Survival Following Pancreatoduodenectomy in England: Perspectives from the HES Database

Zhangdaihong Liu

September 17, 2015

Abstract

Pancreatic cancer is a worldwide common disease with very low survival rate. Recent-year studies have shown the survival rate of the disease has been improved during the last few decades. However, the change of the survival rate of patients after pancreatoduodenectomy (PD, major treatment for pancreatic cancer) has not yet been explored. This project used hospital episodes statistics (HES) data and investigated the postoperative and long-term survival of England patients after PD between the period 2001 and 2014 on 6 variables: gender, age, ethnicity, Charlson score, IMD (index for multiple deprivation) and centre volume. We found that age and Charlson score are the most significant factors that affect patients' survival after PD. IMD and centre volume are also significant to patients' survival. We analysed the importance of centralisation and found the possible optimal annual volume for a centre should be around 30. In addition, we demonstrated the predictive level of three kinds of models and found linear model works better than non-linear model on this dataset.

1 Introduction

Pancreatic cancer is the twelfth most common cancer in the world, and the tenth most common cancer in both the US and UK [1, 2]. However, it accounts for much higher proportion of death among all the cancers, being the fifth most common cause of cancer death in the UK [1, 2, 3]. Though the survival rate of patients with pancreatic cancer has increased over the last few decades [4, 5], it still has a low overall survival rate of around 20% one-year survival, 5% five-year survival and 1% ten-year survial [1, 5, 4, 6]. In most patients (over 80%) no curative treatment is possible at presentation due to one or both of locally advanced disease or metastases [4, 5, 7, 8, 1]. For those patients that can undergo surgery (10-20%) the median survival is around 2 years, with 20% alive at 5 years [9, 10].

Pancreaticoduodenectomy (PD) is the main surgical operation for treating pancreatic cancer. It is a complex procedure associated with high levels of postoperative morbidity and mortality. Patient survivals after PD is affected by many factors including age and comorbidity. Due to the complex nature of the operation and risk of adverse outcome, there has been a global push to centralise surgical services that offer PD. The centralisation of complex cancer surgery is an ongoing project carried out by the UK government since 2001. It aims to improve the outcomes of cancer surgery by centralising the surgical operation to centres with higher annual volume and more specialised surgeons [11]. Centralisation of surgical operations has been considered to play a positive role in improving patient survival [10, 12, 13]. However, increasing centre volume may lead to decrease in the quality of postoperative care for the patients, and the statistical methods that were used in previous studies have been described as overly simplistic, with possible overfitting [14]. Furthermore the effects of centralisation in the UK have to date not been reviewed.

In recent years, the research focus has been on discovering the risk factors for developing pancreatic cancer [15]. However, the emphasis of this project is on pancreatic cancer patients who have had surgical treatment in England during the last 14 years in order to investigate how their survival has been influenced. In particular, this project will address the impact of 'centre volume', a controversial factor related to centralisation. In addition, we will explore the significance of 5 other factors suggested by clinical expertise including gender, age, ethnicity, comorbidity and socioeconomic index. We aim to select a group of features that could offer the most accurate predictions on pancreatic cancer patients after resection. This is an under researched area of pancreatic cancer.

This project will look at 90-day mortality as well as long-term mortality, rather than in-hospital mortality or 30-day mortality, since 90-day mortality provides more comprehensive information on postoperative survival rate.

2 Methods

2.1 Data

The data we used during the project was recorded in the Hospital Episodes Statistics (HES) database in which comorbidity was recorded as Charlson score and socioeconomic index was linked to the index for multiple deprivation (IMD). The data was extracted for patients from England with pancreatic cancer who had the resection in the period between 1st March 2001 and 31st December 2014. Pancreatic cancer was identified using the International Classification of Disease 10th revision (ICD-10) code C25*. We excluded all the patients with incomplete information and age under 18 to separate paediatric cases. Patients with a length of stay (LOS) in hospital less than 5 days were excluded as it was very likely that these patients were miscoded, though patients coded as dying after PD within this time frame were kept in the analysis. The final dataset consisted of a total of 8002 patients with 4466 males and 3536 females.

We also extracted the whole cohort of patients from HES who had PD during the period under consideration using ICD-10 code J56*. We used this dataset to analyse the effect of centre volume on survival, since PD is not only applied to treat pancreatic cancer but also to other diseases involving duodenum and bile duct. This cohort has 13312 patients after excluding the miscoded ones.

Together with age and gender, we analysed:

- IMD (index for multiple deprivation): it was recorded using numbers $1 - 5$, 1=poor and 5=good.
- Charlson score: it was recorded using non-negative integers from minimum value 0 to maximum value 47 in the data. 0 means no comorbidity occurred, whereas larger the score, the worse physical condition of the patient.
- Ethnicity: the way of labelling ethnicity had a change during the study period from number labels to letter labels. We unified the two labels into five sub-features: White, Black, Asian, Mixed ethnicity and Others, and labeled them as 1 to 5.
- Centre volume: the annual number of surgeries performed by the centre where each patient had the surgery. This information was not originally recorded as a variable for each patient, but instead the centre code of where the patient had the operation was provided. Centre volume was extracted from the whole PD cohort regardless of cancer status by counting the number of presence of each centre code in every year and matched back to each pancreatic cancer patient.

During the analysis, gender, ethnicity and IMD were used as categorical variables. Age, Charlson score and centre volume were used as continuous variables in model fitting and also categorised into groups for survival analysis. Age and centre volume were grouped using interquartile intervals, and Charlson score was grouped by clinical convention cutting points (see Table 1). Four groups of centres correspond to low, low-medium, medium-high and high volume centres. We then defined the fifth group, very high volume centre, based upon the top decile with the purpose of finding the threshold of the optimal centre volume.

All the analysis was implemented using Excel and R Studio.

2.2 Survival analysis

Kaplan-Meier curves were plotted for each variable between different groups to study the 90-day mortality and long-term survival using R package *survival* [16, 17]. Kaplan-Meier curve estimates the probability of a subject survives longer than a certain time point. Every time when there are deaths occurring, the curve drops proportionally to the number of deaths in the remaining population. A log-rank test was also performed for each plot to see the significance of distributional difference between the groups. This method offers a straightforward way of visualising the mortality rate to the time points of interest as well as providing a general knowledge of the significance of each variable.

The graph of the 90-day mortality rate over study period was also plotted to observe the general trend of postoperative survival of pancreatic cancer. We looked at the 90-day mortality of different volume groups of centre over the last 14 years as well. This brings another direct sense of what role centre volume played on patients' survival.

Table 1: Variable cutting points. Age is grouped by quartiles into 4 groups; Centre volume is grouped by quartiles and top decile into 5 groups and Charlson score is grouped by clinical conventional cutting points into 4 groups.

Variable	1st quartile	mean	3rd quartile	top decile
Age (years)	59	66	72	
Centre volume (number of operations per annum)	3	11	30	50
Charlson score (conventional cutting points)	5	10	20	

2.3 Model fitting

The aims of model fitting are to select features that are significant to patients' survival and also to predict likely outcome (e.g. high risk) so as to build up a model that could provide survival predictions for individuals. We fitted three models to the data including Cox proportional hazard model, Cox model with lasso penalty and random forests survival model using the *survival*, *glmnet* and *party* packages of R respectively [18, 19, 20, 21]. This is designed to capture both the linear and non-linear structures of the data and to select significant variables that affect prognosis. 3-fold cross validation (divide the whole dataset into 3 subsets; train the model on two sets, predict on the third set and iterate all 3 ways) was applied to each model to estimate model performance and avoid overfitting. The 8002 pancreatic cancer patients were assigned to 3 groups in the following manner: $(3i+1)th$ patients go into group 1; $(3i+2)th$ patients go into group 2 and $(3i+3)th$ patients go into group 3. This way of distribution avoided the survival bias that time causes over the 14 years.

At the end, we calculated concordance index (C-index) of each model using *survival* package in R in order to compare the predictive level of the models. Concordance index is a common way to compare and measure the predictive power of risk prediction models. In survival analysis, C-index returns the proportion of the concordant patient pairs. Two patients are called concordant if the risk scores they acquired from the model are consistent with their survival event, i.e. the patient with the lower risk score should experience the death event at a later time point. For instance, a C-index of 0.5 is no different than tossing a coin, whereas a C-index of 1 is perfectly predictive.

Since the categories have parallel relationships between each other for feature 'ethnicity', this might lead to the model omitting the significance of some of the categories. Thus, we split this feature into 5 sub-features (White, Black, Asian, Mixed and Others) with a sparse matrix. Each patient has an entry 1 under their ethnicity and a 0 under 4 the other sub-features. Therefore, there are 10 variables in total for the model selection.

2.3.1 Cox proportional hazard regression model

Cox proportional hazard regression model (Cox model) is a typical approach of testing the effect of covariates in survival analysis. The effect of a covariate is related to the hazard function and the hazard ratio. The hazard function $h(t)$ is the probability of a survival event happening over a unit of time. It consists of two parts: the baseline hazard function which tells the changes of hazard at a baseline level, denoted by $h_0(t)$, and the effect parameter β, in which hazard varies by the product of β and the predictor variable exponentially.

The hazard function can be expressed as

$$h(t) = h_0(t)\exp(\beta X) \tag{1}$$

where X is an explanatory predictor variable, and it can be continuous or categorical.

The hazard ratio r is defined as :

$$\begin{aligned} r &= \frac{h_0(t)\exp(\hat{\beta}x_1)}{h_0(t)\exp(\hat{\beta}x_2)} \qquad (2)\\ &= \exp((x_2 - x_1)\hat{\beta}), \qquad (3)\end{aligned}$$

where x_1, x_2 are two different values of the same predictor variable and $\hat{\beta}$ is the estimated regression coefficient. Therefore, there is no need to assign the baseline hazard function for Cox model.

For example, assume the predictor is age. If patient 1 is 60 years old and patient 2 is 70, then patients 2 is $\exp(10\hat{\beta})$ times more like to experience the survival event (death) than patient 1.

Before applying the Cox model, a key assumption should be tested, namely the proportional hazard (PH) assumption. This assumption makes sure the hazard ratio keeps constant over time. Besides, for continuous variable, linearity should also be examined.

In medical science, the Cox model is often used as multi-variate model since the patient survival time is very likely to be affected by more than one factor. The variables of interests are normally selected by clinical experience and then fitted into the model to assess their quality. The key point of applying Cox model is to select significant features. Therefore, Akaike information criterion (AIC) and Bayesian information criterion (BIC) were applied to the data for this purpose using the R package *MASS* [22]. AIC offers a relative quality test for a collection of candidate models (models with different combination of features) and calculates the information loss for each model relative to other models, therefore is a means of model selection. The model with the minimum AIC value should be selected. BIC is another model selection method but stricter than AIC. Similarly, the model with the lowest BIC value is preferred. In our case, both AIC and BIC procedures were used to consider all the 10 variables and come up with the optimal model having some/all/one variable(s). Then we calculated the C-index to compare the two criteria.

2.3.2 Cox model with lasso penalty

Lasso (Least Absolute Shrinkage and Selection Operator) regularisation was applied to the Cox model with the advantage of reducing the number of variables by discarding the less relevant ones. This is achieved by using R package *glmnet*. The models underlining this package are called 'Lasso and Elastic-Net Regularized Generalized Linear Models' (GLM). The baseline solution for GLM is the same as for many other statistical approaches, maximum likelihood estimation (MSE) using least squares method. The objective function with lasso penalty term is

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} R_\lambda(\beta_0,\beta) = \min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} \left[\frac{1}{2N} \sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2 + \lambda\|\beta\|_{l_1}\right], \quad (4)$$

where N is the number of observations and λ is the coefficient for lasso penalty.

The *glmnet* package in R provides the following advantages of fitting a GLM:

- It can deal with the sparsity of the predictor variables.
- It has the feature selecting function.
- It is flexible with the penalty term (although lasso was used to reduce the number of features).

2.3.3 Decision tree

'Drawing' a decision tree graph of the data helps to visualise the path of regression/classification. A decision tree has leaves representing class labels or real numbers with probability distribution; The branches are the criteria leading to the leaves and the non-leaf nodes are the input features. In survival analysis with censored data, the leaves show the Kaplan-Meier curves of the patients in the respective classes.

In this project, conditional inference tree was used to show the data structure (see Figure 2 for an example). A conditional inference tree is a decision tree which estimates a regression relationship by binary recursive partitioning in a conditional inference framework [23]. It splits the tree recursively based on the adjusted p-value of the hypothesis test. The global null hypothesis assumes that every input variable is independent of the response variable. The algorithm stops if this hypothesis cannot be rejected, otherwise it chooses the variable with the smallest p-value, and which has the strongest association with the response, to start splitting the tree. Then for each selected variable, it implements the binary split.

2.3.4 Random forests model

The random forests model is an ensemble learning regression/classification method which 'grows' a preset number of decision trees in the data, and can capture the non-linear structure of the data. The trees grow based on bootstrap aggregating, i.e., the whole dataset is resampled into a few subsets with replacement and a decision tree is grown on each of these subsets. Every tree in the forests is grown to its largest extent. The random forests then take the average over all the decision trees so that it avoids overfitting by single decision tree.

The *party* package in R offers an effective way of implementing random forests by fitting conditional inference trees to bootstrap samples. This package has the following advantages:

- It can process censored data in survival analysis.
- It can deal with predictors with different measurement scales, i.e. both categorical and continuous variables can be handled at the same time.

- The aggregation scheme is not simply averaging the outcomes of all the tress but weighted prediction extracted from each tree.

3 Results

During the study period, the median age of pancreatic cancer patients increased from 65 to 67 years old. The number of pancreatic cancer patients who have had surgery each year increased from around 350 in 2001 to over 600 in recent years. However, the 90-day mortality rate has dropped over 5%. Another interesting observation is that the number of surgeons operating PD per year has not decreased but rather fluctuated over the last 14 years, whereas the number of centres providing PDs has decreased significantly due to the centralisation policy from 80 centres in 2001 dropped to under 30 centres in 2014 (all the graphs see supplementary §S1).

3.1 Kaplan-Meier Curve Analysis

Figure 1a-1f present the Kaplan-Meier plots for age groups, Charlson score groups and centre volume groups respectively, showing 90-day survival probability on the left and long-term survival on the right. The p-values suggest that all of the features are significant between their groups in a univariate sense. The older or the higher Charlson score the patient gets, the higher the risk of death becomes. Also patients operated in lower volume centres have higher mortality rate in both postoperative and long-term senses. See supplementary §S2 for the full list of Kaplan-Meier curves for all variables.

3.2 Patients Mortality and Centre Volume

We used the centre volume quartiles and the top decile to group the centres into 5 groups - low, low-medium, medium-high, high and very high. The low and low-median volume groups have experienced significant decrease in the number of surgeries over the past 14 years, whereas the number of surgeries performed by high and very high volume groups has increased rapidly (see supplementary Figure S5).

The annual 90-day mortality rate was calculated for each centre volume group (see supplementary Figure S7). The 90-day mortality is about 10% worse in the low volume centre group than in the very high volume centre group on average. This offers the first insight into the effect of the centralisation on patients postoperative mortality. We observe distinguished behaviour of 90-day mortality in different centre volume groups: the patients in relatively high volume centres have a lower mortality rate than the ones in lower volume centres.

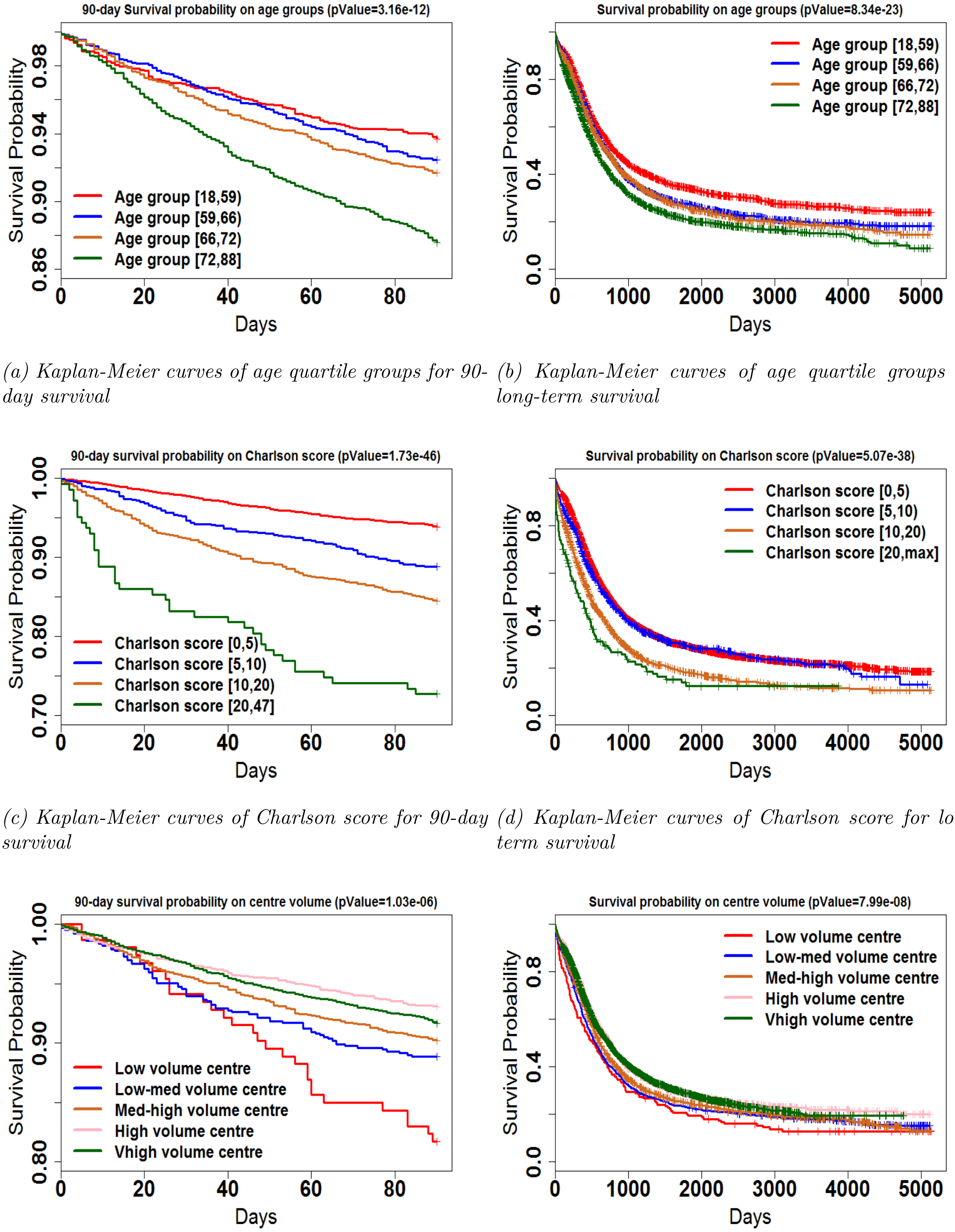
We also considered the Kaplan-Meier curves displaying the 90-day and long-term survival on Figure 1e and 1f. The mortality by the 90th day after surgery in the low volume centre group is over 10% worse than the best performing group. The p-value of the log-rank test suggests that the distributions of each centre volume group are significantly different from each other, which implies that centre volume is an important feature on patient survival. Particularly, the difference between the high and very high volume groups is not obvious in the first 30 days, but the survival of the high volume group outperforms the very high volume group by the end of 90 days as well as the end of the long-term survival. Therefore, we considered next the p-value of log-rank test between these 2 particular groups. It turned out that the p-value of 90-day survival is 0.078 and long term survival is 0.996. Neither of this is significant, which indicates that these two groups are not from different distributions. Thus the threshold of optimal volume for performing PD per centre per year is no bigger than 30 which is the criterion for high volume centre.

In the long-term survival figure (Figure 1f), we can still see the the track of the low volume centre group is significantly lower than the other groups.

In addition, the centres in high and very high volume groups tend to operate on the older patients with worse IMD and Charlson score (see Table 2). This fact should be taken into consideration when comparing the mortality between centre groups as it will lead to even more contrasting behaviour between the best and worst performing groups.

3.3 Cox Proportional Model Fitting

The proportional hazard assumption was examined for each variable before applying the Cox model (details see supplementary §S3). The estimated effect parameter $\hat{\beta}$, hazard ratio and C-index were calculated for each of the 10 variables to assess their individual effect and predictive level on the patient survival time (see Table 3). In Table 3, under the categorical centre volume, the factors show the relationships between every following category and the first one.



(a) Kaplan-Meier curves of age quartile groups for 90-day survival

(b) Kaplan-Meier curves of age quartile groups for long-term survival

(c) Kaplan-Meier curves of Charlson score for 90-day survival

(d) Kaplan-Meier curves of Charlson score for long-term survival

(e) Kaplan-Meier curves of centre volume for 90-day survival on pancreatic cancer patients

(f) Kaplan-Meier curves of centre volume for long-term survival on pancreatic cancer patients

Table 2: Patient Characteristics in Different Centre Volume Groups

	Median age	Mean IMD	Mean Charlson score
Low volume centre	64	3.268	3.275
Low-med volume centre	65	3.296	3.332
Med-high volume centre	66	3.169	4.201
High volume centre	66	3.120	4.394
Very high volume centre	67	3.164	4.728

Table 3: Results for univariate Cox model. $\hat{\beta}$ is the effect parameter. SE is the standard error of C-index. Factors 2-5 under categorical centre volume show their relationship with the first group, i.e. low volume centre group.

	$\hat{\beta}$	Hazard ratio	p-value	C-index	SE(C-index)
Gender	-0.025	0.975	0.359	0.504	0.004
Age	0.167	1.017	<2e-16	0.546	0.004
Charlson score	0.028	1.028	<2e-16	0.549	0.004
Centre volume (continuous)	-0.002	0.998	2.16e-05	0.524	0.004
Centre volume (categorical):	-0.069	0.933	1.77e-08	0.526	0.004
factor (low-med volume centre)2	-0.119	0.888	0.220		
factor (med-high volume centre)3	-0.176	0.839	0.052		
factor (high volume centre)4	-0.322	0.725	4.53e-04		
factor (vhigh volume centre)5	-0.307	0.736	6.32e-04		
IMD	-0.030	0.971	0.002	0.518	0.004
Race:					
White	0.089	1.093	0.010	0.505	0.003
Black	-0.202	0.817	0.147	0.500	0.001
Asian	-0.259	0.772	0.017	0.499	0.001
Mixed	-0.038	0.963	0.311	0.499	0.003
Others	-0.201	0.818	0.134	0.499	0.001

From the C-index column in Table 3, we can see that the Charlson score and age showed relatively high predictive level on survival time; centre volume and IMD are slightly less informative, while the other variables look rather random.

Since Cox model cannot deal with singular matrix and splitting ethnicity into 5 sub-features gives a singular matrix, we excluded every sub-feature of ethnicity and fitted the rest 9 variables into the Cox model for the whole dataset to evaluate their significance. It turned out that sub-feature 'mixed' is the least significant. Therefore, the other 9 were selected for fitting the model (see Table 4 for more details about their significance). In the column of p-value, the ones marked with a * are significant variables. We also examined the dataset using 'centre volume' as a continuous and as a categorical variable. The difference is within a standard error. This implies there is no significant difference between using the two types of that variable. To keep consistency with the centralisation analysis, the categorical 'centre volume' was used.

A 3-folded cross-validation was performed on the data. In each set of cross-validation, AIC and BIC were applied to select significant variables and obtain the predictions for one third of the data. AIC and BIC selected different variables in every subset of the data, therefore led to different predictions. The C-index calculated through the model AIC suggested is 0.5793 with standard error 0.0043. The C-index for Cox model suggested by is 0.5789, also with standard error 0.0043. The difference between two C-indices is within one standard error, therefore there is not a significantly better criterion. More interestingly, in all 3 sets of cross-validation, the 'centre volume' feature was selected by both AIC and BIC to be significant.

3.4 Cox Model with Lasso Penalty Fitting

Fitting the GLM with all 10 variables and 3-fold cross validation output important variables

*Table 4: Multi-variate Cox model. $\hat{\beta}$ is the effect parameter. SE is the standard error of effect parameter. P-values with a * show the significant variables.*

	$\hat{\beta}$	Hazard Ratio	SE($\hat{\beta}$)	p-value
Gender	-0.010	0.990	0.027	0.720
Age	0.017	1.017	0.001	<2e-16 *
Centre volume (categorical)	-0.087	0.917	0.012	2.83e-12 *
factor (low-med volume centre)2	-0.132	0.876	0.097	0.174
factor (med-high volume centre)3	-0.229	0.795	0.090	0.012
factor (high volume centre)4	-0.380	0.684	0.092	3.76e-05*
factor (vhigh volume centre)5	-0.384	0.681	0.090	2.00e-05*
Charlson score	0.028	1.028	0.002	<2e-16 *
IMD	-0.046	0.955	0.010	4.4e-06 *
White (race)	0.046	1.047	0.038	0.223
Asian (race)	-0.231	0.793	0.113	0.041
Black (race)	-0.136	0.873	0.143	0.342
Others (race)	-0.134	0.875	0.138	0.333

Table 5: C-index of the Models

	Cox Model		GLM	Random Forests
	AIC	BIC		
C-index	0.5793	0.5789	0.519	0.481
Standard Error	0.004	0.004	0.004	0.004

(variable with non-zero coefficients) on 3 subsets. The overlapping variables are 'age', 'IMD', 'centre volume', 'Charlson score', 'white (race)', 'Asian (race)' and 'others (race)'. No significance level was provided because of the theoretical design of the function in *glmnet* package. Predictions of type risk probability were also calculated through R and the C-index for GLM is 0.519 with standard error 0.0043.

3.5 Conditional Inference Tree

A single conditional inference tree of the whole dataset was plotted to see the structure of the data as shown in Figure 2. As explained in section 2.3.3, the tree is splitting from the most significant feature (smallest p-value) which is 'age' as shown on the top of the tree. Age 49 is the threshold of dividing the left and right branches. So patients older than 49 go down to the right branch, and the rest go to the left branch. Following this rule until reaching the leaf of the tree, then we can see the Kaplan-Meier curve of that group of patients with the size of the group in the bracket. For instance, if a patient is young (age $\leq$ 32) and has not bad comorbidity (Charlson score $\leq$ 8), then he has a very optimistic long-term survival probability as shown in node 4. However, if a patient is older than 49, then the Charlson score becomes the next important feature to look at. If the Charlson score is not too high ($\leq$ 9), but the patient is very old (age > 71), had the operation in a low volume centre and lives under a deprived condition (IMD $\leq$ 2), then node 16 displays a very poor survival probability.

In the whole graph, 4 variables were selected for splitting the tree: age, Charlson score, centre volume and IMD. This means that they are the most significant features to the response in a conditional inference tree of the whole dataset.

3.6 Random Forests Model Fitting

300 such conditional inference trees were grown in each fold of cross validation to form the random forests. The C-index calculated from the predictions of random forests model is 0.481 with standard error 0.0043. This predictive level is less than random selection.

4 Conclusion

This project has investigated patients' survival after PD between 2001 and 2014 and the possible predictors for prognosis from hospital data (HES) with a particular emphasis on the fact of centralisation. During these 14 years, the 90-day mortality has increased by 5% despite the fact that the

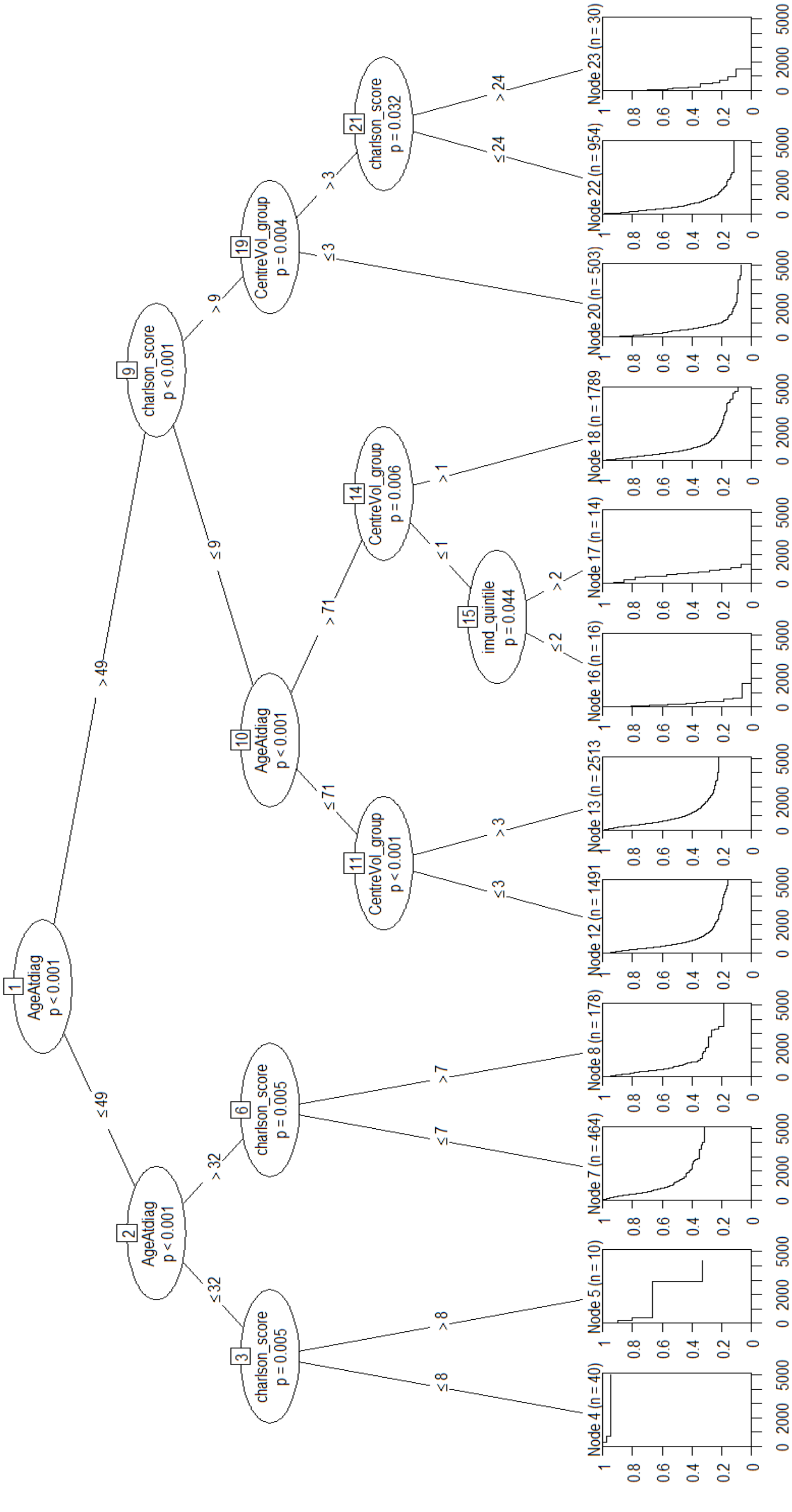


Figure 2: Decision tree graph of the whole dataset. AgeAtdiag is the patient's age at diagnosis; CentreVol_group is the categorical centre volume and imd_quintile is IMD. The number on the branches is the threshold of the binary split for the node above them. The leaves are the long-term Kaplan-Meier survival curves for the patients in respective groups. The sample size of the leave is shown in the bracket.

number of operated patients has doubled. The centralisation has caused the number of centres providing PDs to decrease from 80 to less than 30 with the median number of operations increased from 3.5 per centre in 2001 to 40 per centre in 2014. From the Kaplan-Meier analysis, we found that the optimal annual number of operations for a centre should be around 30 (definition of high volume centre) since the patients' survival is not improved after this threshold, whereas the post-operative survival decreased after this threshold.

The Kaplan-Meier analysis provided a straightforward visualisation of the impact on every selected feature (gender, age, ethnicity, Charlson score, IMD and centre volume). Male patients showed slightly (about 2%) worse mortality rate than the female patients, but this outcome is not significant. Ethnicity showed no significance as well. Patients with old age, high Charlson score, low IMD or operated in lower volume centre generally have poorer prognosis.

Different model fittings offered different angles of analysing the data. The 6 selected features have different levels of influence on the patients' survival: Charlson score and age showed the most significance; Then followed by centre volume and IMD (based on Cox model). Those four variables also appeared in the GLM and the conditional inference tree graph. From the C-index (Table 5), we can see that Cox model showed the best predictive level among all three models. Random forests model performed worst which implies linear model works better than the non-linear one on this dataset.

Further study could investigate more on the threshold of optimal annual centre volume by considering the fact that higher volume centres tend to operate on worse conditioned patients. Moreover, it should take the interactions between variables into consideration when fitting the Cox model. For example, older patients are likely to have higher Charlson score. Considering them independently may over estimate either of the influences.

Acknowledgments

My great gratitude goes to my supervisors Dr Rich Savage (University of Warwick, Systems Biology), Professor Paul Moss (University of Birmingham, School of Cancer Sciences) and Professor Daniel Ray (Queen Elizabeth Hospital, Clinical Informatics Unit) for their guidance and motivation. Thanks also goes to Mr. Keith Roberts (University Hospital Birmingham) for his assistance on clinical aspects of the project and Felicity Evison (Queen Elizabeth Hospital, Clinical Informatics Unit) for her technical assistance on the data.

References

[1] Cancer research uk. `http://www.cancerresearchuk.org`. Accessed: 2015-08-22.

[2] International agency for research on cancer. `http://globocan.iarc.fr`. Accessed: 2015-08-22.

[3] Sara Raimondi, Patrick Maisonneuve, and Albert B Lowenfels. Epidemiology of pancreatic cancer: an overview. *Nature Reviews Gastroenterology and Hepatology*, 6(12):699–708, 2009.

[4] D Hariharan, A Saied, and HM Kocher. Analysis of mortality rates for pancreatic cancer across the world. *HPB*, 10(1):58–62, 2008.

[5] Hirshbetg foundation for pancreatic cancer research. `http://www.pancreatic.org`. Accessed: 2015-08-22.

[6] Audrey Vincent, Joseph Herman, Rich Schulick, Ralph H Hruban, and Michael Goggins. Pancreatic cancer. *The Lancet*, 378 (9791):607–620, 2011.

[7] Giles Bond-Smith, Neal Banga, Toby M Hammond, Charles J Imber, et al. Pancreatic adenocarcinoma. *BMJ*, 344, 2012.

[8] Donghui Li, Keping Xie, Robert Wolff, and James L Abbruzzese. Pancreatic cancer. *The Lancet*, 363(9414):1049–1057, 2004.

[9] Subhankar Chakraborty and Shailender Singh. Surgical resection improves survival in pancreatic cancer patients without vascular invasion-a population based study. *Annals of gastroenterology: quarterly publication of the Hellenic Society of Gastroenterology*, 26 (4):346, 2013.

[10] RF De Wilde, MGH Besselink, Ingeborg van der Tweel, IHJT de Hingh, CHJ van Eijck, CHC Dejong, RJ Porte, DJ Gouma, ORC Busch, and I Quintus Molenaar. Impact

of nationwide centralization of pancreaticoduodenectomy on hospital mortality. *British Journal of Surgery*, 99(3):404–410, 2012.

[11] NHS Choices. Improving outcomes: a strategy for cancer. 2011.

[12] GA Gooiker, W Van Gijn, MWJM Wouters, PN Post, CJH Van De Velde, and RAEM Tollenaar. Systematic review and meta-analysis of the volume–outcome relationship in pancreatic surgery. *British Journal of Surgery*, 98(4):485–494, 2011.

[13] GA Gooiker, Valery Eduard Petronius Paulus Lemmens, MG Besselink, OR Busch, BA Bonsing, I Quintus Molenaar, RAEM Tollenaar, IHJT de Hingh, and MWJM Wouters. Impact of centralization of pancreatic cancer surgery on resection rates and survival. *British Journal of Surgery*, 101 (8):1000–1005, 2014.

[14] Damien J LaPar, Irving L Kron, David R Jones, George J Stukenborg, and Benjamin D Kozower. Hospital procedure volume should not be used as a measure of surgical quality. *Annals of surgery*, 256(4):606–615, 2012.

[15] Albert B Lowenfels and Patrick Maisonneuve. Risk factors for pancreatic cancer. *Journal of cellular biochemistry*, 95(4):649–656, 2005.

[16] Terry M Therneau. *A Package for Survival Analysis in S*, 2015. URL http://CRAN.R-project.org/package=survival. version 2.38.

[17] Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model.* Springer, New York, 2000. ISBN 0-387-98784-3.

[18] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL http://www.jstatsoft.org/v33/i01/.

[19] Torsten Hothorn, Peter Buehlmann, Sandrine Dudoit, Annette Molinaro, and Mark Van Der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, 2006.

[20] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(25), 2007. URL http://www.biomedcentral.com/1471-2105/8/25.

[21] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(307), 2008. URL http://www.biomedcentral.com/1471-2105/9/307.

[22] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S.* Springer, New York, fourth edition, 2002. URL http://www.stats.ox.ac.uk/pub/MASS4. ISBN 0-387-95457-0.

[23] Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15 (3):651–674, 2006.