

## FROM RISK FACTORS TO PREDICTION: BINARY LOGISTIC REGRESSION FOR ENDOMETRIAL CANCER IN A SIMULATED CLINICAL DATASET

Cristina Gena Dascălu<sup>1</sup>, Doriană Agop-Forna<sup>2\*</sup>, Cristina David<sup>3</sup>, Magda Ecaterina Antohe<sup>4</sup>

<sup>1</sup>Medical Informatics and Biostatistics Dept., Faculty of Medicine, "Grigore T. Popa" University of Medicine and Pharmacy, Iași, Romania, [cristina.dascalu@umfiasi.ro](mailto:cristina.dascalu@umfiasi.ro)

<sup>2</sup>Dento-alveolar and Maxillo-Facial Dept., Faculty of Dental Medicine,

"Grigore T. Popa" University of Medicine and Pharmacy, Iași, Romania, [doriana.forna@umfiasi.ro](mailto:doriana.forna@umfiasi.ro)

Obstetrics, Gynaecology and Neonatology Dept., Faculty of Medicine, "Grigore T. Popa" University of Medicine and Pharmacy, Iași, Romania, [cristina.david@umfiasi.ro](mailto:cristina.david@umfiasi.ro)

<sup>4</sup>Implantology and Prosthetic Implant Rehabilitation Dept., Faculty of Dental Medicine,

"Grigore T. Popa" University of Medicine and Pharmacy, Iași, Romania, [magda.antohe@umfiasi.ro](mailto:magda.antohe@umfiasi.ro)

Corresponding author: Doriană Agop-Forna, [doriana.forna@umfiasi.ro](mailto:doriana.forna@umfiasi.ro)

### Abstract

In this paper we discuss about a well-known method for qualitative data analysis – the binary logistic regression. We present the theoretical foundations of this method, conditions for its use and implementation aspects in a statistical software (SPSS 29.0), including predictor selection and processing strategies, model validation via significance tests and model quality evaluation. These theoretical issues are illustrated using a realistic AI-generated dataset of 400 cases of endometrial cases, analysing eight risk factors widely reported in the literature. Two predictive models are compared: one using predictors in their original form, and another based solely on statistically significant binary predictors. The results interpretation is carried out, in order to illustrate in detail how to conduct such an analysis in clinical studies. The example provided aims to emphasize the need for careful interpretation of results and outlines the method's limitations, highlighting also its practical value.

**Key words:** binary logistic regression, qualitative data, predictors.

### Introduction

One of the main objectives of an analytical statistical study is to investigate the relationship between two or more statistical variables. Depending on the nature of these variables (qualitative or quantitative), specific investigation methods are used.

The simplest case is when both variables are quantitative (numerical). In this situation, the nature of the influence between them is determined by calculating a correlation coefficient and constructing a linear regression model to estimate the values of one variable (referred to as the dependent variable) based on the values of the other variable (referred to as the independent variable). This approach can be generalized to cases where there are multiple quantitative independent variables simultaneously influencing the quantitative dependent variable. The first-degree equation

used in simple linear regression ( $y = ax + b$ ) is extended by adding additional variables, resulting in the form  $y = a_1x_1 + a_2x_2 + a_3x_3 + \dots + a_nx_n + b$ . This yields a new mathematical model known as the multiple linear regression model, which in turn has a wide range of practical applications.

However, the main limitation of these models is that they can only be applied when all variables under analysis are quantitative in nature. Ideally, the dependent variable should be continuous, while the independent variables must be at least ordinal. Additionally, the residuals must be normally distributed and exhibit constant variance across all values of the predictors.

In practice, however, there are many situations in which we aim to analyse the combined influence of several independent variables—whether quantitative or qualitative—on a

dependent variable that is not numerical, but qualitative. For this type of problem, statistical theory proposes an alternative method of investigation: logistic regression, which is based on a more complex mathematical equation but offers greater versatility in terms of practical application. The simplest and most intuitive model of logistic regression, which is also relatively easy to understand and apply, is binary logistic regression. This model investigates the influence of several combined independent variables—preferably qualitative, though also quantitative—on a binary qualitative dependent variable (which has only two possible values, such as Yes/No). Binary logistic regression is used when the goal is to predict the presence or absence of a characteristic (typically a specific disease) based on a set of independent predictor variables [1].

Logistic regression does not directly predict the value (0 or 1) of the dependent variable, but rather the probability that a certain event will occur (for example, the risk of developing a disease). This allows us to better understand which factors influence a binary outcome. Another advantage of this method is its flexibility regarding the assumptions required for data analysis. The only essential conditions are the following: the observations should be independent, the predictors should not be strongly correlated with each other, and the sample size should be sufficiently large to yield stable estimates (a commonly used empirical rule recommends a minimum of 10–15 cases for each value of each predictor). Although in essence binary logistic regression is a relatively simple mathematical model, its practical application is prone to methodological or interpretational errors that can have significant consequences.

The objective of the present paper is to highlight some of the model's subtleties and to propose a structured working procedure that leads to valid and easily interpretable results. The procedure is illustrated using a realistic benchmark medical dataset, generated with the aid of artificial intelligence.

### Theoretical background:

Binary logistic regression is a type of Generalized Linear Model (GLM) that establishes a relationship between a set of independent variables and a single binary qualitative dependent variable, whose values can be predicted based on the values of the independent variables [2,3].

Let have  $n$  independent (predictor) variables,  $X_1, X_2, \dots, X_n$ , which meet the conditions mentioned previously; we aim to determine the probability of occurrence of the outcome, denoted as  $P_i$ , for a given case  $i$  from the total number of cases in the analysed dataset (e.g., a patient). To calculate this probability, the model applies a specific transformation of the data known as the logit transformation. An auxiliary continuous variable, denoted as  $Z$ , is introduced. This variable can be interpreted as the "favourable tendency" for the outcome to occur [4,5]—in such a way that  $Z$  varies directly with the probability  $P_i$ : the higher the value of  $Z$ , the higher the probability that the outcome will occur.

In logistic regression, the relationship between  $Z$  and  $P_i$  is described by the following function:

$$P_i = \frac{e^{Z_i}}{1 + e^{Z_i}} = \frac{1}{1 + e^{-Z_i}} \Rightarrow Z_i = \ln \left( \frac{P_i}{1 - P_i} \right).$$

Since  $Z$  is a continuous variable (by the way it is defined), it can be incorporated into a classical regression model. Thus, the relationship between  $Z$  and the predictor variables will be described according to the general equation of multiple linear regression:

$$Z_i = b_0 + b_1 \cdot X_{i1} + b_2 \cdot X_{i2} + \dots + b_n \cdot X_{in} + \text{err},$$

where  $b_j, j \in \{1, 2, \dots, n\}$  are the coefficients of the predictors  $X_{i1}, X_{i2}, \dots, X_{in}$ , and  $\text{err}$  = the model's error amount.

If the variable  $Z$  were observable, calculating the probability  $P_i$  would only require fitting a regression line to the values of  $Z, X_{i1}, X_{i2}, \dots, X_{in}$ , and nothing more. In practice, however,  $Z$  is not an observable variable; therefore, the

probability  $P_i$  is calculated using the following relationship:

$$P_i = \frac{1}{1 + e^{-(b_0 + b_1 \cdot X_{i1} + b_2 \cdot X_{i2} + \dots + b_n \cdot X_{in} + err)}}$$

known in the literature as the logistic regression equation.

The simplest binary logistic regression models use binary variables as predictors, typically selected from risk factors identified through univariate analysis. However, the independent variables included in the model can just as well be multinomial or quantitative. In order to distinguish between quantitative and qualitative variables included in the model, the latter must be explicitly defined as categorical. Professional statistical analysis software provides users with several methods for selecting the predictors to be included in the logistic regression model to be built. Consequently, an important aspect of the analysis is choosing an appropriate selection method. Three main approaches are available: Enter, Forward Stepwise, and Backward Stepwise. The Enter method is the most commonly used. Its main characteristic is that all predictors specified by the researcher are introduced into the model simultaneously. This method has the advantage of being entirely controlled by the researcher and is especially used when there are clear theoretical hypotheses, i.e., when the predictors are supported by theory or by previous literature. Therefore, the Enter method is particularly recommended in clinical and epidemiological research. However, its main disadvantages are that it does not eliminate irrelevant predictors and does not address multicollinearity between them, which reduces the model's predictive power. The Forward Stepwise and Backward Stepwise methods are somewhat similar in their logic, as they involve selecting predictors based on their statistical contribution to the model, keeping only those that are significant. In the Forward Stepwise method, the model is initiated without any predictors, and variables are added one at a time based on their significant contribution to the model, determined using criteria such as the Wald

statistic, likelihood ratio, or AIC. This approach builds a more concise model that includes only significant predictors and establishes a hierarchy among them, helping to identify the most important of them, too. However, a limitation of this method is that it may ignore interactions between predictors that are jointly significant but not individually. In contrast, the Backward Stepwise method starts with all predictors in the model and removes those that do not contribute significantly, also using criteria such as the Wald statistic, likelihood ratio, or AIC. This method is useful for simplifying the model without compromising its predictive power, although it may eliminate important predictors in cases of multicollinearity. These last two methods are particularly useful when dealing with a large number of potential predictors, with no clear indication of which are most important. They are therefore recommended in predictive modeling contexts, such as Machine Learning or Data Mining.

Once the predictors have been classified according to their nature and the appropriate selection method has been chosen, the analysis can begin.

Before interpreting the results, it is essential to verify whether the predictors included are useful and whether the model fits the data well. To this end, two tests are required:

- Omnibus Test of Model Coefficients, which indicates the overall statistical significance of the model—that is, how well the model predicts the categories of the dependent variable compared to a null model without predictors. The model is considered acceptable when the test is statistically significant,  $p < 0.05$ ;
- Hosmer and Lemeshow test, which evaluates how poorly the model predicts the categories of the dependent variable. In this case, the model is considered acceptable if the test is not statistically significant,  $p > 0.05$ .

In the next stage, model validation, we assess how much of the variation in the dependent

variable is explained by the model. For this purpose, we examine the Nagelkerke R Square coefficient, which is expressed as a percentage and reflects the proportion of explained variance—this value should be as high as possible (for example, if  $R^2 = 0.330$ , it means the model explains 33% of the variation in the dependent variable).

Another important indicator of model quality is its classification accuracy, which is identified using the generated classification tables. The global percentage of correct classifications is monitored both before and after applying the model. This percentage should improve after applying the model, thereby justifying the usefulness of the included predictors. Additional performance measures extracted from the final classification table include sensitivity (the percentage of truly positive cases correctly predicted by the model - i.e., subjects for whom the dependent variable equals 1) and specificity (the percentage of truly negative cases correctly predicted - i.e., subjects for whom the dependent variable equals 0).

For the practical interpretation of the model results, the contribution and statistical significance of each predictor are analyzed, retaining only those that contribute significantly to the model. For each of these, the odds ratio (OR) is examined along with its 95% confidence interval, indicating how the odds of the dependent variable being 1 change with each one-unit increase in the predictor's value. Subsequently, the combined effect of the predictors can be specified by calculating the probability of the dependent variable being equal to 1 for each possible combination of predictor values, based on the constructed logistic regression equation. The overall

performance of the model is illustrated using the ROC curve, which is based on the predicted probabilities of the dependent variable. The area under the curve (AUC) is then evaluated; an AUC of 0.5 indicates a null model (predictions are random);  $AUC > 0.7$  suggests acceptable predictive accuracy,  $AUC > 0.8$  indicates good predictive accuracy, and  $AUC > 0.9$  reflects excellent predictive power.

### Method:

Statistical analysis was carried out in SPSS 29.0.

### Results:

#### The Database:

We illustrate the previously described analytical procedure using a realistic dataset generated with the help of AI, focusing on risk factors for endometrial cancer as reported in the scientific literature [6-9]. To obtain realistic data, a request was formulated for the generation of a table with 400 records (including both positive and negative cases), containing eight risk factors that are frequently reported in the literature for endometrial cancer—without specifying individual case details. The AI application returned a dataset consisting of 148 women diagnosed with endometrial cancer (37.0%) and 252 without cancer (63.0%). The selected risk factors, based on their prevalence in the literature, were: Age (in years), BMI, Menarche Age, Parity, Diabetes, Hypertension, Estrogen Therapy, and Family History. The values of these predictors were not generated randomly, but were modelled according to biologically plausible distributions, also derived from published data (table 1).

Table 1. The database's structure: the selected risk factors and the criteria used to generate values (AI)

| Variable     | Values Distribution           | Motivation                                                        |
|--------------|-------------------------------|-------------------------------------------------------------------|
| Age          | N(60, 10)                     | Most cases of endometrial cancer appear after the age of 50 years |
| BMI          | N(30, 5)                      | Obesity is a major risk factor                                    |
| Menarche age | N(13, 1.5)                    | Early menarche is associated with an increased risk               |
| Parity       | 0–3, not uniform distribution | Nulliparity is a recognized risk factor                           |

| Variable         | Values Distribution | Motivation                                               |
|------------------|---------------------|----------------------------------------------------------|
| Diabetes         | Bernoulli(p=0.25)   | Diabetes is associated with a higher risk                |
| Hypertension     | Bernoulli(p=0.35)   | Hypertension is frequently linked to metabolic syndrome  |
| Estrogen therapy | Bernoulli(p=0.15)   | Estrogen therapy without progesterone increases the risk |
| Family history   | Bernoulli(p=0.10)   | Family history is a well-established risk factor         |

### The Analysis:

Using the resulting dataset, we will apply binary logistic regression to analyze the combined effect of the eight risk factors in predicting the likelihood of endometrial cancer.

The standard approach for initiating this type of analysis is through univariate analysis (using Pearson Chi-squared, Student's *t*-test, or the Mann–Whitney U test, as appropriate), which aims to identify independent risk factors for endometrial cancer and their corresponding effects – i.e. odds ratios (ORs). In this case, however, such analysis is redundant, since all eight predictors are already established risk factors based on the way the dataset was generated. Nevertheless, the results obtained are presented in table 2.

Table 2. Univariate analysis of the selected risk factors

|                  |           | Cancer        |        |               |        | Total         |        | t-Student /<br>Pearson Chi-squared | OR<br>(95% CI)            |
|------------------|-----------|---------------|--------|---------------|--------|---------------|--------|------------------------------------|---------------------------|
|                  |           | present       |        | absent        |        |               |        |                                    |                           |
|                  |           | N             | %      | N             | %      | N             | %      |                                    |                           |
| Age              | M ± SD    | 62.22 ± 9.914 |        | 59.10 ± 9.269 |        | 60.26 ± 9.619 |        | t = -3.160<br>p = 0.002**          |                           |
|                  | min ÷ max | 40 ÷ 87       |        | 28 ÷ 99       |        | 28 ÷ 99       |        |                                    |                           |
|                  | median    | 62.00         |        | 59.00         |        | 61.00         |        |                                    |                           |
|                  | IQR       | 54.00 ÷ 69.00 |        | 53.00 ÷ 65.00 |        | 53.00 ÷ 66.00 |        |                                    |                           |
| BMI              | M ± SD    | 31.10 ± 4.703 |        | 29.05 ± 5.080 |        | 29.81 ± 5.035 |        | t = -3.990<br>p <0.001**           |                           |
|                  | min ÷ max | 18.90 ÷ 45.40 |        | 16.50 ÷ 43.20 |        | 16.50 ÷ 45.40 |        |                                    |                           |
|                  | median    | 30.75         |        | 28.95         |        | 29.80         |        |                                    |                           |
|                  | IQR       | 27.95 ÷ 34.20 |        | 25.43 ÷ 32.48 |        | 26.23 ÷ 33.10 |        |                                    |                           |
| Menarche age     | M ± SD    | 12.94 ± 1.454 |        | 13.35 ± 1.494 |        | 13.20 ± 1.490 |        | t = 2.636<br>p = 0.009**           |                           |
|                  | min ÷ max | 8.7 ÷ 16.1    |        | 8.70 ÷ 16.80  |        | 8.70 ÷ 16.80  |        |                                    |                           |
|                  | median    | 13.10         |        | 13.30         |        | 13.20         |        |                                    |                           |
|                  | IQR       | 12.00 ÷ 13.90 |        | 12.30 ÷ 14.40 |        | 12.20 ÷ 14.20 |        |                                    |                           |
| Parity           | 0         | 38            | 25.7%  | 45            | 17.9%  | 83            | 20.8%  | Chi² = 6.219<br>p = 0.101          | -                         |
|                  | 1         | 35            | 23.6%  | 74            | 29.4%  | 109           | 27.3%  |                                    |                           |
|                  | 2         | 53            | 35.8%  | 80            | 31.7%  | 133           | 33.3%  |                                    |                           |
|                  | 3         | 22            | 14.9%  | 53            | 21.0%  | 75            | 18.8%  |                                    |                           |
| Parity 0         | yes       | 38            | 25.7%  | 45            | 17.9%  | 83            | 20.8%  | Chi² = 3.466<br>p = 0.063+         | 1.589<br>(0.974 ÷ 2.593)  |
|                  | no        | 110           | 74.3%  | 207           | 82.1%  | 317           | 79.3%  |                                    |                           |
| Diabetes         | yes       | 51            | 34.5%  | 48            | 19.0%  | 99            | 24.8%  | Chi² = 11.891<br>p <0.001**        | 2.235<br>(1.407 ÷ 3.548)  |
|                  | no        | 97            | 65.5%  | 204           | 81.0%  | 301           | 75.3%  |                                    |                           |
| Hypertension     | yes       | 63            | 42.6%  | 72            | 28.6%  | 135           | 33.8%  | Chi² = 8.169<br>p = 0.004**        | 1.853<br>(1.211 ÷ 2.835)  |
|                  | no        | 85            | 57.4%  | 180           | 71.4%  | 265           | 66.3%  |                                    |                           |
| Estrogen therapy | yes       | 34            | 23.0%  | 29            | 11.5%  | 63            | 15.8%  | Chi² = 9.236<br>p = 0.002**        | 2.293<br>(1.331 ÷ 3.953)  |
|                  | no        | 114           | 77.0%  | 223           | 88.5%  | 337           | 84.3%  |                                    |                           |
| Family history   | yes       | 21            | 14.2%  | 9             | 3.6%   | 30            | 7.5%   | Chi² = 15.152<br>p <0.001**        | 4.465<br>(1.986 ÷ 10.034) |
|                  | no        | 127           | 85.8%  | 243           | 96.4%  | 370           | 92.5%  |                                    |                           |
| Total            |           | 148           | 100.0% | 252           | 100.0% | 400           | 100.0% |                                    |                           |

The previously formulated hypothesis is thus confirmed: univariate analysis reveals

statistically significant differences in the values of all eight selected risk factors between

patients with endometrial cancer and those without, with one exception—parity. The values for this variable were generated under the assumption that nulliparity is a risk factor; although the proportion of women with no births is higher among patients with endometrial cancer, the observed differences do not reach the threshold for statistical significance. In the case of multinomial variables acting as predictors for a dependent variable, in order to evaluate the corresponding risk (odds ratio), binarization of the variable can be performed. We applied this approach to the variable Parity, reducing the original four categories to just two: presence vs. absence of nulliparity, and repeated the univariate analysis. This time, the observed difference between women with endometrial cancer and those without was close to the statistical significance threshold ( $p = 0.063$ ), yet still did not surpass it. In situations where the statistical significance threshold is not met, even though an odds ratio (OR) can technically be calculated, it should not be considered valid. This is because the associated 95% confidence

interval includes the reference value of one, which invalidates the risk estimate.

For the three quantitative predictors included in the study (age, BMI, and menarche age), statistically significant differences were also observed between the values recorded for patients with endometrial cancer and those without. These predictors can be included in the multivariate analysis as continuous variables, but they can also be binarized in order to obtain a more precise estimation of their odds ratios, especially in relation to the other predictors included in the model. The simplest technique for binarizing quantitative variables is by performing ROC curve analysis and identifying the cut-off value that best differentiates between the categories of the dependent variable. The results obtained are shown in table 3. The identification of the cut-off value is done by visually examining the ROC curve (figures 1–3) and selecting the point closest to the top-left corner of the plot. A more objective numerical selection criterion was proposed by Youden, and it is easier to apply in practice: the cut-off value is the one that maximizes the sum of sensitivity and specificity.

Table 3. ROC analysis of the quantitative predictors

| Variable     | AUC coefficient | 95% CI        | p-value | Cut-off value | Sensitivity | Specificity |
|--------------|-----------------|---------------|---------|---------------|-------------|-------------|
| Age          | 0.588           | 0.529 ÷ 0.646 | 0.003** | 67.50         | 0.331       | 0.845       |
| BMI          | 0.617           | 0.561 ÷ 0.673 | 0.000** | 27.15         | 0.838       | 0.389       |
| Menarche age | 0.567           | 0.510 ÷ 0.625 | 0.024*  | 13.55         | 0.662       | 0.440       |

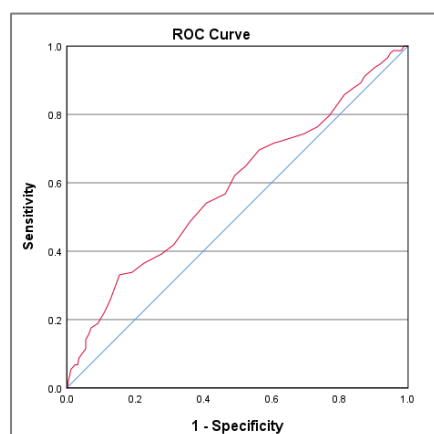


Figure 1. ROC curve: Age (years)

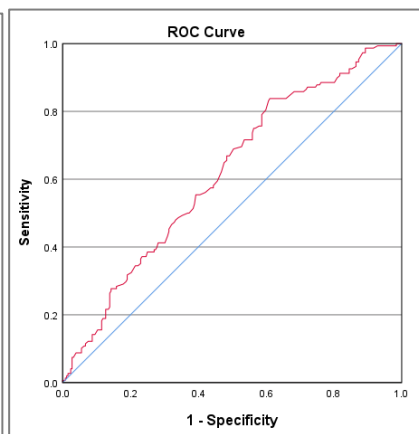


Figure 2. ROC curve: BMI

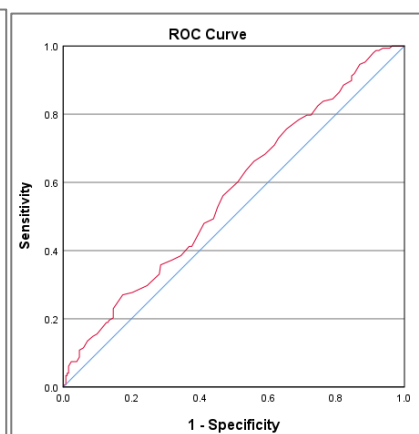


Figure 3. ROC curve: Menarche age (years)

Although the predictive accuracy of the identified cut-off values is modest in this case (a fact that is not particularly relevant, given that the dataset used for the study is hypothetical), these cut-off values were nonetheless suitable for the binarization of the three quantitative predictors. Subsequently, we performed univariate analysis for the binarized variables and calculated the associated odds ratios, as presented in table 4.

Table 4. Univariate analysis of the binarized quantitative predictors

|                            |     | Cancer  |        |        |        | Total |        | t-Student /<br>Pearson Chi-<br>squared | OR<br>(95% CI)  |
|----------------------------|-----|---------|--------|--------|--------|-------|--------|----------------------------------------|-----------------|
|                            |     | present |        | absent |        |       |        |                                        |                 |
|                            |     | N       | %      | N      | %      | N     | %      |                                        |                 |
| Age > 67.50 years          | yes | 49      | 33.1%  | 39     | 15.5%  | 88    | 22.0%  | Chi² = 16.892                          | 2.703           |
|                            | no  | 99      | 66.9%  | 213    | 84.5%  | 312   | 78.0%  | p <0.001**                             | (1.667 ÷ 4.384) |
| BMI > 27.15                | yes | 124     | 83.8%  | 154    | 61.1%  | 278   | 69.5%  | Chi² = 22.611                          | 3.288           |
|                            | no  | 24      | 16.2%  | 98     | 38.9%  | 122   | 30.5%  | p <0.001**                             | (1.984 ÷ 5.449) |
| Menarche age < 13.55 years | yes | 98      | 66.2%  | 141    | 56.0%  | 239   | 59.8%  | Chi² = 4.084                           | 1.543           |
|                            | no  | 50      | 33.8%  | 111    | 44.0%  | 161   | 40.3%  | p = 0.043*                             | (1.012 ÷ 2.352) |
| Total                      |     | 148     | 100.0% | 252    | 100.0% | 400   | 100.0% |                                        |                 |

The next stage involves multivariate analysis, which can be approached in two ways depending on how the investigated predictors are included in the model.

The first approach involves including the predictors in their original form (quantitative or qualitative, multinomial or binary variables). We applied this method, incorporating all eight predictors into the model, including Parity, which was not identified as a significant risk factor for endometrial cancer in the univariate analysis. The resulting model, built using the Enter selection method, demonstrates

acceptable predictive capacity for endometrial cancer (Omnibus Tests of Model Coefficients,  $p < 0.001$ , and Hosmer and Lemeshow Test,  $p = 0.552$ ), and explains 28.6% of the variance in the occurrence of the disease (Nagelkerke R-Squared = 0.286). The percentage of correctly classified cases improves from 63.0% (initial) to 74.8% after applying the model, with a sensitivity of 53.4% (correct classification of true positive cases) and a specificity of 87.3% (correct classification of true negative cases). All eight predictors included in the model are statistically significant (table 5).

Table 5. Multivariate analysis of the investigated predictors

|                     | B      | S.E.  | Wald   | df | Sig.     | Exp(B) | 95% C.I. for EXP(B) |        |
|---------------------|--------|-------|--------|----|----------|--------|---------------------|--------|
|                     |        |       |        |    |          |        | Lower               | Upper  |
| Age                 | 0.050  | 0.013 | 15.381 | 1  | <0.001** | 1.052  | 1.026               | 1.079  |
| BMI                 | 0.105  | 0.025 | 18.001 | 1  | <0.001** | 1.111  | 1.058               | 1.167  |
| Menarche age        | -0.225 | 0.082 | 7.580  | 1  | 0.006**  | 0.798  | 0.680               | 0.937  |
| Parity              |        |       | 10.782 | 3  | 0.013*   |        |                     |        |
| Parity(1)           | -0.984 | 0.347 | 8.058  | 1  | 0.005**  | 0.374  | 0.189               | 0.737  |
| Parity(2)           | -0.538 | 0.324 | 2.752  | 1  | 0.097    | 0.584  | 0.309               | 1.103  |
| Parity(3)           | -1.068 | 0.387 | 7.630  | 1  | 0.006**  | 0.344  | 0.161               | 0.733  |
| Diabetes(1)         | 1.039  | 0.268 | 15.037 | 1  | <0.001** | 2.827  | 1.672               | 4.779  |
| Hypertension(1)     | 0.802  | 0.251 | 10.174 | 1  | 0.001**  | 2.230  | 1.362               | 3.650  |
| Estrogen therapy(1) | 0.921  | 0.311 | 8.781  | 1  | 0.003**  | 2.511  | 1.366               | 4.618  |
| Family history(1)   | 1.743  | 0.454 | 14.731 | 1  | <0.001** | 5.717  | 2.347               | 13.927 |
| Constant            | -4.027 | 1.584 | 6.464  | 1  | 0.011    | 0.018  |                     |        |

The interpretation of these predictors is carried out as follows:

- Age: OR = 1.052 – for every one-year increase in age, the risk of endometrial cancer increases by 5.2%;
- BMI: OR = 1.111 – each additional unit of BMI increases the risk by 11.1%;
- Menarche age: OR = 0.798 – for every additional year in age at menarche, the risk of endometrial cancer decreases by 20.2%;
- Parity (1) vs. reference: OR = 0.374 – women with one child have a 62.6% lower risk of endometrial cancer compared to nulliparous women;
- Parity (2) vs. reference: marginally non-significant predictor;
- Parity (3) vs. reference: OR = 0.344 – women with three children have a 65.6% lower risk of endometrial cancer compared to nulliparous women;
- Diabetes (1): OR = 2.827 – women with diabetes have a 2.827-fold higher risk of endometrial cancer than those without;
- Hypertension (1): OR = 2.230 – women with hypertension have a 2.230-fold higher risk than normotensive women;
- Estrogen therapy (1): OR = 2.511 – women who received estrogen therapy have a 2.511-fold higher risk of endometrial cancer;
- Family history (1): OR = 5.717 – women with a family history of cancer have a 5.717-fold higher risk of developing endometrial cancer.

The role of the investigated predictors in assessing the risk of endometrial cancer can be

visually represented using a Forest Plot, which displays the odds ratios along with their corresponding 95% confidence intervals (fig. 4).

The immediate practical utility of this model lies in the fact that, using the logistic regression equation, it is possible to calculate the probability that a patient will develop endometrial cancer, based on the evaluation of the eight predictors used in the model. The equation identified by the model is:

$$\text{logit}(P) = -4.027 + 0.050 \cdot \text{Age} + 0.105 \cdot \text{BMI} - 0.225 \cdot \text{Menarche age} - 0.984 \cdot \text{Parity}(1) - 0.538 \cdot \text{Parity}(2) - 1.068 \cdot \text{Parity}(3) + 1.039 \cdot \text{Diabetes} + 0.802 \cdot \text{Hypertension} + 0.921 \cdot \text{Estrogen therapy} + 1.743 \cdot \text{Family history}$$

Therefore, for a 62-year-old woman with a BMI of 33, menarche at age 11, who has had three children, has diabetes, hypertension, and a family history of cancer, but has not undergone estrogen therapy, the estimated probability of having endometrial cancer will be:

$$\text{logit}(P) = -4.027 + 0.050 \cdot 62 + 0.105 \cdot 33 - 0.225 \cdot 11 - 1.068 \cdot 1 + 1.039 \cdot 1 + 0.802 \cdot 1 + 0.921 \cdot 0 + 1.743 \cdot 1 = 2.579$$

$$\Rightarrow P = 1/(1+e^{-2.579}) = 0.929$$

Therefore, this probability is 92.9%.

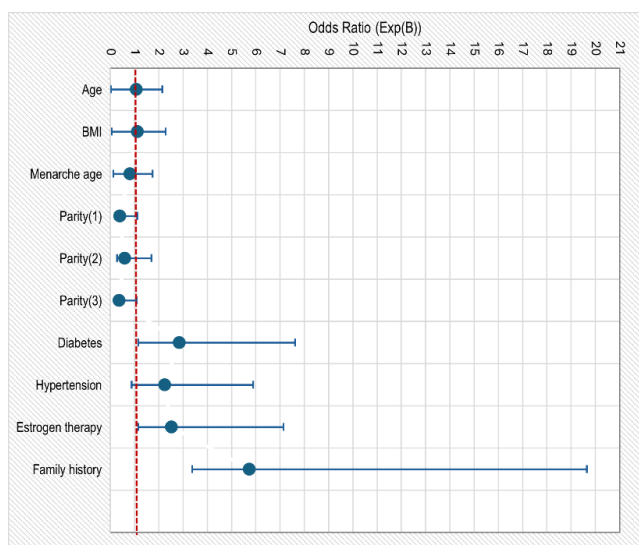


Figure 4. Forest Plot – Predictors of the Endometrial Cancer

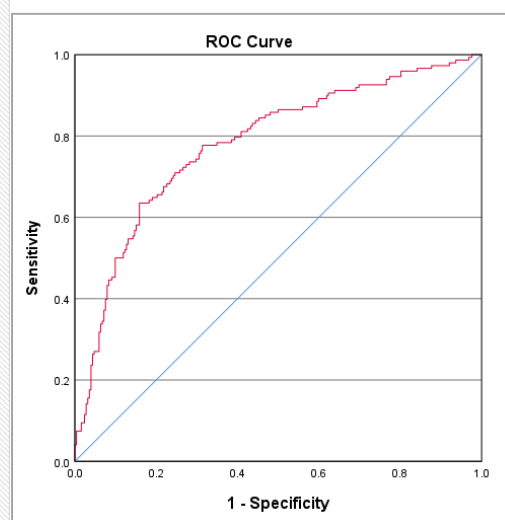


Figure 5. ROC curve of the generated model

The final evaluation of the quality of the constructed model is carried out by plotting the ROC curve for the predicted probabilities (figure 5). The corresponding AUC coefficient has a value of 0.781, which is statistically significant, indicating an acceptable to good performance of the model in predicting endometrial cancer.

A second approach involves including in the model only the statistically significant binary predictors (obtained by binarizing the quantitative and multinomial qualitative variables). In this case, this means that the Parity predictor will no longer be included in the model, as it does not contribute significantly to identifying endometrial cancer.

The resulting model, also constructed using the Enter selection method, shows acceptable predictive performance for endometrial cancer (Omnibus Tests of Model Coefficients,  $p < 0.001$ ; Hosmer and Lemeshow Test,  $p = 0.780$ ), and explains 27.1% of the variance in the occurrence of the disease (Nagelkerke R-Squared = 0.271). The percentage of correctly classified cases improves from 63.0% initially to 71.5% after applying the model, with a sensitivity of 50.0% (correct classification of true positives) and a specificity of 84.1% (correct classification of true negatives). Only 6 out of the 7 predictors included in the model are statistically significant (table 6).

Table 6. Multivariate analysis of the investigated predictors

|                    | B      | S.E.  | Wald   | df | Sig.     | Exp(B) | 95% C.I. for EXP(B) |        |
|--------------------|--------|-------|--------|----|----------|--------|---------------------|--------|
|                    |        |       |        |    |          |        | Lower               | Upper  |
| Age>67.5           | 1.148  | 0.277 | 17.189 | 1  | <0.001** | 3.150  | 1.831               | 5.420  |
| BMI>27.15          | 1.274  | 0.286 | 19.808 | 1  | <0.001** | 3.574  | 2.040               | 6.262  |
| Menarche age<13.55 | 0.450  | 0.243 | 3.429  | 1  | 0.064    | 1.568  | 0.974               | 2.524  |
| Diabetes           | 0.908  | 0.265 | 11.775 | 1  | <0.001** | 2.478  | 1.476               | 4.162  |
| Hypertension       | 0.693  | 0.244 | 8.090  | 1  | 0.004**  | 2.000  | 1.241               | 3.226  |
| Estrogen therapy   | 0.931  | 0.308 | 9.140  | 1  | 0.003**  | 2.537  | 1.387               | 4.640  |
| Family history     | 1.653  | 0.458 | 13.007 | 1  | <0.001** | 5.225  | 2.127               | 12.833 |
| Constant           | -2.792 | 0.347 | 64.571 | 1  | <0.001   | 0.061  |                     |        |

The model shows a slight decrease in overall quality compared to the previous approach, but

the interpretation of the investigated predictors is simpler. Specifically:

- Age over 67.5 years increases the risk of endometrial cancer by 3.150 times;
- BMI greater than 27.15 increases the risk by 3.574 times;
- Menarche age under 13.55 years does not significantly influence the risk of endometrial cancer;
- The presence of diabetes increases the risk by 2.478 times;
- The presence of hypertension increases the risk by 2.000 times;
- Estrogen therapy increases the risk by 2.537 times;

- A family history of cancer increases the risk by 5.225 times.

A graph is also generated to illustrate the odds ratios (ORs) and the confidence intervals of the predictors included in the model. It can be observed that, in this case, all predictors have odds ratios greater than one, with more pronounced risk estimates than in the previous model (fig. 6). The ROC curve corresponding to the probability predicted by the model (fig. 7) is qualitatively similar to the previous one, with a comparable AUC coefficient of 0.768.

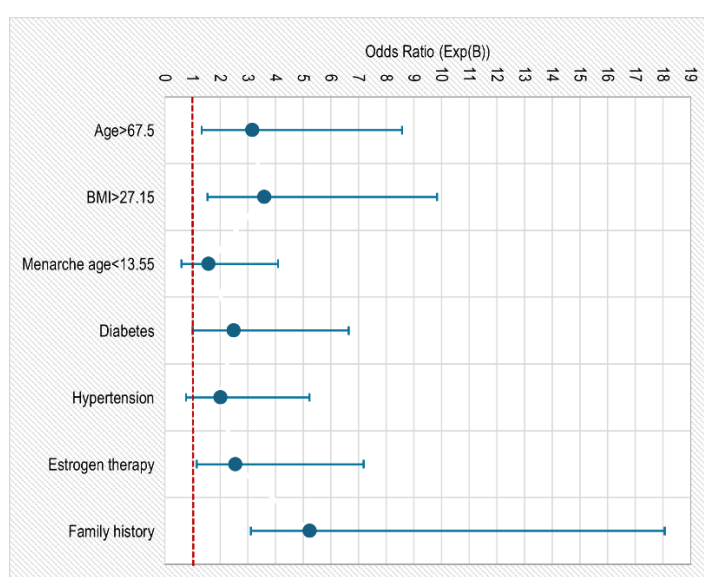


Figure 6. Forest Plot – Predictors of the Endometrial Cancer

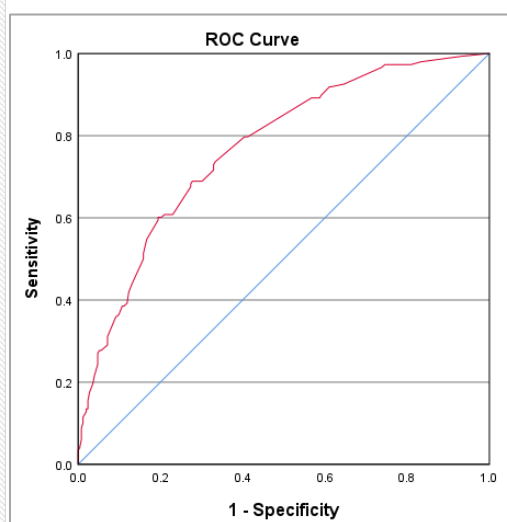


Figure 7. ROC curve of the generated model

The model, however, can be more easily used to calculate the probabilities of endometrial cancer based on the logistic regression equation, as the included predictors are more intuitive to interpret and apply. The equation identified by the model is:

$$\text{logit}(P) = -2.792 + 1.148 * (\text{Age} > 67.5) + 1.274 * (\text{BMI} > 27.15) + 0.450 * (\text{Menarche age} < 13.55) + 0.908 * \text{Diabetes} + 0.693 * \text{Hypertension} + 0.931 * \text{Estrogen therapy} + 1.653 * \text{Family history}$$

Therefore, for the same 62-year-old woman with a BMI of 33, menarche at age 11, who has had three children, has diabetes, hypertension, and a family history of cancer, but has not

undergone estrogen therapy, the estimated probability of having endometrial cancer will be:

$$\text{logit}(P) = -2.792 + 1.148 * 0 + 1.274 * 1 + 0.450 * 1 + 0.908 * 1 + 0.693 * 1 + 0.931 * 0 + 1.653 * 1 = 2.186$$

$$\Rightarrow P = 1 / (1 + e^{-2.186}) = 0.899$$

Therefore, this probability is 89.9%. It can be observed that binarizing the variables and eliminating non-significant predictors altered the prediction very slightly—by only 3 percentage points, which is below the 5% margin of error—a difference that is perfectly acceptable.

## Discussions:

Logistic regression represents an alternative to the classical method of discriminant analysis, which is generally used when working with multiple qualitative dependent variables. In the case of a single dependent variable, logistic regression is better suited to the nature of the data than discriminant analysis, making it the preferred solution. Other notable advantages of this method, which make it preferable in many situations, include the following:

1. After fitting the regression model, the statistical significance of each predictor variable included in the model is assessed—by calculating the Wald statistics and comparing the estimated coefficients to those from a univariate model that contains only that predictor. Variables that do not contribute significantly to the model according to this criterion can thus be identified and excluded, allowing a new fitted model to be generated [10].
2. The model accepts predictors that are not necessarily qualitative; moreover, the dependent variable does not need to be normally distributed, as the model is designed to handle nonparametric distributions—a situation frequently encountered in medical studies, making logistic regression a highly useful investigative tool in such contexts [11].
3. Relative risks are calculated, which are useful for identifying potential risk factors among the predictor variables, as well as the estimated probabilities of disease for each individual case in the dataset—information that is extremely valuable in clinical research studies.

The interpretation of binary or multinomial logistic regression results must, however, be approached with caution, as the predictive model is limited to the variables included—while other relevant predictors may have been overlooked, either intentionally or unintentionally. Another important limitation of the model concerns the sample size, which

must be sufficiently large to ensure reliable results. A minimum requirement is to have at least 100 units in the sample; other studies recommend at least 50 cases per predictor, or 20–30 cases per predictor for samples exceeding 300 units. The commonly recommended sample size, for a significance level of 0.05, a power of 0.80, and a medium effect size, is around 300 units per sample; however, if greater statistical power is desired, the required sample size can easily increase to 1,000 units or more.

It is equally important to control the number of predictors included in the model. When the independent variables are qualitative, a common rule of thumb is to use the smallest group size as a reference and divide it by 10. For example, in a sample with 100 men and 50 women, the reference group would be the women, and based on this rule, no more than 5 predictors could be included in the model.

As with all statistical techniques used for data analysis, it must be remembered that statistical significance does not necessarily imply clinical significance—a point frequently emphasized in clinical studies [12–15]. Thus, *p*-values below the 0.05 threshold, while indicating a statistically significant association, must be interpreted in conjunction with the effect size (evaluated, in the case of qualitative variables, by the OR) and within the clinical context, since a statistically significant effect may still be clinically irrelevant.

Another crucial point—which applies to most statistical techniques—is that logistic regression identifies associations, not causal relationships [16–17]. This has direct implications for how results should be interpreted, which must always be related to the biological, pathophysiological, and clinical context.

## Conclusions:

Binary logistic regression is an analytical tool highly useful in medical research, currently used nowadays in clinical studies due to its ability to estimate the occurrence probability for binary events in the presence of potential

predictors. The method is useful, flexible, and adaptable, because it is able to provide a rigorous framework for informed clinical decision-making, personalized treatment strategies, or optimization of screening programs. One of its many notable advantages is its capacity to compare risks and to assist the medical researchers in creating predictive scores highly applicable in clinical settings,

based on the combined action of carefully selected predictors.

If it is properly applied, binary logistic regression has the capacity to provide significant benefits, even if its usefulness depends heavily on the accurate interpretation of the results, sustained by a clear understanding of the method's limitations and its clinical implications.

## References

1. Munro B.H. *Statistical Methods for Health Care Research*. Philadelphia, New York: Lippincott Press, 1997.
2. Olowe KJ, Edoh NL, Zouo SJC, Olamijuwon J. Comprehensive review of logistic regression techniques in predicting health outcomes and trends. *World Journal of Advanced Pharmaceutical and Life Sciences* 2024, 07(02), 016–026. doi: 10.53346/wjapls.2024.7.2.0039.
3. Harris JK. Primer on binary logistic regression. *Fam Med Com Health* 2021;9:e001290. doi:10.1136/fmch-2021-001290.
4. Hosmer D.W., Lemeshow S. *Applied Logistic Regression*. New York: John Wiley & Sons, 2000.
5. Kleinbaum D.G. *Logistic Regression: A Self-Learning Text*. New York: Springer-Verlag, 1994.
6. Li X, Cheng Y, Dong Y et al. Development and validation of predictive model for lymph node metastasis in endometrial cancer: a SEER analysis. *Ann Transl Med*. 2021;9(7):538. doi: 10.21037/atm-20-5034.
7. Zhang L, Lai H, Lin Q et al. Analysis of risk factors and construction and validation of a predictive model for determining endometrial cancer risk in postmenopausal patients with abnormal bleeding. *Eur J Gynaecol Oncol*. 2023;44(5):39–45. doi: 10.22514/ejgo.2023.077.
8. Forder BH, Ardasheva A, Atha K, Nentwich H, Abhari R, Kartsonaki C. Models for predicting risk of endometrial cancer: a systematic review. *Diagnostic and Prognostic Research* 2025; 9:3. doi: 10.1186/s41512-024-00178-0.
9. Koskas M, Luton D, Graesslin O, Barranger E, Clavel-Chapelon F, Haddad B, Darai E, Rouzier R. Direct Comparison of Logistic Regression and Recursive Partitioning to Predict Lymph Node Metastasis in Endometrial Cancer. *International Journal of Gynecological Cancer* 2015; 25(6): 1037-1043. doi: 10.1097/IGC.0000000000000451.
10. Vittinghoff E, Glidden DV, Shiboski SC, McCulloch CE. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. Springer; 2022.
11. Katz MH. *Multivariable Analysis: A Practical Guide for Clinicians and Public Health Researchers*. 4th ed. Cambridge University Press; 2023.
12. Sperandei S. Understanding logistic regression analysis. *Biochem Med (Zagreb)*. 2014; 24(1):12-18. doi:10.11613/BM.2014.003.
13. Ghioldiş AC, Sârbu V, Pundiche M, Dan C, Butelchin C, Olteanu C, Popescu RC. Clostridium Difficile Infection in Rectal Cancer Patients after Diverted Loop Ileostomy Closure. *Chirurgia*, 2024, 119, 1, 36–43, doi: 10.21614/chirurgia.2024.v.119.i.1.p.36.
14. Vamesu S, Ursica OA, Milea SE, Deacu M, Aschie M, Mitroi AF, Voinea F, Pundiche MB, Orasanu CI, Voda RI. Same Organ, Two Cancers: Complete Analysis of Renal Cell Carcinomas and Upper Tract Urothelial Carcinomas. *Medicina*, 2024, 60, 7, 1126, doi: 10.3390/medicina60071126.

15. Dumitru A, Matei E, Cozaru GC, Chisoi A, Alexandrescu L, Popescu RC, Butcaru MP, Dumitru E, Rugină S, Tocio C. Endotoxin Inflammatory Action on Cells by Dysregulated-Immunological-Barrier-Linked ROS-Apoptosis Mechanisms in Gut-Liver Axis. *Int J Mol Sci*, 2024, 25, 5, 2472, doi: 10.3390/ijms25052472.
16. Boiculese LV, Dimitriu G, Moscalu M. Nearest neighbor classification with improved weighted dissimilarity measure. *Proc. Rom. Acad. Ser. A*, 2009, 10, 205-213.
17. Constantinescu V, Arsenescu-Georgescu C, Matei D, Moscalu M, Corciova C, Cuciureanu D. Heart rate variability analysis and cardiac dysautonomia in ischemic stroke patients. *Clin Neurol Neurosurg*. 2019; 186:105528. doi: 10.1016/j.clineuro.2019.105528.