

ICMR* in patients under 18 years old

Author: MD Hung Dung Van **Start date :** July 17, 2020 **Last Update:** August 16, 2020

*ICMR - Isolated Congenital Mitral Regurgitation

Design Study Objectives

Population: N = 119 patients under 18 years old with heart diseases. These patients either a classic Carpentier-Edwards (CE) ring implant, a Band implant or neither of those implants.

Research questions:

- In the long term, there is a higher chance of patient having to re-operate due to mitral stenosis (MS) occurring post < **26mm (small size) CE ring** Implant.
- In the long term, there is a higher chance of patient having to re-operate due to mitral regurgitation (MR) occurring post **Band** Implant.

Aim:

- to determine whether the above are true by comparing the MS rate and the MR rate between the Band and CE ring groups in the population.
- to calculate the relapse (MS or MR) probability grouped by age and sex.

Methods used :

- Multinomial Logistic regression
- Survival analysis

Descriptive Statistics

Flow chart of the study

About the data

The data set contains 44 explanatory features (excluding the outcome variable) on 119 patients. The outcome/response variable, **Cause_of_redo** is a categorical variable with 3 levels:

- MS
- MR
- NONE

Not all explanatory variables were used in the analysis. In this briefing, the predictor variables used in the multinomial regression analysis are:

1. **Age_group**, four-level categorical variable describing the age groups:
 - 0-4 yrs
 - 5-9 yrs
 - 10-14 yrs
 - 16-18 yrs
2. **GROUP**, three-level categorical variable describing the types of patient with or without implants. More specifically, patients with Band, patients with CE ring and patients with neither:
 - NONE
 - BAND
 - RING
3. **Ring_group**, three-level categorical variable describing the CE ring sub groups, patients with small size ring (<26mm), patients with large size ring (>26mm) and patients without CE ring:
 - NONE
 - <26mm
 - ≥ 26 mm

Data Visualisation (hover over the graphs for more information)

The pie chart describes the distribution between patients with Band implants (28.6%), patients with Ring implants (65.5%) and those with neither implants (5.9%), where the number of cases for each group are 34, 78 and 7, respectively. In the long run, the total redo cases is 24 cases for N = 119 patients, where the redo rates for both Band and CE ring implant are 20.5% each and the redo rate for those with no implants is 0%. By eyeballing, the stacked bar graph on the right shows that the frequency of the redo cases in **0-4 yrs** age group is the highest and that there are no redo cases found in the **10-14 yrs** and **15-18 yrs** age groups. This visualisation can also be used to calculate the redo rate caused by the categories of the outcome variable **Cause_of_redo** for each Age group, where each bar represents the total number of cases for each Age group and each color of the bars represents the categories of the bars.

Figure 1: Descriptive statistics of the data

Moving on to the Age and Gender tornado chart, it is evident that the most frequent age group in the population is the **0-4 yrs** age group, with 18 males and 27 females out of the total of 53 males and 66 females. It is also possible to calculate the redo rate for Female and Male, by dividing the number of redo cases by the total number of cases for each Gender. In summary, the redo rate for Female (22.7%) is higher than the redo rate for Male (17.0%).

Multinomial Logistic Regression

Package used: **nnet**

Step 1: Relevel the baseline groups The baseline/reference group was relevelled to be those with neither a Band nor a Ring implant and belongs to the **0-4 yrs** age and **Female** group.

Step 2: Train the model

The multinomial regression model will attempt to learn the relationship on the training data and be evaluated on the test data. It's important to use new data when evaluating our model to prevent the likelihood of overfitting. In this case, split the data into 80% for training and 20% for test sets. We also need to set a seed so that the samples of training and test sets stay consistent.

Step 3: Fit the model and obtain the results

Next fit the model using the training set in R.

Step 4: Find the p-value of each coefficient

Finally we compute the p-value of each coefficient in order to determine their significant impact on the outcome variable `Cause_of_redo`.

Comment: Overall all variables (except for **SexM** variable and **Age_group5-9**) are statistically significant since their p-values are smaller than the significance level at 5%. The model equation for the first row (MR) is then:

For MR patients:

- β_{15} The log odds of having to re-operate due to **MR** vs. having not to re-operate will increase by 26.53103 if moving from **GROUP** = “**NONE**” to **GROUP** = “**BAND**”.
- β_{16} The log odds of having to re-operate due to **MR** vs. having not to re-operate will increase by 17.15680 if moving from **GROUP** = “**NONE**” to **GROUP** = “**RING**”.
- β_{17} The log odds of having to re-operate due to **MR** vs. having not to re-operate will increase by 8.551873 if moving from **Ring_grp** = “**None**” to **Ring_grp** = “**<26 mm**”.
- β_{18} The log odds of having to re-operate due to **MR** vs. having not to re-operate will increase by 8.604928 if moving from **Ring_grp** = “**None**” to **Ring_grp** = “**>=26 mm**”.

The model equation for the second row is:

For MS patients:

- β_{25} The log odds of having to re-operate due to **MS** vs. having not to re-operate will decrease by 11.91916 if moving from **GROUP** = “**None**” to **GROUP** = “**Band**”.
- β_{26} The log odds of having to re-operate due to **MS** vs. having not to re-operate will increase by 20.27758 if moving from **GROUP** = “**None**” to **GROUP** = “**Ring**”.
- β_{27} The log odds of having to re-operate due to **MS** vs. having not to re-operate will increase by 11.654593 if moving from **Ring_grp** = “**None**” to **Ring_grp** = “**<26 mm**”.
- β_{28} The log odds of having to re-operate due to **MS** vs. having not to re-operate will increase by 8.622985 if moving from **Ring_grp** = “**None**” to **Ring_grp** = “**>=26 mm**”.

For both MR and MS patients,

- β_{11} & β_{21} The log odds of having to re-operate vs. having not to re-operate will increase if moving from **age** = “**0-4**” to **age** = “**5-9**”.
- β_{12} & β_{22} The log odds of having to re-operate vs. having not to re-operate will decrease if moving from **age** = “**0-4**” to **age** = “**10-14**”.
- β_{13} & β_{23} The log odds of having to re-operate vs. having not to re-operate will decrease if moving from **age** = “**0-4**” to **age** = “**15-18**”.

Comments: Based on the above statistical findings, it is safe to assume that:

- For patients with **< 26mm (small size) CE ring** implant, there is a higher chance of re-operation caused by mitral stenosis (MS).
- For patients with **Band** implant, there is a higher chance of re-operation caused by mitral regurgitation (MR)

Summary

Relative Risks

The ratio of the probability of choosing one outcome category over the probability of choosing the baseline category is often referred as relative risk. The output coefficients are represented in the log of odds, hence relative risk can be computed by taking the exponential of the intercepts from the linear equation.

Comments:

- The relative risk ratio switching from **Age_group = 0-4 yrs** to **5-9 yrs** is 4.002148 for redo caused by MR vs. no redo at all.
- The relative risk ratio switching from **Age_group = 0-4 yrs** to **10-14 yrs** 8.88×10^{-9} for redo caused by MS vs. no redo at all.

Predicted probabilities

The predicted probability of the outcome variable's categories for each patient can be computed using the following command.

The probability of n^{th} obs being "NONE", "MR" or "MS" shown in the above table can be interpreted in percentage. For example, the probability of the first observation being "NONE" is 100%, it being "MR" is 0.0% and it being "MS" is 0.0%. Thus we can conclude that this observation/patient did not have to have a re-operation in the long run.

Training vs Test Accuracy

The goal is to create a model that generalizes well to new data. For this reason, the performance of how well the model performs applied to both training and test data was evaluated.

Comment: Accuracy in training dataset is 86.32% and the accuracy of the test set is 87.5% , which is slightly higher than that of the training data. This is not ideal since test accuracy should not be higher than that of training as the model is optimized for the latter. The model performs well overall because it performs well not only on the training data but also on the test (unseen) data.

Age group effect plot

The following plot is used to visualise the predictions from the multinomial logistic model.

Comment: The plot shows the difference in the average age trajectory between the "NONE", "MR" and "MS" groups, with the fitted response line for the "NONE" group being significantly above the latter. The fitted probability lines for "MR" and "MS" are identical in 10-14 yrs and 15-18 yrs age groups, and different in 0-4 yrs and 5-9 age groups where the "MS" line fluctuates.

Survival Analysis

Packages used:

- `survival`
- `survminer`

Method: Kaplan-Meier non-parametric survival estimate. Kaplan-Meier curves are especially useful when the predictor variable is categorical (e.g. treatment A vs treatment B; males vs females).

Measures of interest:

- status: censoring status 1 = censored, 2 = relapsed
- sex: male = M, female = F
- time: time to relapse in months

Results

Redo variable is a two level (Yes and No) categorical variable - whether there is a re-operation done for each individual patient. It is used as an indicator to determine if an individual is censored:

$$\text{Redo} \begin{cases} * \text{Yes to re-operation : not censored} \\ * \text{No to re-operation : censored} \end{cases}$$

Data with a plus sign are censored data and otherwise.

The below visualisations depict the survival rate without grouping and with grouping. The survival rate here is actually the rate of no re-operation required due to MR or MS or neither. We refer this to **disease-free** for short.

Comment: The horizontal axis (x-axis) represents time in months, and the vertical axis (y-axis) shows the probability/proportion of having no disease. The stepwise purple line represent survival curve of the population. A vertical drop in the curve indicates an event. The vertical tick mark on the curve means that a patient was censored at this time.

- At time zero, the survival probability is 1.0 (or 100% of the patients are disease-free).
- At time 300, the probability of survival is approximately 0.50 (or 50%).
- The median survival time for the population is 308 months.

Comment: The log-rank p-value for **Sex** variable is lower than significance level $\alpha = 0.05$ as shown in the plot above (“p = 0.096). This is also supported by the p-value of the Sex coefficient in multinomial logistic regression model. In general, Sex is not a significant predictor and it can be concluded that this variable indeed has no effect on the disease-free curve.

Comment: The log-rank p-value for **Age** variable is statistically significant (“p = 0.0015) at 5% level. The survival curves show the 0-4 yrs age group to have less survival/disease-free advantage in comparison to the other groups. The median survival time for age_group = 0-4 yrs is 204 months, as opposed to 242 months for age_group = 5-9 yrs.

Further analysis to evaluate whether the differences between the age groups are statistically different can be done using a Log-Rank test.

The log rank test for difference in survival gives a p-value of p = 0.001, indicating that the age groups differ significantly for the re-operation rate caused by MR, MS or neither in the long run for those with Band implants, Ring implants or neither.

Table of survival analysis showing the first 10 observations

The following table summarizes a list of important components obtained from the `survfit()` function.

Notations:

- n: total number of subjects in each curve.
- time: the time points on the curve.
- n.risk: the number of subjects at risk at time t
- n.event: the number of events that occurred at time t.
- n.censor: the number of censored subjects, who exit the risk set, without an event, at time t.
- lower,upper: lower and upper confidence limits for the curve, respectively.

Source <http://www.sthda.com/english/wiki/survival-analysis-basics>

References

Sharma, Mohit. 2018. *MULTINOMIAL LOGISTIC REGRESSION USING R*. <https://datasciencebeginners.com/2018/12/20/multinomial-logistic-regression-using-r/>.

n.d. *Survival Analysis Basics*. STHDA. <http://www.sthda.com/english/wiki/survival-analysis-basics>.

n.d. *MULTINOMIAL LOGISTIC REGRESSION / R DATA ANALYSIS EXAMPLES*. UCLA:Statistical Consulting Group. <https://stats.idre.ucla.edu/r/dae/multinomial-logistic-regression/>.