PREDICTIVE ANALYTICS CASE STUDY

DSC630-T301 Predictive Analytics (2233-1)

Soma Vayuvegula

Introduction

What was the problem being solved?

The purpose of the predictive analytics case study was to determine the future of healthcare transformations. Predictive analytics in health care provides many benefits in clinical care and day-to-day operations and administration of health care units. As technology emerges, predictive analytics in the healthcare industry will become more prevalent. Predictive analytics in the healthcare industry is aimed at providing optimal care to patients, investigating current findings to make predictions about the future, enhancement of patient care and accurate disease diagnosis, and improving clinical outcomes.

The purpose of Predictive analytics for colorectal cancer was to predict operative mortality, adjusted for risk, in surgery for colorectal cancer.

Why was this problem significant to solve?

Though the current healthcare system has improved many folds over the decade, we still need to be able to accurately diagnose diseases at the early stage to provide better medical assistance to save lives from deadly diseases. Implementation of predictive analytics in the healthcare domain can help diagnose deadly diseases like cancer (colorectal cancer) at the early stage and the risks involved in treating colorectal cancer. A dedicated predictive model needs to be developed for colorectal cancer that estimates the operative risk for individual patients.

The primary outcome was operative mortality. Operative mortality is defined as death occurring within 30 days of an operative procedure, from whatever cause, occurring either in a hospital or after discharge from the hospital.

How was the data acquired?

The colorectal cancer study of the Association of Coloproctology of Great Britain and Ireland (ACPGBI) was conducted in 73 hospitals. The data were obtained ethically in the following ways:

1. Collected from the participating surgeons on a voluntary basis.
2. Data managers dedicated to colorectal cancer.
3. Patients in the local hospitals were newly diagnosed with colorectal cancer.

Collected data is stored electronically in the MS Access database.

Methods and Results

What steps were taken to prepare the data?

Once the data is collected, data of the patients who underwent elective surgery for colorectal cancer are restrained and those who did not undergo surgery and patients without demographic details were removed.

The data in patients' records are edited to check for missing values or values which are out of range or in case of any data inconsistencies.

Out of 7374 (91.3%) of 8077 patients presenting with colorectal cancer satisfied the inclusion criteria. We excluded from the analysis 499 (6.2%) patients who did not have surgical treatment and 204 (2.5%) patients whose records were incomplete.

How was this problem solved?

The unifactorial logistic regression modeling technique is used to identify independent predictors of operative mortality of colorectal cancer. The dataset has been split into 60%-40% of training and testing datasets, respectively.

To improve the accuracy of the model, the median imputation technique is used to substitute the incomplete data. This method allows it to be used by the multifactorial model, and this reduces the 95% confidence intervals around the model estimates and odds ratios.

What modeling techniques were used?

The main modeling techniques used for this predictive analytics case study were the Unifactorial Logistic regression model. Unifactorial Logistic regression is conducted to assess relationships between colorectal cancer and the variables associated with colorectal cancer; therefore, missing values needs to be filled with appropriate values.

Why did the team choose the methods/models they did?

The reason this model was chosen was primarily that they are supervised modeling techniques (i.e., there is a target to check against for the results), and the goal of the case study was to predict operative mortality of colorectal cancer.

What metrics were used to evaluate the results? Why was this metric chosen?

The metric used to evaluate the model is the Hosmer-Lemeshow statistic. The Hosmer-Lemeshow statistic is a statistical test for goodness of fit for logistical regression models. This evaluation method is frequently used in risk prediction models. As this is an evaluation of the risk of operative mortality of colorectal cancer, Hosmer-Lemeshow statistic is best suited.

The model fitted the data well, as evidence from by the calibration Hosmer-Lemeshow statistic shows development set = 5.98, 8 df, P = 0.649; validation set = 6.069, 8 df, P = 0.640. The area under the receiver operating characteristic curve was 0.775 (95% confidence interval 0.744 to 0.806). On subgroup analysis, the predicted mortality for various types of operations for the validation set (n = 3000), as calculated by the ACPGBI colorectal cancer model, was well within the confidence limits of the observed outcome.

Conclusion

How were the results or model implemented?

The results of the case study are formulated into a simple numerical table derived from the statistical model. Clinicians can predict postoperative death by using a simple numerical table derived from the statistical model of the Association of Coloproctology of Great Britain and Ireland (ACPGBI), which may provide patients and carers with an estimated probability of survival from surgery as part of the decision-making process. This statistical model also is an indirect measure of the quality of care.

What were the actionable consequences of the case study?

After the completion of the case study, Clinicians were provided access to a simple numerical table derived from the statistical model. With the current tendency to discharge patients early, in the future, the need might arise to include a combination of in-hospital and 30-day operative mortality.

What did the team learn from the case study?

As cancer and survival against it have become very critical these days, this study will help the patients and carers in decision-making and proves the quality of care provided by the hospitals.

How should or would the team approach the problem differently in the future?

For this case study, the team used Logistic regression methods since there was a target and features to select from for the best predictions. The current model will help us predict the operative mortality within 30 days, from whatever cause, occurring either in a hospital or after discharge from the hospital. With the current tendency to discharge patients early, in the future, the need might arise to include a combination of in-hospital and 30-day operative mortality. Patients with advanced cancers may not be suitable for cancer surgery instead of more straightforward and safer palliative operations.

References

Operative mortality in colorectal cancer: prospective national study.
BMJ 2003; 327 doi: https://doi.org/10.1136/bmj.327.7425.1196 (Published 20 November 2003)

Hosmer-Lemeshow test. Wikipedia. https://en.wikipedia.org/wiki/Hosmer–Lemeshow_test