

Predicting Quality of Life Following Emergency Laparotomy

Bailey Millns

October 2021

School of Mathematics,

Cardiff University

A dissertation submitted in partial fulfilment of the

requirements for MSc (in Data Science and Analytics)

by taught programme.

|  |  |
| --- | --- |
| **CANDIDATE’S ID NUMBER** | C2098580 |
| **CANDIDATE'S SURNAME** | Mr. Millns |
| **CANDIDATE'S FULL FORENAMES** | Bailey Callan |

**DECLARATION**

This work has not previously been accepted in substance for any degree and is not concurrently submitted in candidature for any degree.

Signed ……………………………………………. (candidate) Date 19/10/2021



**STATEMENT 1**

This dissertation is being submitted in partial fulfilment of the requirements for the degree of MSc

Signed ……………………………………………. (candidate) Date 19/10/2021



**STATEMENT 2**

This dissertation is the result of my own independent work/investigation, except where otherwise stated. Other sources are acknowledged by footnotes giving explicit references. A Bibliography is appended.

Signed ……………………………………………. (candidate) Date 19/10/2021



**STATEMENT 3 –**

I hereby give consent for my dissertation, if accepted, to be available for photocopying and for public viewing, and for the title and summary to be made available to outside organisations.



Signed ……………………………………………. (candidate) Date 19/10/2021

**STATEMENT 4 - BAR ON ACCESS APPROVED**

I hereby give consent for my dissertation, if accepted, to be available for photocopying and for public viewing **after expiry of a bar on access approved by the Graduate Development Committee.**

Signed ……………………………………………. (candidate) Date 19/10/2021



**Executive Summary**

**Introduction**

Emergency laparotomy is a high-risk central surgery associated with a high mortality rate, a prolonged overall length of stay in the hospital, and reduced quality of life following the surgery. There is a lack of research on the quality of life following emergency laparotomy compared to mortality. The purpose of this study is to develop machine learning models whose output can be used to advise patients about potential difficulties during their recovery and to motivate additional research and data collection on postoperative morbidity. Three different models will be used in this project to predict the quality of life following surgery.

**Methodology**

Cardiff and Vale University Hospital identified all patients who underwent an emergency laparotomy from Jan 2016 to December 2019 via their local NELA database. Additional process outcomes (beyond the 30-day NELA period) were obtained retrospectively from patient medical records. All patients still living at the time of the study were sent The European Quality of Life Five Dimension (EQ-5D) questionnaire. After six weeks, with no answer, a telephone call was made to finish the questionnaire. The EQ-5D is an extensively used and well-established generic instrument for evaluating health-related quality of life.

A set of models were developed to predict the outcome of the EQ-5D questionnaire. The questionnaire consisted of two components: the EQ-5D descriptive system and the EQ visual analogue scale (EQ-VAS), which required different types of models to make a prediction. The EQ-5D descriptive system consists of five dimensions (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression) with five levels/categories that the patient assigns to their current health status; this is a multiclassification problem. The EQ-VAS is a number assigned by the patient to their current state of health; this is a regression problem. Additionally, a separate set of models was developed to predict whether a patient would be referred to a mental health or counselling specialist by their general practitioner (GP); this is a binary classification problem. Out of 1054 patients, 262 responded to all EQ-5D questions, 46 were referred to a mental health specialist, and 19 were referred to a counselling specialist.

A neural network and a multinomial logistic regression model were developed to predict one of the EQ-5D’s five levels, with separate models developed for each of the five dimensions. Linear regression, decision trees, and random forest regression models were used to predict the VAS life score. K Nearest Neighbours (KNN), logistic regression and Support Machine Vectors (SVM) were used to predict whether a patient will be referred to a mental health specialist and if they will be referred to a counselling specialist.

**Results**

The multiclass classification models were evaluated using the balanced accuracy metric. The multiclass classification results indicated that the neural network was the best model for: mobility: 0.230 (0.068), anxiety/depression: 0.244 (0.062), and self-care: 0.255 (0.043). The multiclass classification results indicated that the best model was multinomial logistic regression for pain/ discomfort: 0.257 (0.067) and usual activities: 0.303 (0.055). The regression models were evaluated using the Root Mean Square Error (RMSE) metric. Random forest obtained the lowest RMSE value of 21.64 (9.89). The binary classification models were evaluated using the ROC AUC score. The logistic regression model achieved the highest AUC values for both referral to mental health and referral to counselling specialist, with 0.691 (0.072) and 0.661 (0.156), respectively.

**Discussion**

The multiclass classification model’s performance was inadequate and not suitable for clinical settings. The balanced accuracy obtained from the train and test datasets suggests that the models are overfitting the data. This was possibly due to the class imbalance, oversampling and undersampling were applied to the training data, but the testing data retained a class imbalance. Due to the severity of the imbalance, there were very few instances of the minority class in the test data. It was recommended that future work could attempt to equalise the distribution by obtaining more samples of the minority class. It is also possible that the training data features are not highly predictive of the target.

RMSE is a relative measure and context-dependent on the problem, the data, and the dependent variable. The value obtained in the study should be validated using the same data to predict the VAS score following emergency laparotomy; The value obtained in this study appears to be relatively high for an RMSE value, but when validated, it may prove to be a better, more suitable value. The train r squared was significantly higher than the test r squared, suggesting that the model is overfitting. Further research into feature selection techniques and dimensionality reduction techniques may aid in reducing overfitting.

The binary classification model has adequate performance with moderate discrimination. The values of 0.69 and 0.66 for AUC are considered reasonable discriminating abilities when identifying patients who will be referred to a mental health or counselling specialist, respectively. Both models had low precision and low recall, which could be improved with more positive class samples (referred to the mental health/ counselling specialist). A hyperparameter search could be conducted on specific combinations of undersampling the negative class and overweighting the loss on the positive class.

The study had several limitations. Since no baseline EQ-5D score was obtained prior to surgery and different patients were asked to rate postsurgical morbidity at varying time points following surgery, we cannot conclusively state that the surgery was the cause of the score. Due to the study’s exclusion of the deceased, patients with dementia, and patients who do not speak English, there is selection bias. Additionally, if patients are in such poor health (EQ-5D = 5) that they are unable to complete the survey, there may be a non-response bias. Patients who completed the EQ-5D questionnaire may have been removed from the analysis since rows with missing values (less than 5 per cent) were removed from the data.

**Acknowledgements**

I would like to express my gratitude to the colorectal research group at Cardiff and Vale University Hospital, particularly Julie and Louise. Thank you Louise, for the advice, assistance, suggestions and patience throughout the project, and Julie, for your suggestions and feedback, and for being on hand if I needed anything.

Thank you to my university supervisor, Meirion, for your consistent support and guidance during the running of this project. Despite your busy schedule, I appreciate your willingness to have zoom meetings with me even when the topic is not related to the project.

**Contents**

[List of Figures 8](#_Toc85566621)

[List of Tables 10](#_Toc85566622)

[1 Introduction 11](#_Toc85566623)

[1.1 Background 11](#_Toc85566624)

[1.1.1 Emergency Laparotomy 11](#_Toc85566625)

[1.1.2 NELA 14](#_Toc85566626)

[1.1.3 The EQ-5D 17](#_Toc85566627)

[1.1.4 Referral to psychological therapy 18](#_Toc85566628)

[2 Data 20](#_Toc85566629)

[2.1 Data source 20](#_Toc85566630)

[2.2 Data description 20](#_Toc85566631)

[2.3 Data pre-processing 24](#_Toc85566632)

[3 Methods 28](#_Toc85566633)

[3.1 Multiclass classification 28](#_Toc85566634)

[3.2 Regression 31](#_Toc85566635)

[3.2.1 Linear regression 32](#_Toc85566636)

[3.2.2 Decision trees 32](#_Toc85566637)

[3.2.3 Random forest 33](#_Toc85566638)

[3.3 Binary classification 34](#_Toc85566639)

[3.3.1 KNN 34](#_Toc85566640)

[3.3.2 Logistic regression 35](#_Toc85566641)

[3.3.3 SVM 35](#_Toc85566642)

[4 Results 37](#_Toc85566643)

[4.1 K-fold cross-validation 37](#_Toc85566644)

[4.2 Dimensionality reduction 38](#_Toc85566645)

[4.3 Multiclass classification 38](#_Toc85566646)

[4.4 Regression 40](#_Toc85566647)

[4.5 Binary classification 41](#_Toc85566648)

[5 Discussion 44](#_Toc85566649)

[5.1 Multiclass classification 44](#_Toc85566650)

[5.2 Regression 44](#_Toc85566651)

[5.3 Binary classification 45](#_Toc85566652)

[5.4 Limitations and further work 45](#_Toc85566653)

[6 References 47](#_Toc85566654)

[7 Appendix 52](#_Toc85566655)

[7.1 Appendix A: Model functions 52](#_Toc85566656)

[7.1.1 Multiclass classification 52](#_Toc85566657)

[7.1.2 Regression 53](#_Toc85566658)

[7.1.3 Binary classification 53](#_Toc85566659)

[7.2 Appendix B: Precision, recall and false positive rate definitions 55](#_Toc85566660)

[7.2.1 Precision 55](#_Toc85566661)

[7.2.2 Recall 55](#_Toc85566662)

[7.2.3 False positive rate 55](#_Toc85566663)

[7.3 Appendix C: Participant flow diagram 56](#_Toc85566664)

[7.4 Appendix D: Feature list 57](#_Toc85566665)

# List of Figures

[Figure 1: Path of emergency laparotomy patient 13](#_Toc85566513)

[Figure 2: The overall unadjusted 30-day and 90-day ONS mortality rates trend by NELA dataset year. 14](#_Toc85566514)

[Figure 3: Visualisation of EQ-5D target variables. 21](#_Toc85566515)

[Figure 4: Box and whisker plot showing the WIMD 2019 rank of EQ-5D responders and the rest of the patients. 22](#_Toc85566516)

[Figure 5: Visualisation of the referral to mental health and counselling columns. 24](#_Toc85566517)

[Figure 6: Showing class imbalance for target column “Self.care” 25](#_Toc85566518)

[Figure 7: Showing class distribution after oversampling then undersampling the data. 26](#_Toc85566519)

[Figure 8: Architecture of the deep neural network for the multiclass classification. 29](#_Toc85566520)

[Figure 9: Deep neural network for the multiclass classification model. 29](#_Toc85566521)

[Figure 10: Multinomial logistic regression model 31](#_Toc85566522)

[Figure 11: Box and whisker plot showing the average (over all k-folds) balanced accuracy for different hyperparameters. 31](#_Toc85566523)

[Figure 12: Decision tree diagram 32](#_Toc85566524)

[Figure 13: Random forest technique 33](#_Toc85566525)

[Figure 14: KNN algorithm (K=1 and K=100) 34](#_Toc85566526)

[Figure 15: Logistic regression for binary classification. 35](#_Toc85566527)

[Figure 16: SVM algorithm visualisation 36](#_Toc85566528)

[Figure 17: The process of k-fold cross-validation (k=5) 37](#_Toc85566529)

[Figure 18: Scikit-learn pipeline. 38](#_Toc85566530)

[Figure 19: Residuals plot for random forest 41](#_Toc85566531)

[Figure 20: Model evaluation function for multiclass classification. 52](#_Toc85566532)

[Figure 21: Multinomial logistic regression model and hyperparameter tuning. 52](#_Toc85566533)

[Figure 22: Neural network baseline model 52](#_Toc85566534)

[Figure 23: Neural network hyperparameter tuning 52](#_Toc85566535)

[Figure 24: Linear regression model and hyperparameter tuning. 53](#_Toc85566536)

[Figure 25: Decision tree model and hyperparameter tuning. 53](#_Toc85566537)

[Figure 26: Random forest model and hyperparameter tuning. 53](#_Toc85566538)

[Figure 27: Binary classification model evaluation function. 53](#_Toc85566539)

[Figure 28: KNN model and hyperparameter tuning. 54](#_Toc85566540)

[Figure 29: Logistic regression model and hyperparameter tuning. 54](#_Toc85566541)

[Figure 30: SVM model and hyperparameter tuning. 54](#_Toc85566542)

[Figure 31: Patient flow diagram 56](#_Toc85566543)

# List of Tables

[Table 1: Showing the demographics between the EQ-5D patients and the remaining patients 23](#_Toc85566455)

[Table 2: Average results over all K-folds for all EQ-5D target variables 39](#_Toc85566456)

[Table 3: Balanced accuracy score for the training and test data set for all EQ-5D target variables. 39](#_Toc85566457)

[Table 4: Average results over all K-folds for the regression problem 40](#_Toc85566458)

[Table 5: Average results over all K-folds for binary classification problem. 41](#_Toc85566459)

[Table 6: AUC score for the training and test data set for mental health and counselling for the logistic regression model (best model). 42](#_Toc85566460)

[Table 7: Precision and recall for both referral to mental health and counselling for both classes. 42](#_Toc85566461)

[Table 8: List of features used in all models 57](#_Toc85566462)

# 1 Introduction

Each year, approximately 25,000 emergency laparotomies are performed in the UK. Emergency laparotomy is a common surgical procedure associated with a high rate of mortality, complications, and diminished quality of life (NELA 2020). Emergency laparotomy has a higher morbidity and mortality rate than elective procedures. Since 2013, the National Emergency Laparotomy Audit (NELA) has been evaluating the care and outcomes of patients in England and Wales who require this type of high-risk surgery. NELA was established with the goal of enhancing the quality of care for patients undergoing emergency laparotomy. Almost all NELA’s research has focused on mortality, with minimal research paid to the quality of life and morbidity following surgery.

Patients who have undergone an emergency laparotomy often seek the help and support of their general practitioner (GP). Based on the patient’s condition, the GP may then refer the patient for psychological therapy. Researchers at Cardiff and Vale UHB sought to capture this data as it provides a qualitative measure of the quality of life following the surgery.

This project aims to predict a patient’s quality of life following emergency laparotomy. Three models have been developed: a multiclass classification model, a regression model, and a binary classification model. The multiclass classification model predicts one of five EQ-5D scores. The EQ-5D is a health questionnaire in which patients rate their health on a scale of one to five in five different dimensions. Regression analysis is used to predict the VAS life score. The VAS score is a numerical value between zero and one hundred that the patient assigns to their current state of health. The binary classification model predicts whether a patient will be referred by their GP to a mental health specialist. A separate binary classification model predicts whether a patient will be referred to a counselling specialist, by their GP.

## 1.1 Background

### 1.1.1 Emergency Laparotomy

An emergency laparotomy is a major procedure that involves the abdomen being opened. This enables the surgeon to examine the organs and fix any emergency complications. It is referred to as an “emergency” as it is often a life-saving procedure to address an acute abdomen (Pearse RM et al. 2012). Emergency laparotomies are often conducted to treat infections caused by perforated or inflamed intestine, intestinal obstruction, or internal bleeding. Several additional conditions may need an emergency laparotomy, including gall bladder or appendix perforations or infections and abdominal trauma. There is no viable alternative to an emergency laparotomy in the majority of these instances. (Busby and Chitre 2018).

Emergency laparotomy has one of the highest mortality rates of any surgical procedure, almost tenfold that of major elective gastrointestinal surgery (Pearse RM et al. 2012). Despite this, emergency perioperative treatment often falls short of the clinical standards, organisational structures, and care processes that benefit most elective patients.

Diagram, timeline

Description automatically generated

Figure 1: Path of emergency laparotomy patient, source: NELA 2020

### 1.1.2 NELA

The National Emergency Laparotomy Audit (NELA) is a feature of the Healthcare Quality Improvement Partnership’s (HQIP) National Clinical Audit and Patient Outcomes Programme (NCAPOP).

In response to the HQIP’s call for new national audit topic ideas in 2011, NELA was one of the top two (of eleven) national clinical audits chosen for immediate financing. It was commissioned after evidence of a high rate of death among patients receiving emergency laparotomies in hospitals across England and Wales, as well as considerable heterogeneity in treatment and mortality.

The audit aims to enhance the quality of treatment for patients having an emergency laparotomy by collecting high-quality comparative data from all emergency laparotomy providers.

On December 1st, 2012, the NELA project officially commenced. NELA includes both organisational and patient audits to examine the structure, process, and outcome metrics for the quality of care obtained by patients undergoing emergency laparotomy (NELA [no date]).

Since the introduction of NELA in 2013, the 30-day mortality rate has decreased from 11.8 to 9.3 per cent. The decrease has been accomplished by clinicians increasing their detection and documentation of high-risk patients, which is a necessary first step towards ensuring they get the required levels of care (NELA 2020).

A picture containing graphical user interface

Description automatically generated

Figure 2: The overall unadjusted 30-day and 90-day ONS mortality rates trend by NELA dataset year, Source: NELA 2020.

The Portsmouth-physiologic and operative severity score for the enumeration of mortality and morbidity (P-POSSUM) was commonly used as a scoring tool to identify patients at increased risk of 30-day mortality. P-POSSUM was designed to have a wide application across the general surgical spectrum, both in elective and emergency settings (Copeland 2002). NELA recognised a need for a specific scoring tool intended for emergency laparotomy. NELA developed a mortality risk prediction tool specific to emergency laparotomy patients based on similar variables used in the P-POSSUM tool.

N. Eugene et al. 2018 developed and internally validated the NELA model. They compared the NELA models to five other models in terms of discrimination, both in their original development datasets and in the NELA dataset. In the NELA dataset, both the P-POSSUM and Surgical Outcome Risk Tool (SORT) models had a C-statistic of 0.81. The Biochemistry and Haematology Outcome Models (BHOM) had the poorest discrimination, with a C-statistic of less than 0.6. The highest discrimination on the NELA dataset was the NELA model with a C-statistic of 0.86. Barazanchi et al. 2020 also demonstrated the NELA tool to be most predictive of mortality after emergency laparotomy, obtaining a C-statistic of 0.83.

NELA has been a success at improving the quality of care, as evident by reducing mortality rates. However, little research has been conducted into morbidity associated with an emergency laparotomy, perhaps because it is subjective and complex to define. Currently, NELA is contractually obligated to minimise the burden of data input on clinicians, and the registry’s outcomes dataset is limited to short-term results (30 days). NELA records information regarding early postoperative morbidity and the effect of emergency laparotomy on a patient’s longer-term functional capacity and quality of life with poor fidelity, if at all (Saunders et al. 2021).

Clinical perception is that morbidity is high following emergency laparotomy, although this is an under-researched area. Local investigators at Cardiff and Vale UHB sought to establish the morbidity of emergency laparotomy with a postal questionnaire. Free text responses from patients demonstrated a significant effect on patients’ quality of life, with one patient stating: “… I didn’t answer straight away because I was so angry about being left like this, I am grateful for saving my life but to leave me like this is unacceptable”. This quote highlights the motivation for more research into the quality of life following the surgery.

Tolstrup et al. 2018 also recognised that long-term outcomes following the surgery are poorly investigated. Tolstrup et al. 2018 sought to characterise the frequency of chronic postsurgical pain (CPSP), pain-related functional impairment, and gastrointestinal quality of life (QoL) following emergency laparotomy. All patients who underwent any type of gastrointestinal emergency midline laparotomy at Copenhagen University Hospital Herlev were included from May 2009 to May 2013 and June 2014 to November 2015. The eligible study population was contacted by phone to see if they would participate in the survey, and then a questionnaire was mailed. They achieved an excellent response rate of 73%. They learned that CPSP and poor gastrointestinal quality of life were frequent complications of emergency laparotomy. CPSP was associated with younger age and acute postsurgical pain. However, in this study, the strongest independent predictor of CPSP was moderate-severe acute postsurgical pain, with five times the risk compared to patients experiencing mild pain. The main limitation of this study is that patients were asked to describe postsurgical pain they experienced several years ago. Patients may not accurately remember their experience post-discharge, and perhaps subsequent events following discharge may skew their memory. Therefore, the study suffers from recall bias. CPSP’s mechanism of action is not entirely known and promoted for more research into the physical and mental outcomes following the surgery.

The Emergency Laparotomy Follow-Up Study (ELFUS**)** established by Saunders et al. 2021 aimed to record postoperative morbidity in patients undergoing emergency laparotomy and examine the connection between morbidity and subsequent quality of life. They aimed to assess the feasibility of collecting accurate, longitudinal QoL data and morbidity in the setting of emergency laparotomy during the first year of recovery to ultimately provoke more research into a larger-scale future study. This is the first longitudinal study to describe Patient-reported outcomes measures (PROMs) and QoL up to a year following emergency general abdominal surgery. The EQ5D and World Health Organization Disability Assessment Schedule (WHODAS) 2.0 questionnaires were used to collect patient-reported outcomes at 1-, 3-, 6-, and 12-months following surgery. The study achieved excellent response rates, with 100% of patients answering a baseline questionnaire and 70% answering a 12-month follow-up QoL questionnaire. The ELFUS study demonstrates that it is feasible to collect accurate QoL and complication data from emergency laparotomy patients. They reinforced the point that early collection of inpatient morbidity is key to good outcome reporting of postoperative morbidity.

PROMs are one method for assessing the efficacy of interventions and determining the value of resources invested. PROMs are self-reported questionnaires that patients complete to document their health at particular periods to identify health changes over time. They are multidimensional assessments that may include symptoms, functional status, and health-related quality of life. A commonly used generic health questionnaire is The European Quality of Life Five Dimension, or EQ-5D.

### 1.1.3 The EQ-5D

The European Quality of Life Five Dimension is a health questionnaire introduced by the EuroQol group in 2009. The EQ-5D is a brief questionnaire intended to get a broad, ‘generic’ picture of patient-reported health. Its strength is in its brevity and its capacity to compare patient health across patients, illnesses, and treatments. Since its inception almost three decades ago, it has grown to be the most frequently used Patient-Reported Outcomes questionnaire globally, being utilised in population health surveys, clinical research, and routine outcomes assessment in healthcare systems (Devlin and Richard Brooks 2017). The questionnaire consists of 2 pages: the EQ-5D descriptive system and the EQ visual analogue scale (EQ-VAS). The five aspects of the descriptive system are mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each dimension has five levels: no problems, slight problems, moderate problems, severe problems, and extreme problems; this can be encoded as a one to five-digit ordinal score. The patient is asked to tick the box next to the most relevant statement in each of the five dimensions to reflect their health status. This selection yields a one-digit number indicating the level chosen for that dimension, ranging from 0-5, with the highest disease severity indicated by 5.

The EQ-VAS is a question that captures the patient’s self-rated health. The patient is asked to rate their health on a vertical visual analogue scale with endpoints labelled “The best health you can imagine” (100) and “The worst health you can imagine” (0). The VAS is a quantitative health outcome measure that reflects the patients’ subjective judgement (EQ-5D-5L | About. 2017).

A study by Kwong et al. 2018 aimed to report on the follow-up response rate for patients, identify any response biases and explore the feasibility of comparing patients’ outcomes at three months with their retrospectively collected PROMs at baseline. In 11 hospitals, patients undergoing emergency laparotomy were recruited to complete a retrospective questionnaire containing the EQ-5D-3L and the Gastrointestinal Quality of Life Index (GIQLI). The study had a response rate of 74% using mailed follow-up, demonstrating that PROMs could be successfully collected in patients three months after emergency laparotomy. Additionally, they reported that most patients regained their prior level of gastrointestinal health and that their overall health improved. As measured by the EQ-5D score, the improvement in general health status may reflect that emergency laparotomies are performed primarily in life-threatening situations; the improved health outcomes would imply that these procedures are both life-saving and restorative as improving patients’ quality of life. This may be since a significant proportion of emergency laparotomies are performed for conditions associated with chronic symptoms before the acute presentation (such as inflammatory bowel disease or acute colonic perforation in diverticular disease). As such, recall of symptoms in the month prior to surgery may include the effect of chronic disease. The study had a limitation in that PROMs have not been widely used in emergency surgery, and their psychometric properties in such patients have not been established. Further research is required to develop PROMs for use in emergency admissions in a systematic manner, including psychometric testing for use in emergency laparotomy.

Veras et al. 2016 used the EQ-5D to assess the health-related quality of life of elderly patients with various conditions, and it was found to be a reliable instrument for evaluating the specific conditions presented. According to a 2011 study by Noyes and Edwards, the EQ-5D can be used to assess children’s health-related quality of life. The study proposed that the EQ-5D in children be used in conjunction with the child-specific quality of life measures and disease-specific measures to produce more reliable results.

### 1.1.4 Referral to psychological therapy

A study conducted by A Powell-Chandler et al. 2020 motivates the research into the psychological outcomes and referral to psychological therapy following colorectal surgery. Approximately a quarter of patients appear to experience psychological distress following colorectal surgery. There may be a need for increased awareness, preventative strategies, and referral to psychological therapies. There is a lack of information regarding the psychological consequences of colorectal resections. The stress associated with this surgery may result in anxiety, depression, or post-traumatic stress disorder (PTSD). Vulnerability, loss of life, and a perceived lack of support are risk factors for developing these psychological conditions. Patients who have a stoma due to any condition are at risk for anxiety and post-traumatic stress disorder (PTSD) as they adjust to body image issues, sexuality, and faeces. They reported that only younger age and female gender were significant risk variables via multiple linear regression to have poorer psychological outcomes. Any patients identified, in the study, as potentially having anxiety, depression or PTSD, their GP was informed, and they were referred for psychological therapy.

# 2 Data

All data analysis was performed using R, version 4.

## 2.1 Data source

All patients who undergo an emergency laparotomy in England and Wales are entered prospectively into a local NELA database. Cardiff and Vale University Hospital identified all patients who underwent an emergency laparotomy from Jan 2016 to December 2019 via their local NELA database. Additional process outcomes (beyond the 30 days of NELA) were retrospectively collected from patient medical records. All patients who were still alive at the time of the study were approached by post to complete an EQ5D questionnaire. If no response was obtained six weeks from postage, a telephone was made to complete the questionnaire via telephone. This project was registered as a Service Evaluation at the Cardiff and Vale University Health Board. All data was stored confidentially onto an excel spreadsheet and inputted by one of two authors of the clinical team. The excel spreadsheet was read into R for analysis.

## 2.2 Data description

There were a total of 1054 adults included in the study. A total of 262 patients answered all EQ-5D questions. Deceased, patients with dementia and patients who do not speak English were excluded from the EQ-5D study; a complete participant flow diagram can be found in appendix C. Responders were more likely to be more affluent based upon having a higher mean Welsh Index of Multiple Deprivation (WIMD) (t-test, p=0.000) and have a higher Charlson score (t-test, p= 0.030).

Chart, bar chart

Description automatically generated

Figure 3: Visualisation of EQ-5D target variables.

Diagram

Description automatically generated

Figure 4: Box and whisker plot showing the WIMD 2019 rank of EQ-5D responders and the rest of the patients.

Table 1: Showing the demographics between the EQ-5D patients and the remaining patients

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | **EQ-5D Responders (n=262)** | **Remaining Patients**  **(n=792)** |
| **Age** | | Mean | 62.72 | 62.13 |
| Standard deviation | 14.64 | 18.40 |
| **Male** | | Proportion | 0.41 | 0.48 |
| **ASA** | 1 | Proportion | 0.10 | 0.90 |
| 2 | 0.44 | 0.30 |
| 3 | 0.35 | 0.40 |
| 4 | 0.11 | 0.18 |
| 5 | 0.00 | 0.03 |
| **BMI** | | Q1 | 23.35 | 23.00 |
| Mean | 26.22 | 26.80 |
| Q3 | 31.05 | 31.88 |
| Proportion Missing | 0.27 | 0.38 |
| **WIMD 2019 Rank** | | Q1 | 556.50 | 345.00 |
| Mean | 1322.00 | 863.00 |
| Q3 | 1673.00 | 1573.50 |
| **New Cancer Diagnosis** | No | Proportion | 0.76 | 0.65 |
| Yes | 0.14 | 0.12 |
| Pre-existing | 0.05 | 0.05 |
| Missing values | 0.05 | 0.19 |
| **New IBD Diagnosis** | No | Proportion | 0.86 | 0.74 |
| Yes | 0.5 | 0.02 |
| Pre-existing | 0.05 | 0.05 |
| Missing values | 0.05 | 0.19 |

A total of 46 patients were referred to a mental health specialist, with 69% of those referred having a history of mental illness prior to having an emergency laparotomy. A total of 19 patients were referred to a counselling specialist, with 47% having a prior history of mental illness.

Chart, bar chart

Description automatically generated

Figure 5: Visualisation of the referral to mental health and counselling columns.

## 2.3 Data pre-processing

**2.3.1 Missing values**

Missing data was prevalent in the dataset, with many ways in which the missingness was represented. It was essential to understand why the values were missing and determine the appropriate strategies for rectifying the missingness. A subset of the data was made to contain the 262 EQ-5D responders. The project aims to build a prediction tool that can give information to the patient as they are discharged, so data or columns collected after discharge were removed from the subset.

Imputing missing values in columns with more than 5% missing values could introduce bias into the data (Alice 2015). Therefore, columns with more than five per cent missing were removed from the subset. Multiple imputation is a preferred technique over single imputation (Jakobsen 2017). Multiple imputation attempts to introduce variability into imputed data in order to generate a range of possible responses from which to work; it does so by creating several plausible imputed data sets and appropriately combining the results obtained from each of them (SurveyMethods 2011). Jakobsen 2017 recommends that only Missing At Random (MAR) missing values should be replaced via multiple imputation. Missing Completely At Random (MCAR) and Missing Not At Random (MNAR) missing values should not be imputed and removed from the data. It was determined that columns with less than five per cent of missing data were not MAR. The rows containing the missing values were removed.

For the referral to a mental health specialist and referral to a counselling specialist data subsets, the procedure mentioned above was followed. Columns with more than five per cent missing were removed from the subset. No MAR columns were contained in the subset, so the rows containing the missing values were removed.

**2.3.2 Class imbalance**

The EQ-5D target columns were imbalanced; as illustrated in figure 6, we can see that the class distribution of the Self.care column is skewed. The class distribution for all EQ-5D target variables can be found in figure 3. The referral to a mental health specialist and referral to a counselling specialist were also severely imbalanced, as seen in figure 5.

**Chart

Description automatically generated**

Figure 6: Showing class imbalance for target column “Self.care”

There are many possible reasons for the imbalance shown in figure 6; one reason could be that a patient is less likely to fill in a questionnaire if their physical health is very poor. Since this is a severe imbalance, we cannot treat this as a typical classification problem, and oversampling or undersampling or a combination of the two should be utilised. Oversampling via SMOTE and undersampling via RandomUnderSampler was used to correct the imbalance. SMOTE works by synthesising samples from the minor class rather than creating duplicates. The algorithm compares two or more similar instances (using a distance measure) and modifies each instance’s attributes randomly within the difference between the neighbouring instances (Brownlee 2015).

Chart, icon

Description automatically generated

Figure 7: Showing class distribution after oversampling then undersampling the data.

Nitesh V. Chawla et al. 2002 demonstrated that combining the technique of over-sampling the minority class and undersampling the majority class results in superior classifier performance than just under-sampling the majority class. Therefore, the training data was undersampled via RandomUnderSampler after being oversampled via SMOTE. RandomUnderSampler randomly selects and removes samples from the majority class, thus decreasing the number of instances in the transformed data that belong to the majority class (Pykes 2020). Data leakage is where information of the hold-out test set leaks into the dataset used to train the model (Brownlee 2020). To avoid this, the oversampling and undersampling was conducted via scikit-learn’s pipeline class, as the pipeline fits the transformation on the training part of each CV- split only.

**2.3.3 Feature scaling**

Feature scaling is a critical pre-processing step that is required to standardise or normalise the input data. When one of the columns has a very high value compared to the others, the impact of the column with the higher value will be much more significant than the impact of the other low valued columns. Even if they are more critical in determining the output, features with a large magnitude will weigh significantly more than features with a small magnitude. As a result, the prediction may fail to produce the desired results. There are two methods for scaling features that are frequently used: normalisation and standardisation. We chose to standardise the input data because it is less susceptible to outlier values. The mean of the features was shifted to zero, and the standard deviation was shifted to one (Gupta 2020). The feature scaling was applied to the training data via the scikit-learn pipeline, which avoids data leakage.

# 3 Methods

All model development was performed using Python, version 3.

## 3.1 Multiclass classification

Multiclass classification refers to classification tasks that have more than two class labels. Two models were built for this problem: an artificial neural network and a multinomial logistic regression. The models were built to predict one of the class values of the target variables, the EQ-5D categorical variables: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression.

**3.1.1 Neural network**

Deep learning algorithms are based on extracting representations at a high degree of abstraction from vast amounts of input data. Deep learning aims to replicate the human brain by enabling systems to cluster data and make predictions. Deep learning’s central idea is a feedforward neural network in which multiple non-linear hidden layers represent the degrees of abstraction. Deep learning is defined as having more than three layers, including input and output. Thus, a deep neural network’s fundamental structure consists of an input layer, multiple hidden layers, and an output layer. The input layer comprises a collection of nodes that represent the input characteristics. Each node in the hidden layer applies a weighted sum to the values in the bottom layer. The total is then supplied to the activation function of a node in order to determine the layer’s output values. Finally, the output layer accepts values from the previously hidden layer and converts them to output values (IBM Cloud Education 2020) (Brownlee 2016).

Diagram

Description automatically generated

Figure 8: Architecture of the deep neural network for the multiclass classification.

Text

Description automatically generated

Figure 9: Deep neural network for the multiclass classification model.

In this instance, we present a deep, fully connected neural network with two hidden layers, as seen in figure 9. The input layer’s “input\_dim” parameter is set to the same value as the number of features being used, in this case, 77. In the two hidden layers, the “RELU” (Rectified Linear Units) function is utilised. RELU is a technique that is often used in neural networks. Dummy variables were used to encode the features, which supports the usage of the RELU function. RELU is defined as follows:

(1)

If the input value is higher than zero, the output value equals the input value (Brownlee 2019).

In the output layer, the “Softmax” function is used. The Softmax function normalises the outputs, converting them from weighted sum values into probabilities that sum to one. The following is the definition of the Softmax function:

(2)

Where  represents the ith element of the input to softmax, which corresponds to the class i and k is the number of classes. The model was compiled with the loss parameter set to “sparse\_categorical\_crossentropy”, as multiple class labels exist.

Hyperparameter tuning was applied to the “batch\_size” and “epoch” parameters. The batch\_size hyperparameter specifies the number of samples to process before updating the model’s internal parameters, and the considered values were 10 and 100. The number of epochs is a hyperparameter that specifies how many complete passes the model will make through the training dataset, and the considered values were 50, 100 and 200. More hyperparameter tuning to a more extensive set of values and other parameters might have been performed; however, computational power was limited.

**3.1.2 Multinomial logistic regression**

By default, logistic regression cannot solve multiclass classification problems. To enable the prediction of multiple class labels, the logistic regression model must be modified. Changing from binomial to multinomial probability in logistic regression requires modifying the loss function used to train the model (e.g. log loss to cross-entropy loss) and changing the output from a single probability value to one probability for each class label. The LogisticRegression class, from scikit-learn, can be set up for multinomial logistic regression by setting the “multi\_class” argument to “multinomial” and the “solver” argument to a solver that supports multinomial logistic regression, such as “lbfgs”. Logistic regression does not directly support imbalanced classification, and hence the logistic regression model must be adjusted to account for the skewed class distribution. The logistic regression, by default, has all errors in each class having equal weighting. These weightings may be changed according to the relative significance of each class. The “class\_weight” parameter can provide the class weights to the model. The parameter was set to “balanced”, which uses best-practice heuristics to calculate the class weights (Brownlee 2020). The model is illustrated below in figure 10.

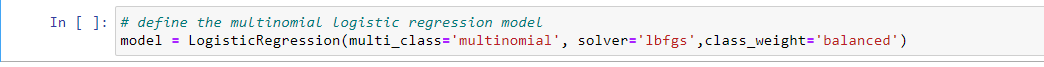


Figure 10: Multinomial logistic regression model

Hyperparameter tuning was applied to the penalty weighing values via the “C” parameter. C is the regularisation parameter. When C is small, this increases the regularisation strength, resulting in simple models that underfit the data. For large values of C, this reduces the power of regularisation, which allows the model to become more complex and thus overfit the data (Shivers 2012). The considered values were: 0.0000, 0.0001, 0.0010, 0.0100, 0.1000 and 1.0000. The best hyperparameter was chosen based on obtaining the highest average balanced accuracy.

Chart, box and whisker chart

Description automatically generated

Figure 11: Box and whisker plot showing the average (over all k-folds) balanced accuracy for different hyperparameters.

## 3.2 Regression

Regression analysis refers to a set of statistical techniques for estimating the relationships between a dependent variable and one or more independent variables (Corporate Finance Institute [no date]). Three models were built for this problem: linear regression, decision trees and random forest. The models were built to predict the VAS score, a continuous variable.

### 3.2.1 Linear regression

Linear regression is a well-known technique in machine learning and statistics; a single output y can be calculated from a linear combination of the input variables (x). Our model is a linear model fitted by minimising a regularised empirical loss with SGD. SGD stands for Stochastic Gradient Descent: the loss gradient is estimated for each sample when the model is updated along the way with a decreasing strength schedule (aka learning rate) (Brownlee 2016). The scikit-learn model was built using default parameters but with hyperparameter tuning on the “max\_iter” parameter, with values 500, 1000, 3000 and 5000 considered. The max\_iter parameter specifies the maximum number of epochs, which indicates how many times the learning algorithm will work through the entire training dataset (Brownlee 2018).

### 3.2.2 Decision trees

A decision tree is a supervised learning algorithm. Decision trees are constructed from a set of nodes, each of which represents a distinct feature. Generally, the first node of a decision tree is referred to as the root node. The depth of a tree is defined as the total number of levels in the tree, exclusive of the root node. A branch represents a decision and can be visualised as a connection between different nodes.

Diagram, schematic

Description automatically generated

Figure 12: Decision tree diagram, source: [Chauhan 2020](https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html)

Decision trees incrementally divide large data sets into smaller data groups until they reach a point where a label can describe them; at that point, the node becomes terminal. Decision trees use multiple algorithms to split a node into two or more sub-nodes (Raj 2020).

The scikit-learn model was built using default parameters but with hyperparameter tuning on the “max\_depth” parameter, with values from one to ten considered in increments of one. The max\_depth parameter controls the maximum distance between the root node and any terminal node(Scikit-Learn [no date]).

### 3.2.3 Random forest

Random forest is a supervised learning algorithm that aggregates the result of many decision trees, using ensemble approaches (bagging), and then outputs the most optimal result, with no interaction between individual trees. The random forest algorithm solves the problems associated with decision trees. Decision trees are prone to overfitting, primarily when used on datasets with many features. Training a large number of decision trees and then calculating the mean/mode of prediction for all trees helps prevent overfitting. Also, decision trees often seek locally optimum solutions rather than globally optimal ones (Raj 2020).

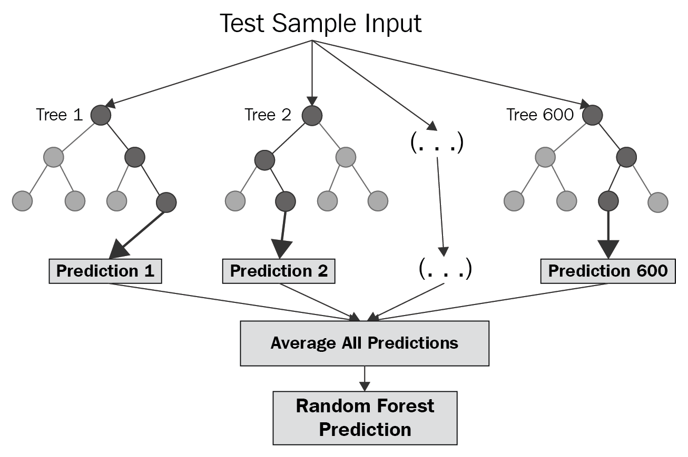


Figure 13: Random forest technique, source: [Chakure 2019](https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f)

Our model was constructed below, with hyperparameter tuning on the “n\_estimators” parameter, with values 100, 250, 400, 500, 1000 and 2000 considered. The n\_estimators parameter controls the number of trees in the forest (Scikit-Learn [no date]).

## 3.3 Binary classification

Binary classification refers to classification tasks that have two class labels. In our problem, we are trying to predict whether a patient is referred to a mental health or counselling specialist by their GP due to the surgery. Three models were built for this problem: K Nearest Neighbors (KNN), logistic regression and Support Vector Machine (SVM).

### 

### 3.3.1 KNN

KNN is a supervised learning method that may be used for regression and classification. By computing the distance between the test data and all the training points, KNN attempts to predict the suitable class for the test data, and then selects the K spots that are closest to the test data. The KNN method evaluates the likelihood that the test data belongs to each of the ‘K’ training data classes, and the class with the most significant probability is chosen (Christopher 2021).

Chart, map, scatter chart

Description automatically generated

Figure 14: KNN algorithm (K=1 and K=100), source: [DataSklr 2020](https://cf-my.sharepoint.com/personal/millnsbc_cardiff_ac_uk/Documents/Documents/MSc%20Data%20Science%20and%20Analytics/Dissertation/FINAL/:%20https:/www.datasklr.com/select-classification-methods/k-nearest-neighbors)

The model was created using the scikit-learn library using default parameters but with hyperparameter tuning applied to the “n\_neighbors” and “weights” parameters. n\_neighbors specifies the maximum number of neighbours to use when performing queries, as illustrated in figure 14, where K=1 and K=100 are used. For the value of n\_neighbors, values between 1 and 31 were considered in increments of 1. Two weight values were considered: “uniform” and “distance”. Uniform assigns equal weight/ dominance of all neighbours, and distance assigns weight to points based on their inverse distance. In this case, neighbours closer to a query point have a more significant influence than those further away.

### 3.3.2 Logistic regression

Binomial logistic regression measures the relationship between the categorical dependent variable and explanatory variables (features) by estimating probabilities using a logistic (sigmoid) function. The positive class or outcome is mapped to one, while the negative class or outcome is mapped to zero. The fitted model estimates how likely an example is to belong to class one. As with the multinomial logistic regression, the “class\_weight” parameter equals “balanced” since our dataset is imbalanced. Using scikit-learn, the model can be seen below in figure 15.

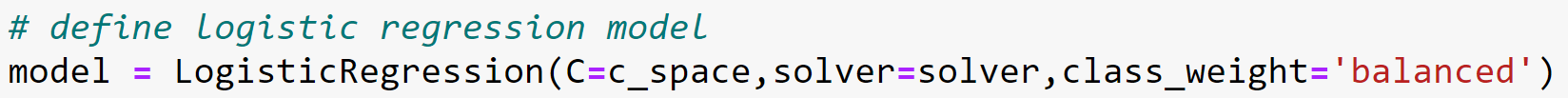


Figure 15: Logistic regression for binary classification.

Hyperparameter tuning was applied to the “solver” parameter and the “C” parameter. C is the regularisation parameter, the same as the multinomial logistic regression. The considered values of the C parameter were 15 values between -5 and 8, which were generated with the np.logspace function. The considered values for the solver parameter were: “newton-cg”, “lbfgs”, “sag”, “saga” and “liblinear”. Each solver attempts to minimise a cost function by determining the parameter weights (Jeff Hale 2019). A detailed explanation of each solver value can be found [here](https://stackoverflow.com/questions/38640109/logistic-regression-python-solvers-definitions).

### 3.3.3 SVM

The SVM algorithm tries to find a hyperplane that maximises the separation of the data points to their potential classes in n-dimensional space; see an illustration below in figure 16.

A picture containing website

Description automatically generated

Figure 16: SVM algorithm visualisation, source: [baeldung 2021](https://www.baeldung.com/cs/svm-multiclass-classification)

The model was created using the scikit-learn library using default parameters but with hyperparameter tuning applied to the “C” and “gamma” parameters. C is the regularisation parameter, the considered values were: 0.1, 1, 10, 100 and 1000. The gamma parameter indicates the extent to which a single training example has an influence, with low values indicating “far” and high values indicating “close.” The gamma parameters can be considered the inverse of the radius of influence of the support vectors selected by the model (Scikit-Learn [no date]). The following values were considered for the gamma parameter: 0.001, 0.01, 0.1 and 1.

# 4 Results

Due to the stochastic nature of the algorithms, each model was run several times with a different random seed before an average result was obtained.

## 4.1 K-fold cross-validation

All models were evaluated using k-fold cross-validation. K-fold cross-validation is the gold standard for determining how well machine learning models perform in applied machine learning. It is used to mitigate against overfitting in prediction models, mainly when the quantity of data available is restricted. We performed cross-validation on our dataset by partitioning it into a given number of folds, analysing each fold, and averaging the total error estimate. In machine learning, a value of k equal to five or ten is very typical (Goyal 2021). Five was chosen as the value of k. This balances the number of folds required to sufficiently smooth the results and the processing time required to run each model five times effectively. Because mobility=5 and selfcare=5 had fewer than five class values, the number of splits in the cross-validation folds in those models was decreased to four and three, respectively. To determine the best model with the best hyperparameter, all models were evaluated using k-fold cross-validation. Each model’s different hyperparameters were evaluated across all K folds. The hyperparameter with the best average score over the k folds would be chosen as the best for that model. This model is then compared to all other models for that task, and the model with the highest average score across all k folds is determined to be the best.

Diagram

Description automatically generated

Figure 17: The process of k-fold cross-validation (k=5), source: Goyal 2021

## 4.2 Dimensionality reduction

After dummying the features, the feature count increased to approximately 330. The data contained many features compared to the number of rows, which can lead to the curse of dimensionality. “As the number of features or dimensions grows, the amount of data we need to generalise accurately grows exponentially.” - Charles Isbell, Georgia Tech.

The curse of dimensionality states that the error increases as the number of features increases. It refers to the fact that high-dimensional algorithms are more difficult to design and frequently have a running time exponential in the dimensions. While adding dimensions theoretically allows for the storage of more information, in practice, it rarely helps due to the increased possibility of noise and redundancy in real-world data (Choudhury 2019).

The decision was made to reduce the feature count to 23. The feature selection decision was based on domain knowledge; appendix D contains a list of features. This was a necessary step toward dimensionality reduction. When the features were dummied, the number of features increased to 77 for the multiclass classification and regression problems and 96 for the binary classification problem, necessitating further dimensionality reduction.

Principal Component Analysis, or PCA, is a dimensionality reduction technique in which highly correlated data is transformed into a set of uncorrelated components referred to as principal components. The principle components with a lower dimension capture the majority of the information in the high-dimensional dataset (Great Learning Team 2020). PCA was added into the scikit-learn pipeline, with default parameters. In addition, k-fold cross-validation reduces the effect of high dimensionality (Spruyt 2014). The complete scikit-learn pipeline is shown below in figure 18.

****

Figure 18: Scikit-learn pipeline.

## 4.3 Multiclass classification

Both models were evaluated using the balanced accuracy metric. Average accuracy provides an overall measure of how well the model predicts the whole collection of data correctly. Accuracy is not a reliable metric to evaluate models with imbalanced data sets and often misleads one into thinking the model has excellent performance when, in reality, it does not (Grandini et al. 2020).

Balanced accuracy overcomes the shortcomings of average accuracy and can deal with imbalanced datasets (Brodersen et al. 2010). The balanced accuracy formula is simply an average of recalls. Recall indicates how well the model accurately identifies true positives. The recall for each class is calculated, and then the average is taken across all classes to get the balanced accuracy score. When the dataset is balanced, accuracy and balanced accuracy tend to converge to the same result. However, balanced accuracy is insensitive to imbalanced class distribution because each class is equally weighted and significant (Grandini et al. 2020).

The model with the highest balanced accuracy for each EQ-5D target variable would be chosen as the best model. Table 2 shows the model results for each target variable.

Table 2: Average results over all K-folds for all EQ-5D target variables



The multinomial logistic regression model outperforms the neural network predicting pain/discomfort and usual activities. The neural network performs better at predicting the other target variables. The overall performance of the models was poor. The balanced accuracy of the training and test sets for the best models were then explored. For this, the data was split into training and test data, a 70:30 split. The balanced accuracy for each model is shown in Table 3. All models obtain a much higher balanced accuracy for the training set than the test set, which suggests overfitting. The neural network results indicate that there is a greater degree of overfitting.

Table 3: Balanced accuracy score for the training and test data set for all EQ-5D target variables.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Neural Network** | | | **Multinomial Logistic Regression** | |
| **Mobility** | **Anxiety/ Depression** | **Self-Care** | **Pain/ Discomfort** | **Usual Activities** |
| **Train** | 0.908 | 0.953 | 0.655 | 0.552 | 0.538 |
| **Test** | 0.250 | 0.188 | 0.190 | 0.338 | 0.292 |

## 4.4 Regression

Three models were built to predict the VAS: linear regression with stochastic gradient descent, decision tree and random forest. The models were evaluated using the root mean square error (RMSE). RMSE is formally defined by:

(3)

Where  are the predicted values,  are the observed values, and  is the number of observations. RMSE is the standard deviation of the residuals (prediction errors). Residuals measure how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it indicates the degree to which the data is concentrated around the best fit line. RMSE, when compared to MAE, has the advantage of penalising significant errors (Wesner 2016). The best model would be the model that achieved the lowest RMSE value. Table 4 shows the RMSE obtained from each model.

Table 4: Average results over all K-folds for the regression problem

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Linear Regression** | **Decision Tree** | **Random Forest** |
| **RSME- Mean (Standard deviation)** | 24.96 (10.25) | 22.32 (10.77) | 21.64 (9.89) |

The random forest regressor obtains the lowest RMSE value, and so it is determined to be the best model. A residual plot is typically used to find problems with a regression. For this, the data was split into training and test data, a 70:30 split. Figure 19 shows a residual plot; we can observe that the train r squared is significantly higher than the test r squared, which suggests that the model is overfitting.

Chart, scatter chart

Description automatically generated

Figure 19: Residuals plot for random forest

## 4.5 Binary classification

All models were evaluated using the AUC metric, the same as the C-statistic metric for binary outcomes. To define the Area Under the Curve, we first define the Receiver Operating Characteristics curve (ROC). The ROC curve is a graphical plot that shows how a binary classifier’s diagnostic ability changes as its discrimination threshold varies. The ROC curve is constructed by comparing the true positive rate (recall) and false positive rate (FPR) at different threshold values. AUC provides a “summary” for the performance of the ROC curves. The higher the AUC score, the better a classifier performs for the given task. Huang and Ling 2005 showed theoretically and empirically that AUC is a better measure than accuracy and provides an effective method to evaluate an imbalanced dataset.

The model with the highest AUC value would be chosen as the best model for referral to mental health and then referral to counselling. The binary classification results are shown in Table 5.

Table 5: Average results over all K-folds for binary classification problem.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Mental Health** | | | **Counselling** | | |
| **KNN** | **Logistic Regression** | **SVM** | **KNN** | **Logistic Regression** | **SVM** |
| **ROC AUC- Mean (Standard Deviation)** | 0.626 (0.057) | 0.691 (0.072) | 0.651 (0.091) | 0.587 (0.112) | 0.661 (0.156) | 0.660 (0.136) |

Logistic regression obtained the highest average AUC for predicting a referral to a mental health and counselling specialist over all K folds. The AUC values for training and test were then compared to look for evidence of overfitting. Table 6 shows the AUC scores on the training and test data sets.

Table 6: AUC score for the training and test data set for mental health and counselling for the logistic regression model (best model).

|  |  |  |
| --- | --- | --- |
|  | **Mental Health** | **Counselling** |
| **Train** | 0.638 | 0.729 |
| **Test** | 0.581 | 0.641 |

Table 9 does not suggest the model is overfitting, as both training and test sets obtain similar AUC scores. The precision and recall measures for predicting each class were then explored (class 1 = referral to mental health/ counselling) for the logistic regression model. A definition of precision and recall can be found in appendix B. The precision and recall for both classes and both problems can be found in table 7.

Table 7: Precision and recall for both referral to mental health and counselling for both classes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Mental Health** | | **Counselling** | |
| **Class 0** | **Class 1** | **Class 0** | **Class 1** |
| **Precision** | 0.958 | 0.138 | 0.986 | 0.061 |
| **Recall** | 0.891 | 0.308 | 0.870 | 0.400 |

Precision and recall are both low in both problems for class 1 but high for class 0. This shows that the model is constantly classifying the class as class 0.

# 5 Discussion

This study is one of the first to aim to predict the life-changing effects of an emergency laparotomy on its survivors. The purpose of this study was to develop models that the information produced by the models whose output could be used to advise patients about the difficulties they may face during their recovery to allow early signposting to relevant support and motivate additional research and data collection on morbidity following surgery.

## 5.1 Multiclass classification

We proposed two multiclass classification models: a deep neural network and a multinomial logistic regression, to predict the score of one of five EQ-5D parameters. The model performed poorly on this problem, with an average balanced accuracy of 0.25 across the five dimensions. As a result, this model is not recommended for use in a clinical setting. It is possible that the data lacked sufficient predictive power to produce a good result. Table 3 suggests that the models are overfitting the data. This could be because the training and test sets have a distributional mismatch. Oversampling and undersampling were applied to the training data, whereas the testing data still had a class imbalance. In all five of the EQ-5D target variables, the class value equal to five was the minority class, recommending that future work attempt to equalise the distribution. A shorter time between discharge and receiving the questionnaire and determining how to represent patients who are in such poor health that they cannot complete a questionnaire can all contribute to improving the class distribution. Statistically based feature selection and more dimensionality reduction can also help reduce overfitting. In addition, a stepwise selection technique could be utilised to continually add features to the model until the model performance does not improve. This may help with the overfitting. We were also trying to predict the EQ-5D without a baseline measurement obtained before the surgery. Without a baseline measure, it is impossible to differentiate between the impact of the operation on the patient and the EQ-5D changes induced by the surgery. This may have impacted the model’s performance.

## 5.2 Regression

Three regression models were proposed to predict the EQ-5D VAS measure: linear regression, decision trees, and random forest. The theory predicted that the random forest would be the best regressor, and indeed, the random forest had the lowest root mean square error (RMSE), 21.64 (9.89). RMSE is a relative measure and context-dependent on the problem, the data, and the dependent variable. Additional research should be conducted using the same data to predict the VAS score following emergency laparotomy and compare it to the result obtained in this study. The residual plot (figure 19) shows that the train r squared is significantly higher than the test r squared, suggesting that the model is overfitting. Feature selection and more dimensionality reduction can help reduce overfitting. A stepwise regression technique could be used to incrementally add features to the model until the model’s performance does not improve; this could help with overfitting.

## 5.3 Binary classification

Finally, three binary classification models were proposed to predict whether a patient will be referred for mental health or counselling therapy by their GP: KNN, logistic regression, and SVM. The logistic regression model achieved the highest average AUC measure of 0.691 (0.072) and 0.661 (0.156) for predicting mental health and counselling, respectively, over all K folds. Table 6 does not indicate that the model is overfitting, as both the training and test sets achieve comparable AUC values. In Table 7, we see that the logistic regression model has high precision and recall for class zero and a low precision and recall for class one for both problems. This shows that the model is classifying most patients as class zero due to the class imbalance. The model has adequate performance with moderate discrimination. The values of 0.69 and 0.66 for AUC are considered to have a reasonable discriminating ability when it comes to identifying patients who will be referred to a mental health or counselling specialist. The precision and recall values for class one could be improved with more class one samples. A hyperparameter search could be conducted on specific combinations of undersampling class zero and overweighting the loss on class one. This problem may benefit from a superior algorithm like XGBoost or LightGBM, capturing more information from the dataset.

## 5.4 Limitations and further work

There are several limitations of the study and models presented. Because no baseline EQ-5D score was obtained before surgery and different patients were asked to score postsurgical morbidity at varying lengths of time following surgery, we cannot conclusively state that the effect of the surgery is the cause of the score. In the future, a baseline EQ-5D score should be obtained in order to create a model for calculating the EQ-5D delta, with a fixed time for obtaining the postoperative EQ-5D score. Also, the questionnaire could be sent earlier to perhaps increase the response rate. Kwong et al. 2018 reported a 74.1 per cent response rate after sending the questionnaire after three months.

Selection bias is present, as the study excludes the deceased, patients with dementia, and patients who do not speak English. Further work must be conducted to determine how this critically ill population group may be included in emergency surgery research frameworks to guarantee broader generalisability. Additionally, there could be non-response bias if patients are in such poor health (EQ-5D = 5) that they are unable to complete the survey. Patients who completed the EQ-5D questionnaire may have been removed from the analysis since rows with missing values were removed from the data.

One of the main aims of this study was to motivate further research into postoperative morbidity. Although the model results for this study were inadequate and should not be used in a clinical setting, there is the possibility that meaningful results can emerge from this data. The proposed problems should be broken down into smaller projects, emphasising the interaction of the features and feature selection, this may increase model performance and yield results that can advise patients on their recovery. Additional hyperparameter tuning can be applied to the models, as computing power was limited.

# 6 References

A Powell-Chandler et al. 2020. *Psychological sequelae of colonic resections*. PISA Trial Management Group. Available at: <https://orca.cardiff.ac.uk/128688/1/PISA%20FINAL.pdf>.

Alice, M. 2015. Imputing Missing Data with R; MICE package. *Datascience+* October 4th. Available at: <https://datascienceplus.com/imputing-missing-data-with-r-mice-package/>.

Ashwin, R. 2020. Sowing the Seeds of Decision Tree Regression. *Towards Data Science* June 6th. Available at: <https://towardsdatascience.com/sowing-the-seeds-of-decision-tree-regression-2bb238dfd768>.

baeldung 2021. Multiclass Classification Using Support Vector Machines. August 25th. Available at: <https://www.baeldung.com/cs/svm-multiclass-classification>.

Barazanchi, A. et al. 2020. *Evaluating and Improving Current Risk Prediction Tools in Emergency Laparotomy*. ournal of Trauma and Acute Care Surgery. Available at: <https://www.researchgate.net/publication/340661975_Evaluating_and_Improving_Current_Risk_Prediction_Tools_in_Emergency_Laparotomy>.

Brodersen, K.H. et al. 2010. *The Balanced Accuracy and Its Posterior Distribution*. Department of Computer Science, ETH Zurich, Switzerland: Institute for Empirical Research in Economics, University of Zurich. Available at: <https://ieeexplore.ieee.org/document/5597285>.

Brownlee, J. 2015. 8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset. *Machine learning mastery* August 19th. Available at: <https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/>.

Brownlee, J. 2016a. Linear Regression for Machine Learning. *Machine learning mastery* March 25th. Available at: <https://machinelearningmastery.com/linear-regression-for-machine-learning/>.

Brownlee, J. 2016b. Multiclass Classification Tutorial with the Keras Deep Learning Library. *Machine learning mastery* June 2nd. Available at: <https://machinelearningmastery.com/multi-class-classification-tutorial-keras-deep-learning-library/?unapproved=625230&moderation-hash=c646d5ce769c4f95c57d5abdf09d426b#comment-625230>.

Brownlee, J. 2018. Difference Between a Batch and an Epoch in a Neural Network. *Machine learning mastery* July 20th. Available at: <https://machinelearningmastery.com/difference-between-a-batch-and-an-epoch/>.

Brownlee, J. 2019. A Gentle Introduction to the Rectified Linear Unit (ReLU). *Machine learning mastery* January 9th. Available at: <https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/>.

Brownlee, J. 2020a. Cost-Sensitive Logistic Regression for Imbalanced Classification. *Machine learning mastery* January 27th. Available at: [Cost-Sensitive Logistic Regression for Imbalanced Classification](https://doi.org/Cost-Sensitive%20Logistic%20Regression%20for%20Imbalanced%20Classification).

Brownlee, J. 2020b. How to Avoid Data Leakage When Performing Data Preparation. *Machine learning mastery* June 22nd. Available at: <https://machinelearningmastery.com/data-preparation-without-data-leakage/>.

Bruna de Veras et al. 2016. *Health-related quality of life in elderly: a review of the EQ-5D use*. Available at: <https://www.researchgate.net/publication/313814797_Health-related_quality_of_life_in_elderly_a_review_of_the_EQ-5D_use>.

Busby, C. and Chitre, V. 2018. What is an Emergency Laparotomy? Available at: <https://www.jpaget.nhs.uk/media/294350/NU-62-Emergency-Laparotomy-sheet.pdf>.

Chai, T. 2014. *Root mean square error (RMSE) or mean absolute error (MAE)?– Arguments against avoiding RMSE in the literature*. Available at: <https://www.researchgate.net/publication/272024186_Root_mean_square_error_RMSE_or_mean_absolute_error_MAE-_Arguments_against_avoiding_RMSE_in_the_literature#pf4>.

Chakure, A. 2019. Random Forest Regression. *Medium* June 29th. Available at: <https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f>.

Choudhury, A. 2019. Curse Of Dimensionality And What Beginners Should Do To Overcome It. *Analytics India Mag* May 22nd. Available at: <https://analyticsindiamag.com/curse-of-dimensionality-and-what-beginners-should-do-to-overcome-it/>.

Christopher, A. 2021. K-Nearest Neighbor. *Medium* February 2nd. Available at: <https://medium.com/swlh/k-nearest-neighbor-ca2593d7a3c>.

Copeland, G.P. 2002. *The POSSUM System of Surgical Audit*. Available at: <https://jamanetwork.com/journals/jamasurgery/fullarticle/212013>.

Corporate Finance Institute [no date]. Regression Analysis. Available at: <https://corporatefinanceinstitute.com/resources/knowledge/finance/regression-analysis/>.

D. I. Saunders et al. 2021. *Emergency Laparotomy Follow-Up Study (ELFUS): prospective feasibility investigation into postoperative complications and quality of life using patient-reported outcome measures up to a year after emergency laparotomy*. Perioperative Medicine. Available at: <https://perioperativemedicinejournal.biomedcentral.com/articles/10.1186/s13741-021-00193-5>.

DataSklr 2020. K-Nearest Neighbors. *DataSklr* January 28th. Available at: <https://www.datasklr.com/select-classification-methods/k-nearest-neighbors>.

Devlin, N.J. and Richard Brooks 2017. *EQ-5D and the EuroQol Group: Past, Present and Future*. Office of Health Economics. Available at: <https://pubmed.ncbi.nlm.nih.gov/28194657/>.

EQ-5D-5L | About. 2017. Available at: <https://euroqol.org/eq-5d-instruments/eq-5d-5l-about/>.

Goyal, C. 2021. Importance of Cross Validation: Are Evaluation Metrics enough? *Analytics Vidhya* May 21st. Available at: <https://www.analyticsvidhya.com/blog/2021/05/importance-of-cross-validation-are-evaluation-metrics-enough/>.

Grandini, M. et al. 2020. *METRICS FOR MULTICLASS CLASSIFICATION: AN OVERVIEW*. University of Bologna. Available at: <https://arxiv.org/pdf/2008.05756.pdf#:~:text=Accuracy%20is%20one%20of%20the,computed%20from%20the%20confusion%20matrix.&text=The%20formula%20of%20the%20Accuracy,confusion%20matrix%20at%20the%20denominator.>

Great Learning Team 2020. Understanding Curse of Dimensionality. *Great learning* October 1st. Available at: <https://www.mygreatlearning.com/blog/understanding-curse-of-dimensionality/#:~:text=Principal%20Component%20Analysis%2C%20or%20PCA,referred%20to%20as%20principal%20components.>

Gupta, M. 2020. What, When and Why Feature Scaling for Machine Learning. *Medium* September 2nd. Available at: <https://medium.com/technofunnel/what-when-why-feature-scaling-for-machine-learning-standard-minmax-scaler-49e64c510422>.

Huang, J. and Ling, C.X. 2005. *Using AUC and Accuracy in Evaluating Learning Algorithms*. The University of Western Ontario. Available at: <https://ieeexplore.ieee.org/abstract/document/1388242>.

IBM Cloud Education 2020. Deep Learning. *IBM* May 1st. Available at: <https://www.ibm.com/cloud/learn/deep-learning>.

J. Noyes and Edwards, R.T. 2011. *EQ-5D for the Assessment of Health-Related Quality of Life andResource Allocation in Children: A Systematic Methodological Review*. Bangor University. Available at: <https://reader.elsevier.com/reader/sd/pii/S1098301511033249?token=447C0405E7371905FB70C9AD208D4E917B21D86FAA5859C78BDD94EC4C4B47E5EC7933BF52FB763AD73D47DAFBBFB77F&originRegion=eu-west-1&originCreation=20210928213650>.

Jakobsen, J.C. 2017. *When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts*. BMC Medical Research Methodology. Available at: <https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-017-0442-1>.

Jansson, K.-Å. and Granath, F. 2011. *Health-related quality of life (EQ-5D) before and after orthopedic surgery*. Department of Orthopedic Surgery at Karolinska University Hospital. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3230002/>.

Jeff Hale 2019. Don’t sweat the solver stuff. September 27th. Available at: <https://towardsdatascience.com/dont-sweat-the-solver-stuff-aea7cddc3451>.

Kwong, E. et al. 2018. *Feasibility of collecting and assessing patient-reported outcomes for emergency admissions: laparotomy for gastrointestinal conditions*. BMJ Open Gastroenterol. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6203065/>.

N. Eugene et al. 2018. *Development and internal validation of a novel risk adjustment model for adult patients undergoing emergency laparotomy surgery: the National Emergency Laparotomy Audit risk model*. NELA. Available at: <https://www.bjanaesthesia.org.uk/article/S0007-0912(18)30578-6/fulltext>.

NELA 2020. Sixth Patient Report. Available at: <https://www.nela.org.uk/Sixth-Patient-Report#pt>.

NELA [no date]. About the Audit: Background. Available at: <https://www.nela.org.uk/NELA_Background>.

Nitesh V. Chawla et al. 2002. *SMOTE: Synthetic Minority Over-sampling Technique*. University of South Florida. Available at: <https://arxiv.org/pdf/1106.1813.pdf>.

Pearse, R.M. et al. 2012. *Mortality after surgery in Europe: a 7 day cohort study*. European Surgical Outcomes Study (EuSOS) group for the Trials groups of the European Society of Intensive Care Medicine and the European Society of Anaesthesiology. Available at: <https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(12)61148-9/fulltext>.

Pykes, K. 2020. Oversampling and Undersampling. *Towards Data Science* September 10th. Available at: <https://towardsdatascience.com/oversampling-and-undersampling-5e2bbaf56dcf>.

Raj, A. 2020. A Quick and Dirty Guide to Random Forest Regression. *Towards Data Science* June 18th. Available at: <https://towardsdatascience.com/a-quick-and-dirty-guide-to-random-forest-regression-52ca0af157f8>.

Scikit-Learn [no date][a]. RBF SVM parameters. Available at: <https://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html>.

Scikit-Learn [no date][b]. sklearn.ensemble.RandomForestRegressor. Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>.

Scikit-Learn [no date][c]. sklearn.tree.DecisionTreeClassifier. Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>.

Shivers, M. 2012. What is the influence of C in SVMs with linear kernel? *StackExchange* . Available at: <https://stats.stackexchange.com/questions/31066/what-is-the-influence-of-c-in-svms-with-linear-kernel>.

Spruyt, V. 2014. The Curse of Dimensionality in classification. *Computer vision for dummies* April 16th. Available at: <https://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/>.

SurveyMethods 2011. When to Use Single Imputation or Multiple Imputation. *SurveryMethods* October 7th. Available at: <https://surveymethods.com/when-to-use-single-imputation-or-multiple-imputation/>.

Tolstrup, M.-B. et al. 2018. *Chronic Pain, Quality of Life and Functional Impairment After Emergency Laparotomy*. World Journal of Surgery. Available at: <https://link.springer.com/article/10.1007%2Fs00268-018-4778-4>.

Wesner, J. 2016. MAE and RMSE — Which Metric is Better? *Medium* March 23rd. Available at: <https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d>.

# 7 Appendix

## 7.1 Appendix A: Model functions

### 7.1.1 Multiclass classification

#### Model evaluation function

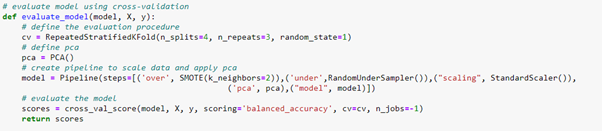


Figure : Model evaluation function for multiclass classification.

#### Multinomial logistic regression

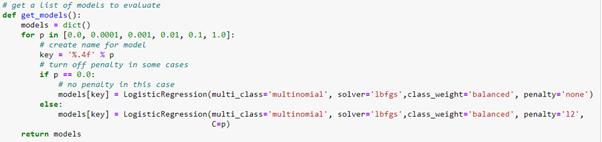


Figure : Multinomial logistic regression model and hyperparameter tuning.

#### Neural network

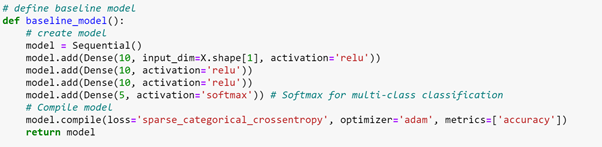


Figure : Neural network baseline model

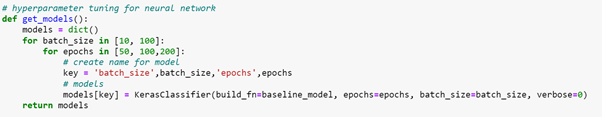


Figure : Neural network hyperparameter tuning

### 7.1.2 Regression

#### Linear regression



Figure : Linear regression model and hyperparameter tuning.

#### Decision tree



Figure : Decision tree model and hyperparameter tuning.

#### Random forest

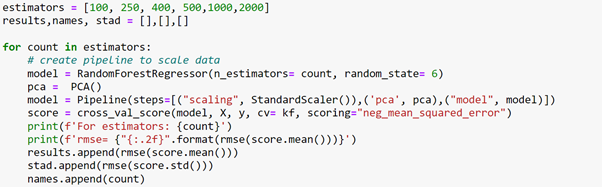


Figure : Random forest model and hyperparameter tuning.

### 7.1.3 Binary classification

#### Model evaluation function

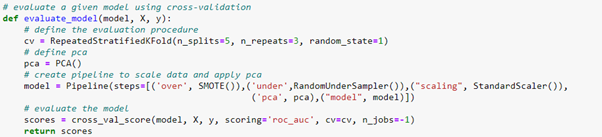


Figure : Binary classification model evaluation function.

#### KNN



Figure : KNN model and hyperparameter tuning.

#### Logistic regression



Figure : Logistic regression model and hyperparameter tuning.

#### SVM

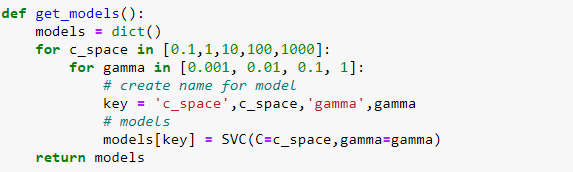


Figure : SVM model and hyperparameter tuning.

## 7.2 Appendix B: Precision, recall and false positive rate definitions

### 7.2.1 Precision

(4)

### 7.2.2 Recall

(5)

### 7.2.3 False positive rate

(6)

## 7.3 Appendix C: Participant flow diagram

**Diagram

Description automatically generated**

Figure : Patient flow diagram

## 7.4 Appendix D: Feature list

Table : List of features used in all models

|  |
| --- |
| **Feature list** |
| RTT |
| Procedure.2 |
| Surgeon.grade |
| Chronic.Pain.issues |
| Anastomotic.Leak |
| Main.procedure |
| Number.of.Procedures |
| Pre.op.PPOSSUM.mortality |
| Free.Gas.. |
| Approach |
| Crohns.Disease..F. |
| WIMD.2019.Rank |
| Gender |
| Malignancy.status |
| PMH\_of\_mental\_health\_binary |
| Hb.closest.to.discharge.death |
| Anastomotic.Leak..F. |
| Pre.Op.risk.category |
| ASA.score |
| Post.op.destination |
| Pre.Op.risk.category |
| Pus.. |
| Perforation |