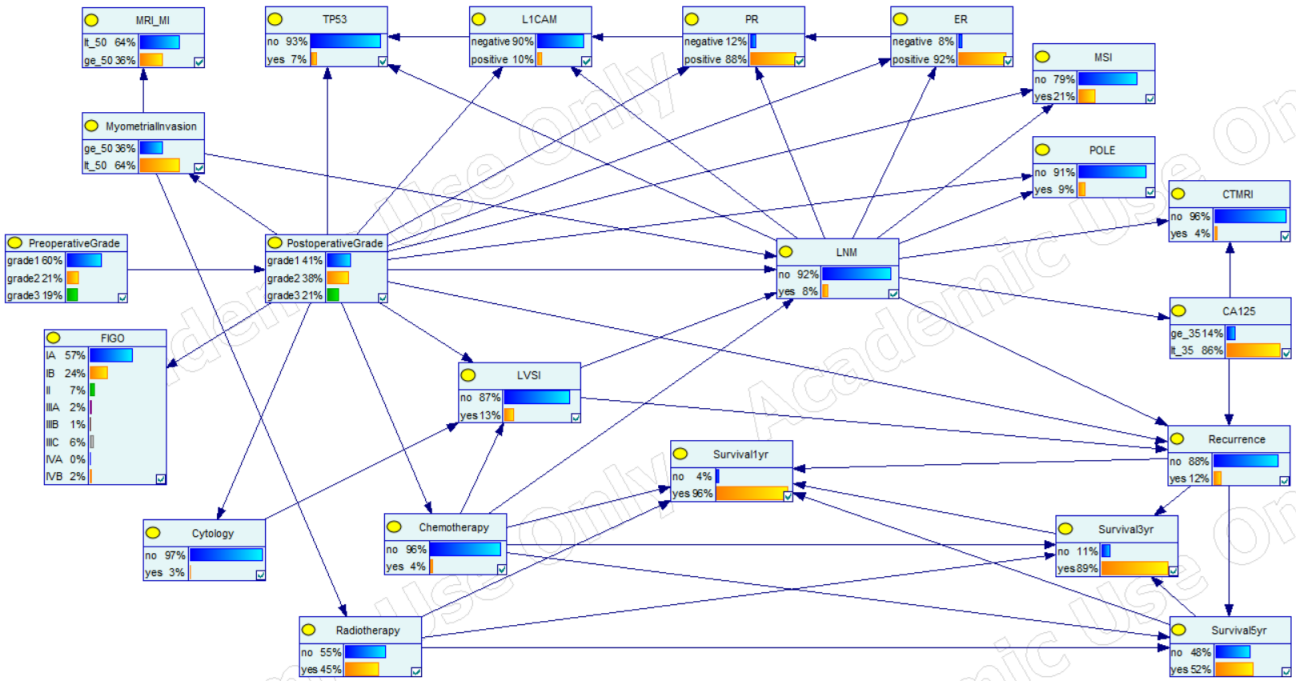


Figure 7: Visual representation of the updated Bayesian network.