Developing and validating a multivariable prediction model for predicting costs of colon surgery

Anas Taha[1], Bassey Enodien[2], Vincent Ochs[3], Daniel M. Frey[2], Stephanie Taha-Mehlitz[4]

1. Department of Biomedical Engineering, Faculty of Medicine, University of Basel, Allschwil, Switzerland
2. Department of Surgery, GZO-Hospital, Wetzikon, Switzerland
3. Faculty of Economics,University of Basel, Basel, Switzerland
4. Clarunis, Department of Visceral Surgery, University Center for gastrointestinal and Liver Diseases, St. Clara Hospital and University Hospital Basel, Basel, Switzerland

Correspondance : anas.taha@unibas.ch

Abstract

Hospitals are faced with a burden to predict, calculate and to manage various cost affecting parameters in regard with the patients and their treatments. Accuracy in cost prediction is further affected if a patient suffers from other health issues which hinder the traditional prognosis. This can lead to an unavoidable deficit in the final revenue of these medical centers. This study aims to determine whether machine learning (ML) algorithms can predict the cost factors based on patients with colorectal surgery. For the forecasting, multiple predictors will be taken into the model to provide a tool which can be helpful for hospitals to manage their costs which can end up in operating more cost effectively. This proof of principle will lay the groundwork for an efficient ML-based prediction tool based on multicenter data from a range of international centers in the subsequent phases of the study. With a % MAPE result of 18 – 25.6 the prediction if our model showed decent results in order to forecast the costs in regard to various diagnosed factors and surgery approaches. There is an urgent need for further studies on predicting cost factors, especially for cases with anastomotic leakage to minimize unnecessary costs for the hospitals.

1. Introduction

1.1 Background

Colorectal cancer is one of the most prevalent cancers in the world today based on diagnoses, with about 1.8 million cases being diagnosed and about 0.7 million related deaths occurring annually. The disease accounts for 10 percent of newly diagnosed cancers which means that it is a considerable social and economic burden for many nations around the world (1). One of the treatment options used for colorectal cancer is surgery. Surgical procedure is aimed to re-establish mural and luminal intestinal continuity. Over time, there have been improvements in the way the disease is treated. But the already existing disorders in patients or comorbidities have limited the surgical procedures. The time required to dress the patients for surgery and their comorbidities are the contributing factors for increased surgical cost. However, despite many improvements, there are still significant other complications which occur during especially after a surgical procedure. To avoid this patient is required to be in a necessary post-operative care, generally for 5 and 7 days after a surgical operation. Some of the post-operative risk factors will further add to surgical cost, but their prediction is very vague due to absence of sufficient datasets. These involve leak of liminal contents from a surgical join (2), delirium or prolonged ileus (3), other emergency surgery, longer inter-operative time, peritoneal contamination, and rectal anastomosis.

The comorbidities and longer stays, results in a cost burden for patients. It is for this reason that prediction models are now being updated to determine the costs for anastomotic insufficiency. Prediction models are normally used to estimate the probability of achieving a particular outcome in future (4). A large number of prediction models have been developed but only a small number are used because not all models accurately predict the desired outcome (5). In this study, the focus is to develop and validate a multivariable prediction model to predict costs for patients undergoing Colon Surgery while considering their stay period in hospital. This will help determine the cost burden due to variable stay time in ICU and hospital during the complete surgery. The medical context is prognostic in that is focused on predicting the likely cost of overall expenditure involved in colon surgery for the clinical center.

1.2 Rationale

The rationale for developing and validating the multivariable model is it will help to accurately predict the costs associated with colon surgery. Accurate prediction will help inform actions taken by patients, practices employed by hospitals, and policies enacted by government. The results that come with the use of the model will also help in planning. In short, developing and validating the multivariable model will help provide insight into the costs involved in colon surgery which will, in turn, revise and develop strategies to manage it in a better way.

1.3 Objectives

- To develop predicting models for the final costs in patients based on multiple predictors.

- To test the models in terms of their ability to accurately predict the final costs associated with colon surgery in patients.

2. Methods

2.1 Overview and Data Collection

Data was extracted from a registry of patients who underwent Colocolic anastomosis for various reasons like tumors, diverticulitis, mesenterial ischemia, iatrogenic or traumatic perforation, or inflammatory bowel disease at the Hospital of Wetzikon from 1st January 2013 to the 31st December of 2019. Furthermore, this study was completed based on the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement checklist for prediction model development (6). Utilizing this data, we developed a machine learning model with the aim of predicting the costs for colon surgery in patients.

2.2 Ethical Considerations

The registry data was approved by an institutional review Board, where the patients' informed consent was waved. The Registry was registered at [Req 2021-01107]

2.3 Predictors and Outcome Measures

Recorded variables include Insurance (General/Semi-Private/Private) Age, Operation procedure (Hartmann/Hemicolectomy left and Extended Hemicolectomy left/Hemicolectomy right and Extended Hemicolectomy right/Sigmoidectomy), Surgical Approach (Open/Laparoscopic), Diagnosis (Tumor/Diverticulitis), Final costs (the sum of all cost factors), Hospitalization (in days), Intensive care (in days), Operation time (in minutes), Anesthesia time (in minutes).The data on the Final cost which is the sum of all the costs incurred during stay in hospital for surgery were collected in CHF (Swiss Franc).

2.4 Model Development

Data was randomly split into two sets, 80% of the data was put into a training set to build the models and 20% was utilized for a test set to internally validate the models and assess their performance. The two sets had approximately the same class distribution (Gaussian). The following 11 predictors were chosen to predict the final costs based on regression and clinical insights: Age, Gender, Insurance, Diagnosis, Operation, Surgery Approach, Hospitalization, Intensive care , Operation and Anesthesia time(9).

A variety of machine learning models were developed including Generalized Boosted Regression, Random Forest, Decision trees. An interaction depth of 3 and total number of 500 trees were chosen. As the type of the random forest, the regression model was chosen. The Classification/Predictive performance was measured using the Mean Absolute Percentage Error (MAPE), where a result of < 10% is classified as highly accurate, <20% denotes as a good forecast 20%-50% as a reasonable forecast and everything >50% as an inaccurate forecast Abidin and Jaffar (10). The MAPE factor, also known as mean absolute percentage deviation (MAPD) is

a measure which is used for the accuracy of a forecasting prediction. Continuous data are reported as mean ± SD or median (IQR) and categorical data as numbers (percentages). Hyperparameters were tuned, and the final model was selected based on MAPE. The final model was the random forest based on its performance.

The analysis was carried on Using R 4.0.4. The Random Forest Library was used for the random forest models, the Metrics Library was used for the calculation of the performance measurements, the gbm library was used for the generalized boosted regression models and the rpart Library was used for the other models.

2.5 Deployment

The best performing model will be deployed as a web-based, user-friendly application using RShiny to predict the final costs which consists of the different cost factors (accessible via: http://anas2mosis.shinyapps.io/updated)

3. Results

3.1 Cohort

A total of 344 patients were included for our study. This number includes all patients from the center which suffered from the diagnosed factors of section 3 and had to undergo the type of operations mentioned. The mean age was 67.5 ± 13.8 (range from 28 to 94). Furthermore, 214 (73.6%) patients were male, while 130 (37.8%) were female. Table 1 provides all baseline variables and the outcome. Continuous variables will be recorded as mean ± STD (min value, max value). Moreover, categorical variables will be recorded as numbers (percentage). No missing values were detected. Table 1 provides all baseline variables and the outcome.

Table 1. Variable and outcome characteristics

Variable	*Overall*	*Mean (SD)*	*min*	*Max*	*Range*	*Median (IQR)*
Insurance	344	-	-	-	-	-
General	280	-	-	-	-	-
Semi-Private	49	-	-	-	-	-
Private	15	-	-	-	-	-
Age	344	67.45(13.79)	28	94	66	20
Diagnose	344	-	-	-	-	-
Tumor	167	-	-	-	-	-
Diverticulitis	177	-	-	-	-	-
Operation	344	-	-	-	-	-
Hartmann	20	-	-	-	-	-
Hemicolectomy left	22	-	-	-	-	-
Hemicolectomy right	86	-	-	-	-	-
Sigmoidectomy	216	-	-	-	-	-
Surgery approach	344	-	-	-	-	-
Open	145	-	-	-	-	-
Laparoscopic	199	-	-	-	-	-
Hospitalization	344	9.37(9.78)	2	84	82	7
Icu Days	344	1.48(6.07)	0	70	70	0
Operation Time	344	171.67 (82.25)	23	571	548	82.5
Anesthesia	344	114.26 (59.86)	45	486	441	49.25
Final Costs	344	-32747.82 (45922.03)	- 520591.14	- 7484.74	513 106.4	1386 6.06

3.2 Model Performance

During internal validation, the performance of all three models was tested and stated with their mean value and their 95% confidence intervals (Table 2). The Random Forest Classifier provided the highest MAPE for predicting the final costs (21.4). Thus, it was the model with the best Internal Validation performance and subsequently used for predicting the costs (11). In comparison, the Decision Tree and General Boosted Regression Model displayed results for MAPE of only 25.5 and 29.7, respectively. Therefore, the average of the MAPE for the final costs is around 21.4 which means that on average the forecast of this prediction model regarding the final costs are off by 21.4%. Since a MAPE value of <20% is considered as being "good", our result is showing decent results. The percentage of the Variance of the Random Forest Classifier which got explained in the models varied from 73.81% to 81.05%. Specific Feature importance according to the Random Forest Classifier is displayed as Gini Index in

Figure 1, while Figure 2 shows the prediction of the Random Forest Classifier compared to the actual observed values from the test data set for the final cost factor.

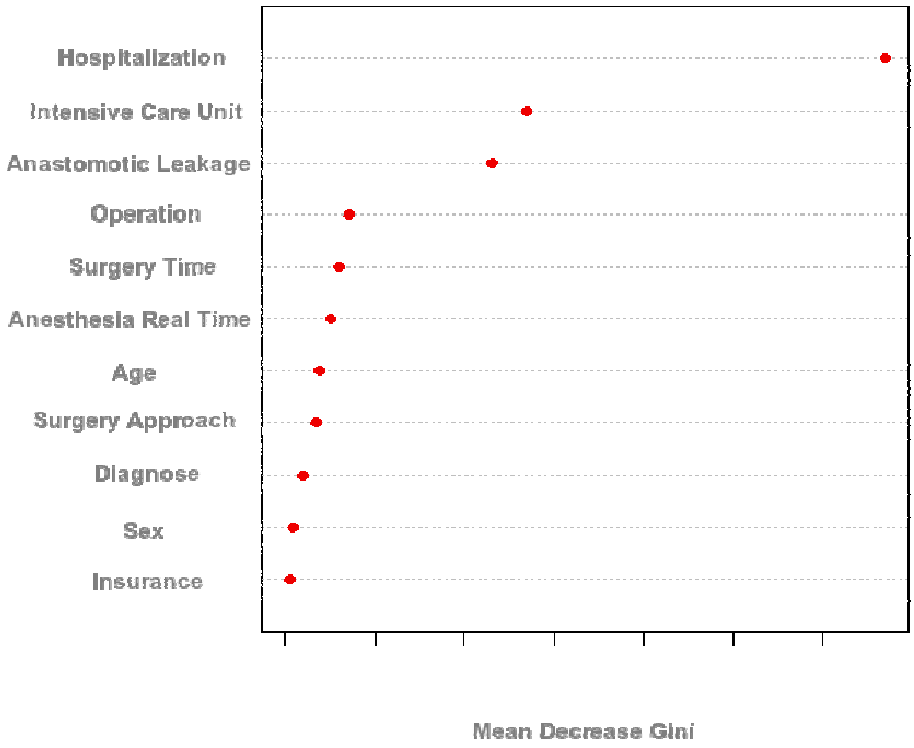

Figure 1. Total decrease in node impurities, measured by the gini index from splitting on the variable, averaged over all trees.

In figure 1 one can see those factors like intensive care unit (in days) and hospitalization (in days) are the best predictors in our model, which could be explained as being variables which are often correlated with complications of an operation and end up in being more expensive. The factor hospitalization can be explained as being a good predictor for costs because the overall costs for a hospital will increase since the patient is not feeling better after a surgery. Mean Decrease in Gini is the average (mean) of a variable's total decrease in node impurity, weighted by the proportion of samples reaching that node in each individual decision tree in the random forest. A higher Mean Decrease in Gini indicates higher variable importance.

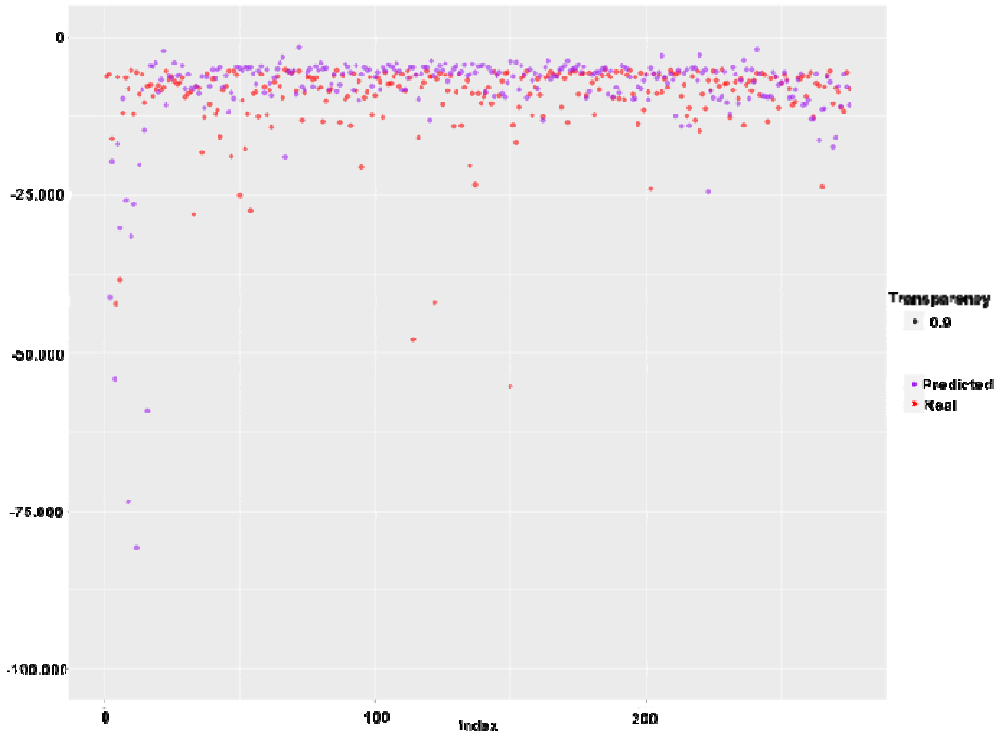

Figure 2. Predicted vs. Real observations of the model.

4. Discussion

4.1 Limitations of the study

The main limitation of the study is a lack of representative sample. A representative sample is a sample selected in a way that makes it a representative for the target population. In this case, the focus was on patients undergoing colon surgery. However, in the sample dataset for only 344 individuals met this criterion. This implies that the sample was not selected in the manner that made it a representative of patients undergoing colon surgery. The larger the dataset, more accurate are the results. The limited number of individuals with common reason for higher cost implies that it was not possible to effectively test the developed models in terms of their ability to predict costs associated with the disease. For such models, there is need for adequate and detailed data in order to ensure they are tested thoroughly. Additionally, an overall increase of the sample size could end up in having more precise models by looking at the values of Table 2. Especially the events per predictor should be bigger.

Table 2. Internal validation performance for the 3 developed models

Classifier	MAPE (%) Final Costs
Random Forest	21.4 (17.2-26.8)
Decision Tree	25.2 (21.4-26.3)
Generalized Boosted Regression	29.7 (25.2-34.2)

Note: Scores Reported as Mean (95% Confidence Interval)

4.2 Interpretation of results

As indicated, in this study, three models were developed and tested. They are Random Forest, Decision tree, and Generalized Boosted Regression. An examination of the results shows that Random Forest has the lowest percentage for all the costs examined on MAPE.

The lowest MAPE % for Random Forest model is an indication that this model is the most accurate when it comes to predicting costs associated with surgeries when compared to the other two models examined. Normally, MAPE is a measure of error. It is used to measure the accuracy of a forecast (12). In calculating MAPE, the difference between the actual value and the forecast value is determined and expressed as a percentage. This means that if the difference between the actual value and the forecast value is small, the percentage is small (13). On the other hand, if the difference between the actual value and the forecast value is big, MAPE percentage is big. This implies that a small MAPE percent is an indication of the forecast value being near the actual value. In other words, the forecast value is more accurate (14). In the case of the three models, since Random Forest model has the lowest MAPE percent value for all the costs when compared to the other models considered, it is the most effective model in terms of predicting the cost.

The question is, why is Random Forest the most effective predictive model when compared to the Decision Tree and Generalized Boosted Regression models? This question can be answered by examining the model. Random Forest model is a machine learning technique which is used to solve classification and regression problems (15). This model uses ensemble learning, a technique which combines many classifiers to obtain solutions to complex problems. A random forest algorithm comprises of multiple decision trees. The forest that is generated by the algorithm is trained through bootstrap aggregating or bagging (16). Bagging is a meta-algorithm which improves the machine learning algorithms' accuracy.

The random forest algorithm establishes the result from the predictions of decision trees. It predicts by taking the mean or average of the prediction output of the various trees (17). This implies that the predicted outcome by the algorithm becomes more accurate when the number of decision trees is increased.

One of the features of the Random Forest model which makes it more accurate in predicting cost outcome is it reduces the overfitting problem normally experienced when using the decision tree model. As indicated, the model is uses ensemble learning method and is based on bagging (15). This means that the model creates many decision trees and then considers the outcomes of all the trees in its final prediction. This enhances the accuracy of the prediction by the model.

However, despite the higher accuracy of Random Forest model when compared to the Decision Tree and Generalized Boosted Regression models, the model does not have the highest possible accuracy when considered alone. Normally, when examining the accuracy of a prediction using MAPE, as indicated, the result of less than 10% is considered highly accurate. A MAPE score of less than 20% denotes a good forecast while that between 20 and 50% is considered reasonable forecast (12). Looking at the results, it shows that Random Forest model

gives mostly reasonable forecasts rather than accurate forecast. The model gave an outcome of over 20% when analyzed using the MAPE. This means that while it is the most accurate model when compared to the other models, when considered alone, it has only considerable accuracy and that it does not accurately predict the cost incurred.

A number of similar studies have been carried out on the Random Forest model in terms of its accuracy of predicting outcome. For example, in their study, Dimopoulos et al. (2018) examine the prediction accuracy of Random Forest model when applied for residential mass appraisals (18). In examining the prediction accuracy of the model, its results were compared to that of Generalized Boosted Regression model. It was established that the Random Forest model exhibited higher prediction accuracy. The results in the study by Dimopoulos et al. (2018) are similar to those of this study which also show that the Random Forest model has a higher level of prediction accuracy when compared to other studies. However, the shortcoming with the study by Dimopoulos et al. (2018) is it compares the Random Forest model to only one other model. This does not provide adequate insight into the prediction accuracy of the model. A comparison with other models other than Generalized Boosted Regression model would have helped determine whether Random Forest Model is the most accurate prediction model or whether there are other more accurate models.

Another similar study is that of Guo et al. (2020) developed and tested a shot accuracy prediction model based on the Random Forest classification algorithm (19). To evaluate the performance of the model, the researchers compared it to other models which include logistic regression, support vector machine, decision tree, k-nearest neighbors, Adaboost and gradient boosting decision tree. It was established that the Random Forest-based prediction model had the highest level of accuracy when compared to all the other prediction models examined. This implies that Random Forest provides the most accurate outcomes when it is used for prediction. The results in the study by Guo et al. (2020) are in line with those of this study since it was also established that the Random Forest model is the most accurate when compared to other models. The study by Guo et al. (2020) provides better insight into the accuracy of Random Forest model because it compares it multiple models (19). It is an indication that the Random Forest model is one of the most accurate prediction models that can be used to predict costs for surgery.

Lastly, the results are in line with those of Song et al. (2021) who also established that the Random Forest model has a higher accuracy when compared to other models (20). In the study, Song et al. (2021) built and tested machine learning models for predicting pressure ulcer nursing adverse event in order to find an optimal model which accurately predicts the occurrence of pressure ulcers (20). The models used to predict the occurrence of pressure ulcer are support vector machine, decision tree model, artificial neural network model, and random forest model. The results obtained show that all the models perform fairly well in terms of predicting the occurrence of pressure ulcers. However, when they were compared, it was found that Random Forest model was a better predictor of the occurrence of pressure ulcer.

Basically, the results in this study show that all the models have reasonable accuracy as the MAPE for each of them for all the costs highlighted is below 50%. This means that all of the

models can be used to predict the costs to some level of accuracy. However, when compared, it can be seen, that Random Forest model is a more accurate predictor. These results are also evident in similar studies which also show that Random Forest model is a more accurate prediction model.

4.3 Implications

One of the implications of the results is that hospitals and other concerned parties can employ Random Forest model to forecast costs not only for colon surgery but also the costs of other risks and conditions, mentioned previously. A look at the results show that Random Forest model is provides more accurate predictions when compared to other models like Generalized Boosted Regression and Decision Tree models. It means that for concerned parties to achieve more accurate results when predicting the costs of conditions or any other outcome, they have to employ the Random Forest model.

Another implication is that there is need for further research about the model in terms of the ways of enhancing the accuracy of the Random Forest model. The results show that for the final costs examined, the accuracy is more than 20%. This is only reasonable accuracy. However, it is way before the desired value. As indicated, the MAPE value of less than 20% is indication of a good forecast while that of less than 10% shows that the forecast is highly accurate. While achieving a highly accurate forecast is unlikely, any good prediction model should give a good forecast. With Random Forest model being the most accurate model, it implies that it should be developed further to ensure its accuracy is improved so that it gives more credible results when used to predict outcomes meaning further research is needed on the model.

5. Conclusion

Essentially, colorectal cancer is a common within patients who have undergone a surgical procedure for cancer. This post operative conditions increases the cost burden for patients and hospitals. However, there is no way of predicting these costs so that a patient or healthcare can prepare adequately to handle the condition. This study thereby aimed to develop and validate a prediction model so that the costs can accurately be predicted, and strategies developed to eliminate or cover them. Three models were developed and tested to determine the most accurate one. They are Random Forest, Decision tree, and Generalized Boosted Regression Model. Out of the three models, the results obtained based on MAPE analysis showed that the Random Forest model is the most accurate. For all the costs, this model had the lowest MAPE percentage. However, apart from one cost, prediction for all the other cost was over 20%. The results imply that the Random Forest model should be adopted for prediction. However, the fact that MAPE results were mostly over 20% for the model means that research should be done on how to improve its accuracy further.

Author Contributions: Conceptualization A.T., data collection B.E.; analysis V.O, A.T.; visualization A.T. writing—original draft preparation A.T, V.O., writing—review and editing D.F., S.T. M; All authors have read and agreed to the published version of the manuscript

Funding: This research received no external funding.

Ethics: Cantonal Ethics Committee Zurich, Switzerland BASEC Nr. 2021-01107

Data Availability Statement: The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Xi, Y. and Xu, P., 2021. Global colorectal cancer burden in 2020 and projections to 2040. *Translational Oncology*, *14*(10), p.101174.
2. Soeters, P.B., de Zoete, J.P.J.G.M., Dejong, C.H.C., Williams, N.S. and Baeten, C.G.M.I., 2002. Colorectal surgery and anastomotic leakage. *Digestive surgery*, *19*(2), p.150.
3. Karliczek, A., Harlaar, N.J., Zeebregts, C.J., Wiggers, T., Baas, P.C. and Van Dam, G.M., 2009. Surgeons lack predictive accuracy for anastomotic leakage in gastrointestinal surgery. *International journal of colorectal disease*, *24*(5), pp.569-576.
4. Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V. and Fotiadis, D.I., 2015. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, *13*, pp.8-17.
5. Mosavi, A., Ozturk, P. and Chau, K.W., 2018. Flood prediction using machine learning models: Literature review. *Water*, *10*(11), p.1536.
6. Collins, G.S., et al., Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. 2015. 102(3): p. 148-158.
7. Gessler, B., O. Eriksson, and E. Angenete, Diagnosis, treatment, and consequences of anastomotic leakage in colorectal surgery. Int J Colorectal Dis, 2017. 32(4): p. 549-556
8. Rahbari, N.N., et al., Definition and grading of anastomotic leakage following anterior resection of the rectum: a proposal by the International Study Group of Rectal Cancer. Surgery, 2010. 147(3): p. 339-51.
9. Bolenz, Gupta, Roehrborn, Lotan, Predictors of costs for robotic assisted laparoscopic radical prostatectomy, Urologic Oncology: Seminars and Original Investigations, Volume 29, Issue 3, 2011, Pages 325-329.
10. Abidin, S. and Jaffar, M.M. (2014) Forecasting Share Prices of Small Size Companies in Bursa Malaysia. Applied Mathematics and Information Sciences, 8, 107-112
11. Sushmita, Newman, Marquardt, Ram, Prasad, De Cock, Teredesai, 2015. Population Cost Prediction on Public Healthcare Datasets. In Proceedings of the 5th International Conference on Digital Health 2015. Association for Computing Machinery, New York, NY, USA, 87-94.
12. Coleman, C.D. and Swanson, D.A., 2007. On MAPE-R as a measure of cross-sectional estimation and forecast accuracy. *Journal of Economic and Social Measurement*, *32*(4), pp.219-233.
13. Hyndman, R.J. and Koehler, A.B., 2006. Another look at measures of forecast accuracy. *International journal of forecasting*, *22*(4), pp.679-688.

14. Rayer, S., 2007. Population forecast accuracy: does the choice of summary measure of error matter?. *Population Research and Policy Review*, *26*(2), pp.163-184
15. Sinha, P., Gaughan, A.E., Stevens, F.R., Nieves, J.J., Sorichetta, A. and Tatem, A.J., 2019. Assessing the spatial sensitivity of a random forest model: Application in gridded population modeling. *Computers, Environment and Urban Systems*, *75*, pp.132-145
16. Bharathidason, S. and Venkataeswaran, C.J., 2014. Improving classification accuracy based on random forest model with uncorrelated high performing trees. *Int. J. Comput. Appl*, *101*(13), pp.26-30.
17. Wang, L., Liu, Z.P., Zhang, X.S. and Chen, L., 2012. Prediction of hot spots in protein interfaces using a random forest model with hybrid features. *Protein Engineering, Design & Selection*, *25*(3), pp.119-126.
18. Dimopoulos, T., Tyralis, H., Bakas, N.P. and Hadjimitsis, D., 2018. Accuracy measurement of Random Forests and Linear Regression for mass appraisal models that estimate the prices of residential apartments in Nicosia, Cyprus. *Advances in Geosciences*, *45*, pp.377-382.
19. Guo, J., Yang, L., Umek, A., Bie, R., Tomažič, S. and Kos, A., 2020. A Random Forest-Based Accuracy Prediction Model for Augmented Biofeedback in a Precision Shooting Training System. *Sensors*, *20*(16), p.4512.
20. Song, J., Gao, Y., Yin, P., Li, Y., Li, Y., Zhang, J., Su, Q., Fu, X. and Pi, H., 2021. The Random Forest Model Has the Best Accuracy Among the Four Pressure Ulcer Prediction Models Using Machine Learning Algorithms. *Risk Management and Healthcare Policy*, *14*, p.1175.