UMEÅ UNIVERSITY

Statistical analysis of corrective and preventive maintenance in medical equipment

Author: Linn von Schewelov
Cooperation Partner: Centre for Biomedical Engineering and Radiation Physics

Master's Thesis , 30 ECTS

Master of Science in Industrial Engineering and Management, 300 ECTS

Industrial Statistics

Spring 2022

Abstract

Maintenance of medical equipment plays an important role in ensuring the healthcare quality so that the care can be conducted with minimal risk. Preventive maintenance is performed to maintain the equipment in satisfactory operating condition, while corrective maintenance is made when there is an unpredicted maintenance requirement. This study aims to determine what effect preventive maintenance has on corrective maintenance. A correlation analysis, regression analysis and survival analysis are performed on work-order data from 2000-2021. The results obtained indicate that increasing the number of preventive maintenances made to medical equipment will decrease the number of corrective maintenances required for the medical equipment.

Keywords: Correlation Analysis, Regression Analysis, Survival Analysis, Medical equipment, Corrective Maintenance, Preventive Maintenance, Least squares, Cox Proportional Hazards model

Sammanfattning

Statistisk analys av förebyggande och avhjälpande underhåll för medicin teknisk utrustning

Underhåll av medicin teknisk utrustning har en viktig roll för att säkerställa sjukvårdens kvalitet och att vården kan bedrivas med minimal risk. Förebyggande underhåll utförs i syfte att bevara utrustningen i tillfredställande skick medan avhjälpande underhåll utförs när det uppstår ett oförutsägbart underhållsbehov. Syftet med denna studie är att bestämma vilken effekt förebyggande underhåll har på avhjälpande underhåll. En korrelations analys, regressionsanalys och överlevnadsanalys utförs på arbetsorderdata från 2000-2021. De erhållna resultaten indikerar att ett ökande antal förebyggande underhåll på medicinsk teknisk utrustning kommer att minska antalet avhjälpande underhåll som krävs för den medicinsk tekniska utrustningen.

Acknowledgement

I would like to thank my supervisor from Umeå University, Alp Yurtsever, for his assistance and support throughout this thesis. I would also like to thank Per Hallberg at the Center for Biomedical Engineering and Radiation Physics who arranged the collaboration, making the work possible. I wish to extend my special thanks to my supervisor at the University Hospital of Umeå, Jonas Arrefalk, for his commitment, support, and for providing valuable information about medical equipment. Finally, I wish to thank Urban Wiklund for telling me about survival analysis and how it can be used.

Contents

1 Introduction

1.1 About the Centre for Biomedical Engineering and Radiation Physics at the University Hospital of Umeå

The Centre for Biomedical Engineering and Radiation Physics (CMTS) at the University Hospital of Umeå (NUS) conducts research, development, and education in collaboration with Umeå University (Region Västerbotten 2022). CMTS has a department of *Biomedical Engineering* (MT) which is responsible for acquiring, installing and maintaining of medical equipment. The medical technicians at MT control that they have received the correct medical equipment and that the performance of the medical equipment is sufficient. The suppliers or the medical technicians conduct the maintenance of the medical equipment. The maintenance is done to prevent future deviation or malfunction and repair sudden incidents, malfunctions or deviations. The two different types of maintenance are defined below.

- Corrective maintenance (CM). This type of maintenance is made when there is an unpredicted maintenance requirement or an incident, failure or malfunction that forces prompt attention and change in the maintenance schedule[1].
- Preventive maintenance (PM). This type of maintenance is executed on non-failed systems to prevent future failures and to maintain it in satisfactory operating condition (Pecht & Kang 2018, 729).

The preventive maintenance of medical equipment is executed according to an interval provided by the supplier (Sveriges Kommuner och Landsting, 2009). The maintenance plays an important role in ensuring healthcare quality so that the care can be conducted with minimal risk (Revenäs and Derneborg 2019, 77). However, resources are sometimes limited, and there is not enough time to execute preventive maintenance according to the schedule provided by the supplier. In these cases, medical staff follow a ranking index that depends on a model called 'Prioritisation model for Preventive Maintenance' PMFU (Sveriges Kommuner och Landsting, 2009). This model prioritises medical equipment depending on risk, maintenance requirements, and Business-critical functions, resulting in a prioritisation value on a scale from 1 to 30(Sveriges Kommuner och Landsting, 2009). In case of additional limitations on time and resources, the hospital staff can deviate from this ranking-index and calculate their PMFU-value to get a new prioritisation value[2].

1.2 Problem background

There is a perception at MT that employees value the effects of preventive maintenance on life expectancy of the medical equipment highly[2]. However, the effect of preventive maintenance on the life expectancy has not been examined. Neither has it been proven that preventive maintenance significantly impacts life expectancy. The total cost of service and support in the hospital year 2020 amounted to about 123 million SEK[1]. Since the cost for service and support is high, it is valuable to examine whether the maintenance is significant and how much effect it has on the life expectancy of the medical equipment.

1.3 Problem Description

The purpose of the project is to evaluate the current preventive maintenance to determine how the knowledge and time of the medical technicians can be utilised. The goal is to determine whether preventive maintenance affects corrective maintenance. If there is an effect, a secondary

[1]Arrefalk, Jonas; Project manager at CMTS. 2022. Meeting 19 Jan
[2]Arrefalk, Jonas; Project manager at CMTS. 2022. Meeting 26 Jan

goal is to determine how significant this effect is for all the different medical equipment. If the time constraints allow, there is an additional goal to build a model that predicts when preventive maintenance should be executed for each medical equipment respectively.

The main task of this thesis work is to answer the primary question. If the time constraint allows for further investigation, the secondary questions are evaluated.

1.3.1 Primary Question

Does recurring preventive maintenance result in less corrective maintenance?

1.3.2 Secondary Questions

- Has the deviation from national prioritisation of preventive maintenance affected the number of corrective maintenance or deviations?
- Does recurring preventive maintenance result in less deviations in healthcare?

1.4 Limitations

This project performs a correlation, regression, and survival analysis of the historical maintenance data provided by the university hospital of Umeå. Correlation alone cannot answer whether preventive maintenance causes less number of corrective maintenance. This fundamental question requires performing controlled experiments on the equipment to get sensor data from the different equipment. Performing physical experiments is beyond the scope of this work due to the time constraint. Data with insufficient information about medical equipment will be left out. Therefore, medical equipment with too few observations or insufficient information on unscheduled- or preventive maintenance is left out of the model. The Biomedical Engineering department changed its computer system in the year 1999[3]. Consequently, the data used to solve this problem will be limited to observations logged after 2000-01-01.

1.5 Related Works

Che-Ani and Ali (2019) studied the relationship between corrective- and preventive maintenance in medical equipment. They work with a similar data set as we do. They have access to work-orders and perform a correlation analysis to make conclusions. Che-Ani and Ali (2019) use information about the asset per year. Their approach in the correlation analysis is used when choosing method for the correlation analysis in this study.

There are literature that study the remaining useful life (RUL) estimate. RUL is a form of survival analysis that can be useful to model the life of the assets. Liao, Zhao and Guo (2006) show that the RUL models can provide accurate predictions of the RUL. Their study uses logistic regression and Cox proportional hazards model to measure the RUL of bearings which is a widely used mechanical element. However, in their study, they have access to sensor data that can provide knowledge about the well-being of the machinery. The data accessed in this project did not provide this information. Huang et al. (2015) perform a literature review of RUL in equipment and machinery. They focus on the support vector machine based estimation of RUL. Zhu, Chen and Peng (2019) investigate the RUL of machinery and present a deep feature learning method for RUL estimation. They use time-frequency representation and multiscale convolutional neural network. Their study provides a CNN-based method that show enhanced performance in the prediction accuracy of the remaining useful life.

[3] Arrefalk, Jonas; Project manager at CMTS. 2022. Meeting 21 Jan

1.6 Data

The data consist of 412 216 work orders, where each work order or observation has 23 features. These features, together with their description, are found in Table 1. After discovering many faulty values of the feature *INSTALLDATE*, another data set was extracted from the NUS data system. This data set contained the install-date from a different database and the latest work made for each asset. This data set consists of approximately 204 000 observations. Due to secrecy in healthcare, the data is anonymous before extraction. There is no personal data, only information concerning the equipment. There is, for example, information about what type of event that occurred but not what caused the event.

Table 1: Features from the workorder data set with description and comments.

Variable name	**Description**	**Comment**
WONUM	work-order number	Unique for observations. No observations with the same WONUM.
WORKTYPE	Workorder-type	31 valid types. Exists 'old' types that are uninterpretable.
LOCATION	Location of asset	Code for location.
LOCATION_DESCRIPTION	Description of location	Description of coded location.
ASSETNUM	Asset Number	Unique for each asset. Exists more than one observation with the same Asset-number.
ASSET_DESCRIPTION	Description of asset	Description of type of asset. Not unique for assets.
MANUFACTURER	Name of manufacturer	
PLUSCMODELNUM	Model-name of asset	The manufacturers model-number of the asset.
PLUSCVENDOR	Service Vendor name	Company responsible for service.
INSTALLDATE	Date of installment	Installment of asset.
PURCHASEPRICE	Purchaseprice	Purchase-price of asset (SEK).
PMNUM	Serialnumber of asset	Only exists for observations of type 'Preventive Maintenance'.
WOPRIORITY	Proirity of asset at NUS	When worktype 500 or 501: PMFU-value (1-30) created by NUS
ALNVALUE	National proirity of asset	PMFU-value (1-30) from national PMFU-index.
REPORTDATE_	Report date of WO	Date of work-order created. Often 30 days before target completedate.
TARGSTARTDATE_	Target startdate of WO	
TARGCOMPDATE_	Target completedate of WO	
SCHEDSTART_	Scheduled startdate of WO	
SCHEDFINISH_	Scheduled finishdate of WO	
ACTSTART_	Actual startdate of WO	
ACTFINISH_	Actual finishdate of WO	
WO_STATUS	Status of Work-order	

1.7 Outline

This thesis is organised as follows: Section 2 reviews some basic notations in statistics that provide the theoretical background needed to describe the method used, how the results are generated, and how they can be interpreted. Section 3 presents the methods used, how the data is handled and how the results are produced. Section 4 presents the results from the chosen methods and how the results can be interpreted. Finally, Section 5 presents our findings and concludes our results. These sections are divided into three parts: the correlation analysis, the regression analysis, and the last part concerning the survival analysis.

2 Theory

This section presents the theoretical background used or relevant to the problem and the methods. We divide this section into four parts. We begin by defining some basic notions needed in this study and establishing the notation. Then, we present the theory in three subsections concerning correlation analysis, regression analysis and survival analysis.

Shapiro-Wilk test

The Shapiro-Wilk test can be used to identify whether a population being sampled is normally distributed. The test compares the ordered sample values with the corresponding order statistics from the normal distribution, whose probability density function f is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{1}$$

where $-\infty < x < \infty$, μ is the mean and σ is the variance . The normal distribution is denoted by $N(\mu, \sigma^2)$. In case of a Shapiro-Wilk test, the test statistic, W, is given by

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where $x_{(i)}$ are the ordered random sample values of n observations, and a_i are constants generated from the means, variances, and covariances from a normally distributed sample. The null hypothesis of this test is that the population is normally distributed.The alternative hypothesis is that the population is not normally distributed.

H_0 : The data is normally distributed

H_1 : The data is not normally distributed

Small values of the test statistic, W, indicate that the sample is not normally distributed and hence the null hypothesis can be rejected. Larger values of the test statistic indicate that you cannot reject the null hypothesis. For further information, see original article by Shapiro and Wilk (1965).

Outlier

An outlier is an observation that differs significantly from other observations in a set of data. There are various indicators and methods for identifying an outlier. For data from a normal distribution, the Grubbs test can be performed to detect outliers. The test statistics is

$$G = \frac{max_{i=1,..,N}|Y_i - \bar{Y}|}{s} \tag{2}$$

where $\bar{Y}$ is the sample mean, and s is the standard deviation. The null and alternative hypothesis are:

H_0 : there are no outliers in the data set

H_1 : There are outliers in the data set.

The null hypothesis is rejected if the test statistic

$$G > \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{\alpha/(2n),n-2}}{n-2+t^2_{\alpha/(2n),n-2}}}$$

where $t^2_{\alpha/(2n),n-2}$ denotes the upper critical value of the t-distribution with $f = n - 2$ degrees of freedom and a significance level of $\frac{\alpha}{(2n)}$. (Grubbs 1950)

2.1 Correlation

Correlation denotes the relationship or the association between two or more variables. It measures the strength of an association between variables and their direction. The correlation coefficient ranges from -1 to +1. A correlation coefficient with a value of +1 suggests that the variables are perfectly related in a positive manner while a value of -1 suggests that the variables are perfectly related in a negative manner. A zero correlation indicates that there no linear relationship between the two variables.

Correlation analysis should not be used when data is repeated measures of the same variable from the same individual at the same or varied time points. Spearman's correlation coefficient is more robust to outliers than the Pearson's correlation coefficient (Gogtay & Thatte 2017)

Karl Pearson Correlation Coefficient

The Karl Pearson's product-moment correlation coefficient, r, is a measure of the strength of a linear association between two variables. If n pairs of random variables X and Y in a random sample are denoted by $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, then the sample correlation coefficient between X and Y is given by

$$r(x, y) = \frac{\sigma_{xy}}{\sqrt{\sigma_x \sigma_y}} \tag{3}$$

where

$$\sigma_{xy} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n}$$

$$\sigma_x = \sqrt{\sum \frac{(x_i - \bar{x})^2}{n}}$$

where $\bar{x}$ is the sample mean of x and σ_y is defined analogously to σ_x. The correlation coefficient can take any value from -1 to 1. (Upton & Cook 2014)

Spearman's rank correlation

The Spearman's correlation, ρ, measures the strength and direction between two ranked variables. The correlation is given by:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{4}$$

where d_i is the difference in paired ranks and n is the number of cases. (Laerd, n.d.) The Spearman correlation can be used when the assumptions of the Pearson correlation are violated (Gogtay & Thatte 2017). Spearman's method assesses the monotonic relationship between variables, while Pearson's method determines the linear relationship between two variables (Laerd, n.d.). A monotonic function implies that the relationship is constantly increasing or decreasing. Spearman's correlation is therefore less restrictive compared to Pearson's (Laerd, n.d.).

2.2 Regression

While correlation analysis can identify relationships between variables, regression analysis can model the relationship between the response variable and one or several other explanatory variables. Regression analysis can be used to predict the response variable by using the explanatory variables. In regression models, the expected value of one variable Y is presumed to be dependent on one or more other variables. The variable Y is variously known as the response variable. The x-variables are variously known as predictor variables or explanatory variables.

Simple linear regression model

The simple linear regression model is the simplest of all statistical regression models. The model states that the response variable Y is related to the explanatory variable X by

$$Y = \alpha + \beta X + \epsilon \tag{5}$$

where the parameters α and β correspond to the intercept and the slope of the line and ϵ denotes a random error. (Upton & Cook 2014)

Least squares method

Least squares method is the process of estimating the unknown parameters of a model by minimising *the residual sum of squares*(RSS). If every observation is given equal weight, then this is ordinary least squares (OLS). With n pairs of observations $(x_1, y_1), \ldots, (x_n, y_n)$, the ordinary least squares estimates are the values for α and β that minimise the following expression

$$\sum_{j=1}^{n}(y_j - \alpha - \beta x_j)^2 \tag{6}$$

Multiple linear regression model

Multiple linear regression model is an extension of the simple linear regression model. For p number of X-variables and n observations, the model is

$$E(Y_j) = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + ... + \beta_p x_{pj}, \text{ j = 1,2,...,n} \tag{7}$$

where $\beta_0, \beta_1, .., \beta_p$ are unknown parameters. An equivalent presentation is

$$Y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + ... + \beta_p x_{pj} + \epsilon_j, \text{ j = 1,2,...,n} \tag{8}$$

where ϵ_j are random errors. In matrix terms the model is expressed as

$$E(Y) = \mathbf{X}\boldsymbol{\beta} \tag{9}$$

where $\mathbf{Y}$ is the $n \times 1$ column vector of random variables, $\boldsymbol{\beta}$ is the $(p+1) \times 1$ column vector of unknown parameters, and $\mathbf{X}$ is the $n \times (p+1)$ design matrix. This can also be expressed as

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{10}$$

where ϵ is an $n \times 1$ vector of random errors. In practice the explanatory variables may be related as in the **general polynomial regression model** expressed as

$$Y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{1j}^2 + ...\beta_p x_{1j}^p + \epsilon_j, \text{ j = 1,2,..,n} \tag{11}$$

Usually it is assumed that the random errors, and hence the Y-variables, are independent and have common variance σ^2. In this case, the ordinary least squares estimates are obtained. (Upton & Cook 2014)

Adjusted R^2

The adjusted R^2 is similar to R^2 and accounts the number of explanatory variables in the models that may vary in the model. Adjusted R^2 is given by

$$\text{Adjusted } R^2 = 1 - \frac{RSS(n-1)}{TSS(n-p-1)} = 1 - \frac{(\sum_{j=1}^{n}(y_j - E(y_j))^2)(n-1)}{(\sum_{j=1}^{n}(y_j - \bar{y})^2)(n-p-1)} \tag{12}$$

(Upton & Cook 2014)

Mean absolute error (MAE)

If $y_1, y_2...y_n$ are n observed values and $\hat{y}_1, \hat{y}_2...\hat{y}_n$ are the corresponding predicted values by some model, then the mean absolute error is

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j| \tag{13}$$

Mean square error(MSE)

The mean square error is given by

$$MSE = \frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2 \tag{14}$$

and the root mean square error (RMSE) is given by taking the root of the mean squared error. (Upton & Cook 2014)

2.2.1 Residual Standard Error (RSE)

The residual standard error (RSE) is calculated as:

$$RSE = \sqrt{\frac{nMSE}{df}}$$

where MSE is the mean square error and df is the degrees of freedom (calculated as the number of observations - number of model parameters). The smaller the RSE, the better a regression model fits the observed values.

Akaike's information criterion (AIC)

AIC is an index used to aid in choosing between competing models. Its definition is

$$AIC = -2L_m + 2m \tag{15}$$

where L_m is the maximised log-likelihood and m is the number of parameters in the model. AIC takes into account the statistical goodness of fit and the number of parameters that have to be estimated to achieve this particular degree of fit. Lower values of AIC indicate the preferred model with the fewest parameters that still provide an adequate fit to the data. (Everitt 2002)

Collinearity

Collinearity indicates situations where the explanatory variables are related by a linear function. Collinearity can be a problem and it can be hard to interpret the model since the regression coefficients have influence of other variables. Approximate collinearity can also be a problem when estimation regression coefficients.(Everitt 2002)

Backward selection

Selection methods are used for selecting a subset of explanatory variables when conducting a regression analysis. One of the most commonly used selection methods is backward elimination. The criterion used for assessing whether a variable should be removed from an existing model in backward elimination is the change in the residual sum-of-squares produced by the exclusion of the variable. An 'F -statistic' known as the *F-to-remove* is calculated as

$$F = \frac{RSSm - RSS_{m-1}}{\frac{RSS_{m-1}}{(n-m-2)}}$$

RSS_m and RSS_{m-1} are the residual sums of squares when models with *m* and *m - 1* explanatory variables have been fitted. The F-to-remove is calculated and compared with a preset term. A calculated F less than a corresponding *F-to-remove* leads to a variable being removed from the current model. In the stepwise procedure, those variables currently in the model are considered for removal by the backward elimination process.(Everitt 2002)

Bootstrap

Bootstrap is a data-based simulation method for statistical inference The basic idea of the procedure involves sampling with replacement to produce random samples of size n from the original data. Each of these samples is known as a bootstrap sample and each provides an estimate of the parameter of interest. Repeating the process a large number of times provides the required information on the variability of the estimator and an approximate 95% confidence interval can, for example, be derived from the 2.5% and 97.5% quantiles of the replicate values. (Everitt 2002)

2.3 Survival analysis

Survival analysis is used to investigate the time it takes for an event to occur. Typically, it is used to predict a patient's lifetime undergoing some sort of treatment and analyse the treatment impact. Censoring is a type of missing data problem common in survival analysis. It occurs when you track the subject through the end of the study and the event never occurs. It could also happen due to the subject dropping out of the study for reasons other than failure, or some other loss to follow-up. If the sample is censored, you only know that the individual survived up to the loss to follow-up, but you do not know anything about survival after that.

Kaplan-Meier estimate

The Kaplan-Meier estimate is a nonparametric statistic for estimating the survivor function, $P(t)$ (the probability that a component survives until time t) from observations of lifetimes when some observations are censored. A formal Kaplan-Meier estimate is defined by the following:

- Let $L_1 < L_2 < ... < L_{k-1}$ be the distinct observation limits that are less than the age $t = L_k$ at which $P(t)$ is being estimated. Let $n_{j+1} = n(L_j + 0)$ be the number of survivors observed beyond L_j where $L_j + 0$ means that losses at L_j have been subtracted off,and δ_j the number of failures observed in the interval $(L_{j-1}, L_j]$, excluding nonzero values of the left endpoints, with $L_0 = 0$. Then

$$\hat{P}(t) = \prod_{j=1}^{k}(1 - \frac{\delta_j}{n_j})$$

(Kaplan & Meier 1958)

Cox proportional hazards model

The purpose of the model is to evaluate simultaneously the effect of several factors on survival. The model examines how specified factors influence the rate of a particular event happening at a particular point in time. This rate is commonly referred as the hazard rate.

The Cox model is expressed by the hazard function denoted by $\lambda(t)$. This function can be interpreted as the probability of an event occurring at time t. It can be estimated as:

$$\lambda(t;X) = \lambda_0(t)e^{X\beta}$$

where β is a $p \times 1$ vector of unknown parameters and $\lambda_0(t)$ is an unknown function giving the hazard function for the standard set of conditions $X = 0$ and t represents the time. The quantities $\exp(\beta_i)$ are called hazard ratios (HR) and are usually obtained by maximum likelihood estimation.The set of individuals at risk at time $t - 0$ is called the risk set at time t and denoted $R(t)$; this consists of those individuals whose failure or censoring time is at least t. For the particular failure at time $t_{(i)}$, conditionally on the risk set $R(t_{(i)})$, the probability that the failure is on the individual as observed is

$$\frac{e^{x_{(i)}\beta}}{\sum_{l\in R(t_{(i)})} e^{x_{(l)}\beta}}$$

Each failure contributes a factor of this nature and hence the required conditional log likelihood is

$$L(\beta) = \sum_{i=1}^{k} x_{(i)}\beta - \sum_{i=1}^{k} \log[\sum_{l\in R(t_{(i)})} e^{x_{(l)}\beta}] \tag{16}$$

For $\epsilon, \eta = 1, .., p$ we can derive the following from equation (21)

$$\frac{\partial L(\beta)}{\partial \beta_\epsilon} = \sum_{i=1}^{k} \left(x_{(\epsilon i)} - \frac{\sum x_{(\epsilon l)} e^{(x_l \beta)}}{\sum e^{(x_l\beta)}} \right) \tag{17}$$

where the sum being over $l \in R(t_{(i)})$. Similarly

$$\frac{\partial^2 L(\beta)}{\partial \beta_\epsilon \partial \beta_\eta} = \sum_{i=1}^{k} \left([\frac{\sum x_{\epsilon l} x_{\eta l} e^{(x_l\beta)}}{\sum e^{(x_l\beta)}}] - \frac{\sum x_{(\epsilon l)} e^{(x_l\beta)}}{\sum e^{(x_l\beta)}} \frac{\sum x_{(\eta l)} e^{(x_l\beta)}}{\sum e^{(x_l\beta)}} \right) \tag{18}$$

Maximum-likelihood estimates of β can be obtained by the use of equations (22) and (23). Significance tests about subsets of parameters can be derived in various ways, for example by comparison of the maximum log likelihoods achieved.

The Cox proportional hazards model makes two assumptions:

- Proportional Hazard assumption - the survival curves for different explanatory variables must have hazard functions that are proportional over the time t. This means that the explanatory variable only changes the chance of failure - not the timing of periods of high hazard.
- Linearity assumption - the relationship between the log hazard and each covariate is linear, which can be verified with residual plots.

A value of β_i greater than zero, or equivalently a hazard ratio greater than one, indicates that as the value of the *i:th* covariate increases, the event hazard increases and thus the length of survival decreases. A hazard ratio above one indicates a covariate that is positively associated with the event probability, and thus negatively associated with the length of survival.

(Cox 1972)

Martingale residuals

Martingale residuals can be used to determine the functional form of each of the covariates in the Cox model. In order to check the linearity assumption, Martingale residuals can be used

$$r_{M_i} = \delta_i - \hat{\lambda}_0(t_i)e^{X_i^T\hat{\beta}}$$

where $\hat{\lambda}_0(t_i)$ is the estimate of the baseline cumulative hazard at t_i, and δ_i is the event indicator for subject i (Mohammed 2019). The null martingale residuals are computed from a null model, with no covariates. The null martingale residuals can show the ideal functional form of the covariates. To find out if the covariates should be of linear form, each covariate is plotted against the null martingale residuals. Each of the covariates are plotted against the null martingale residuals. A LOWESS curve is added to the plot. Locally Weighted Scatterplot Smoothing (LOWESS) is a method of regression analysis which creates a smooth line through a scatterplot. When curvature is present in the null martingale plot, you might need to add the square or the logarithm of the covariate to the model. (NCSS Statistical Software, n.d)

Schoenfeld residuals

Schoenfeld residuals are used to check the proportional hazards assumption

- For any subject $i \in D(t_k)$, which is the set of d_k failures at time t_k, is the difference between the covariate for that subject and the weighted average of covariates in the risk set is $X_i - \bar{X}(\beta, t_k)$.
- The sum of the Schoenfeld residuals over all d_k subjects who fail at t_k is

$$r = \sum_{i \in R(t_k)} \delta_{ik}[X_i - \bar{X}(\beta, t_k)] \tag{19}$$

where $\delta_{ik} = 1$ if the subject fails at time t_k and zero otherwise, $R(t_k)$ is the risk set at t_k.

(Mohammed 2019)

2.3.1 Extended Cox

The extended Cox model is not used in this work but included for future works.

If the proportional hazards assumption is not fulfilled, the covariates must be interacted with a time function. Covariates may change their values over time. Such variables are referred to as time-dependent covariates. Time-dependent covariates may be used in Cox models with extreme caution since the standard Cox model typically cannot be used to predict the survival curve over time. In the case of time-dependent covariates an extended cox model is appropriate. Let T be the failure time of interest, and let X be a set of possibly time-dependent covariates. $X(t)$ is used to denote the value of X at time t, and $\bar{X}(t) = \{X(s) : 0 \leq s \leq t\}$ to denote the history of the covariates up to time t. The conditional hazard function of T given $\bar{X}$ is

$$\lambda(t|\bar{X}) = Pr(T \in [t, t + dt)|T \geq t, \bar{X}(t)] \tag{20}$$

where $(t, t + dt)$ is a small interval from t to $t + dt$. Let there be p_1 covariates that meet the proportional hazards assumption and p_2 covariates that do not meet the assumption, then the obtained model is

$$\lambda(t, X(t)) = \lambda_0(t)exp[\sum_{a=1}^{p_1} \beta_a X_a + \sum_{b=p_1+1}^{p_2} \beta_b X_b + \sum_{b=p_1+1}^{p_2} \delta_b X_b g_b(t)] \tag{21}$$

where $g_b(t)$ is a time function. Different functions can be used, like: $g_b(t) = 0, g_b(t) = t, g_b(t) = ln(t)$. (Hartina Husain et al 2018)

3 Method

This section explains how the problem is approached. The first part explains how the data was handled and prepared. There are three parts explaining how the correlation analysis, regression analysis and survival analysis are conducted.

3.1 Motivation and Approach

During the literature study one option is to estimate the remaining useful life(RUL) of the assets since there is literature in this area. Since all literature on RUL estimation found has access to sensor data, this was disregarded and a correlation analysis was performed. Che-Ani and Ali (2019) performed a correlation analysis on a similar data set and problem, therefore the correlation method in this study is approached in the same way as in Che-Ani and Ali (2019).

After the correlation analysis, a natural next step is to perform a regression analysis. Liao, Zhao and Guo (2006) analysed the RUL with logistic regression and have access to sensor data. Since we do not have access to sensor data and want to find a relationship between preventive- and corrective maintenance, we use a different approach to the regression where we model the number of corrective maintenance instead of the RUL. Three different approaches to the regression problem are used. The main difference between these approaches is how the data is represented in the problem.

Most of the literature found on survival analysis was performed to study patient treatments. For example, Hashim and Weiderpass (2019) conducted a survival analysis to estimate the survival in cancer patients that undergo treatment and Husain et al (2018) studied the survival in patients with breast cancer. Our aim is not to study patient treatment or survival probability of patient's. Still, if we view the medical equipment in the same way as a patient, then our problem is similar to the problems Hashim and Weiderpass (2019) and Husain et al. (2018) has studied. As in the studies of patient treatment and survival probability, our maintenance equipment have a start of life and an end of life. Maintenance made to the equipment can be viewed as the treatments done to the patients. Therefor a survival analysis in performed as well.

3.2 Data preparation

The data types in the data set are changed to an appropriate type for all variables. That is, *character* for all variables except for the variables of type date and the priority- and price-variables which are of type integer. The data set with work-orders is merged with the data set containing a more accurate install-date variable. Non-valid work types are removed and the observations of the work types that are relevant are kept. A list of all the valid work types can be found in Appendix. The work types and their corresponding work type codes used in this project are

500 Preventive Maintenance

603 Corrective Maintenance

700 Settlement

Incomplete observations are removed by only keeping observations where the work-order status is 'COMPLETE' or 'CLOSED'. The date-features are investigated and observations with inaccurate dates are either replaced by another appropriate date or removed. In this project, observations after 2000-01-01 are used and observations where the asset's life started earlier than 2000-01-01 are removed. A feature called *group* is added to the data set which explains, for each observation, which type the asset belongs to. The type is determined by the brand and the model-number of the asset.

The medical technicians at NUS add the PMFU value manually in their data system whenever they decide to deviate from the national prioritisation scale. Therefore, there are missing values in the original PMFU feature. Since this is done manually, we can assume that whenever there is a missing value, NUS will use the national prioritisation PMFU value. The missing values in the PMFU feature are therefore replaced with the value of the national prioritisation PMFU value and the PMFU difference feature is set to zero in these cases.

3.3 Correlation Analysis

Data modification

The data is modified for the correlation analysis by creating a new data frame with number of corrective- and preventive maintenance per asset. A feature was added to the data set containing a group number based on the manufacturer and the model of the asset. This feature is used by identifying assets that are of the same type. Spearman's and Pearson's correlation coefficients were calculated for the ten biggest groups using equation 4 and 3. When referring to the ten biggest groups, these are the groups that have the largest number of assets hence largest number of observations. A scatter-plot of the number of preventive- and corrective maintenance visualises the relationship and colour-coded by group.

After reviewing the results, a different approach is used by integrating a method from a similar problem found in an article by Che-Ani and Ali (2019). The number of preventive- and corrective maintenance per year was calculated for each asset. Che-Ani and Ali (2019) estimated the mean of preventive- and corrective maintenance per year in order to calculate the correlation. Therefore, the mean number of preventive- and corrective maintenance within a group is calculated for each year. Spearman's and Pearson's correlation coefficients are calculated using the mean per year within a group. The mean per year within each group was visualised in a plot as well as the correlation coefficients.

Outliers and normality

The Shapiro-Wilk test, q-q plots and Grubbs test are used to identify outliers and check for normality. In consultation with the supervisor, Jonas Arrefalk at NUS, groups with insufficient data are replaced with the next biggest group. The maintenance mean per year and correlation coefficients are calculated and displayed in a scatter plot.The outliers are identified and removed. After removal and addition of new groups, Spearman's and Pearsons correlation coefficients are calculated and displayed in a plot.

3.4 Regression Analysis

For the regression analysis, the least squares method is used to fit a polynomial regression model. Backward selection is used to eliminate features. Although many features may be included and presented in the models, the important features to answer the primary and secondary questions are the preventive maintenance and the difference in PMFU value. Therefore these features are the ones that is commented and discussed in the result and the discussion part.

Adjusted R^2, AIC and RSE are used to compare models. The bootstrap method is used to construct a 95 % confidence interval of the adjusted R^2. To find a model that explains the observations relatively well, different approaches are used. The main difference between these approaches is how the data is presented and hence the formulation of the model. Three different approaches to fitting a polynomial regression model are explained below.

Model 1

The data is constructed by summarising information about one asset in one row. All information available in a lifetime for an asset is represented by one observation. A description of all variables included in the data set is found in Table 2. We fit a model to estimate the number of corrective maintenance made to the asset during its lifetime. All features in table 2, except *CM* and asset number, are used as explanatory variables as well as their polynomial.

Table 2: Description of variables included in the dataset for model 1.

Variable name	**Description**
Assetnum	the assetnumber
Group	the name of the group that the asset belongs to
PM	total number of preventive maintenance made on the asset during its lifetime
CM	total number of corrective maintenance made on the asset during its lifetime.
dif_pmfu	difference between the national prioritisation and the hospitals prioritisation
purprice	purchase price of the asset
firstPM	number of days until the first preventive maintenance occured
time_alive	number of days the asset has been alive
PM_timeRatio	the number of maintenance made divided with number of days alive

Model 2

Compared to the data set in model 1, the number of maintenance is replaced with the cumulative maintenance at time *t*. There is an observation for each maintenance made on an asset in this data set. In Table 2 a description of all variables included in the data set is presented. We fit a model to estimate the number of corrective maintenance made to the asset at time *t*. All features in Table 3, except *CM_cumul* and assetnumber, are used as explanatory variables as well as their polynomial.

Table 3: Description of variables included in the dataset for model 2.

Variable name	**Description**
Assetnum	the assetnumber, unique in the dataset
Groupname	the name of the group that the asset belongs to
PM_cumul	total number of preventive maintenance made at time t
CM_cumul	total number of corrective maintenance made at time t
dif_pmfu	difference between the national prioritisation and the hospitals prioritisation
purprice	purchase price of the asset
time t	number of days the asset has been alive
PM_timeRatio	the number of maintenance made at time t divided with number of days alive at time t.

Model 3

The third approach to the problem is influenced by the method used in the correlation analysis. The maintenance is grouped by year and which group each asset belong to. The mean of preventive- and corrective maintenance per year within a group is represented in one observation. The year-variable is represented by first year alive, second year alive, and so on until there are no observations within the group for the next year. A description of the variables in the dataset is found in Table 4 . We fit a model to estimate the mean number of corrective maintenance made to the asset during year *y* for group *g*. All features in Table 4, except *CM*, are used as explanatory variables as well as their polynomial.

Table 4: Description of variables included in the dataset for model 3.

Variable name	**Description**
PM	mean number of preventive maintenance during year y in group g.
CM	mean number of corrective maintenance during year y in group g.
Groupname	name of group g.
Year	Year y.

3.5 Survival Analysis

Survival analysis is typically used in the presence of some sensor data that can provide measurements on the physical conditions and the tear on the assets or well-being of the patient, see article written by Husain et al. (2018). Unfortunately, we do not have access to such measurements in this experiment. Instead, we conduct a survival analysis with the present work-order data. We use two different methods: The first one is the Kaplan-Meier Analysis which evolved into a Cox Proportional Hazards model. The advantage of a Cox model compared to a Kaplan-Meier is that Kaplan-Meier curves are good for visualising differences in survival between two categories, but do not work well for assessing the effect of quantitative variables. Cox proportional hazards regression can assess the effect of both categorical and continuous variables, and model the effect of multiple variables at once.

Since a majority of the observations in the data set are assets still in function, it is an advantage to use survival analysis since it can handle censored data. This means that the model uses observations where the event has not occurred yet. In this case, the model uses observations where an asset is still in function even if we are modelling the probability of failiure.(Mohammed 2019) When referring to an asset being dead or alive, an asset that is in function is the same thing as it being alive and an asset that is not in function is the same thing as the asset being dead. The data set used in the regression analysis of model 1 is used for the survival analysis. A variable, *status*, is added to the data set which explains if the asset is dead or alive today. The variable is a factor variable used for the censoring and it can take one of two values 'alive' or 'dead'.

In the Kaplan-Meier analysis the first corrective maintenance was seen as an event and the explanatory variable was which group the asset belonged to. The groups are formed based on whether the asset had a lot of preventive maintenance done to it or not. We create the three groups as follows:

g_{min} if the asset has had no preventive maintenance done to it before the first corrective maintenance occurred.

g_{med} if the asset has had one preventive maintenance done to it before the first corrective maintenance occurred.

g_{max} if the asset has had more than one preventive maintenance done to it before the first corrective maintenance occurred.

For the Cox proportional hazards model, the response variable is the probability of the asset dying after time, t. The estimated risk of failure at time t is derived by maximum likelihood estimates of β in the expression

$$h(t) = h_0(t)e^{\beta_0+\beta_1 x_1+..+\beta_p x_p}$$

The same data set used for the Kaplan Meier analysis is used but an additional variable is added. Since the response variable is not the number of corrective maintenance made to the asset we can use this variable as an explanatory variable. Hence, $CM\_timeRatio$ is added which represents the number of maintenance made divided with the number of days the asset has been alive.

To validate the model, the Martingale residuals and the Schoenfeld residuals are calculated and plotted. The Martingale residuals should not display patterns in the plot for the model to fulfil the linearity assumption. For each covariate in the model, the scaled Schoenfeld residuals are plotted against time and a smoothing spline that fits the plot is added together with a 95% confidence interval. A pattern of systematic deviations from a horizontal line in the plot shows an indication that the covariate does not fulfil the proportional hazards assumption. A p-value to test whether the slope in the plot is equal to 0, i.e horizontal or not, is provided in the plot. If the p-value is less than 0.05 then the null hypothesis that the slope is equal to zero is rejected and the covariate in question does not fulfil the proportional hazards assumption.

4 Result

In this section, examples from the data preparation, representation of the data, and results from the different methods are presented. In regression and survival analysis, we particularly focus on preventive- and corrective maintenance features as well as the PMFU feature in our discussions, as these three features seem to be more relevant to answering the primary and secondary questions.

4.1 Data Preparation

The features *installdate, reportdate, targetstartdate, targetcompletedate, schedulestart, schedulefinish, actuallstart, actuallfinish* (see Table 1 for definition) are changed to datatype *date*. The variables installdate and datemoved_min is compared and since the variable installdate has more faulty dates, the datemoved_min feature are used as the variable for the start of the life for the asset. The observations that are not of work-type 500, 603 or 700 are removed form the data set. Before removing these observations the different work-types are displayed. There are some observations where the work-type code is added manually and there has been an spelling mistake. These mistakes are corrected before the work-types is removed.

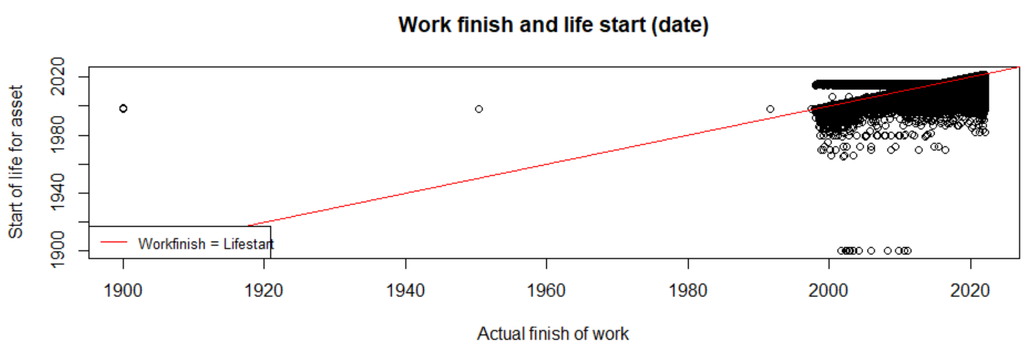


Figure 1: Data cleaning process of date-variables. Actual finish-date and Installation date of original data.

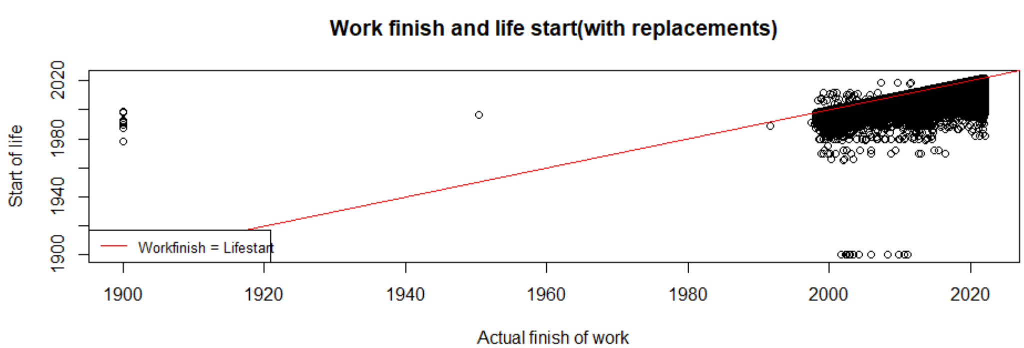


Figure 2: Data cleaning process of date-variables. Actual finish-date and Installation date with replaced dates for non-valid observations.

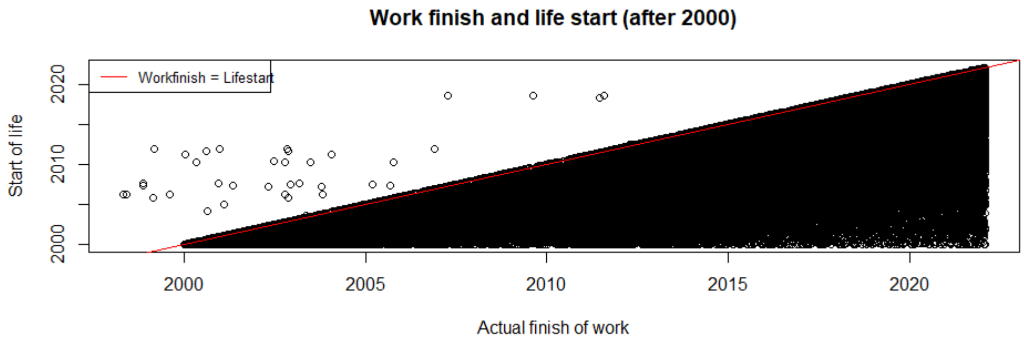


Figure 3: Data cleaning process of date-variables. Actual finish-date and Installation date of observations with install-date after 2000-01-01.

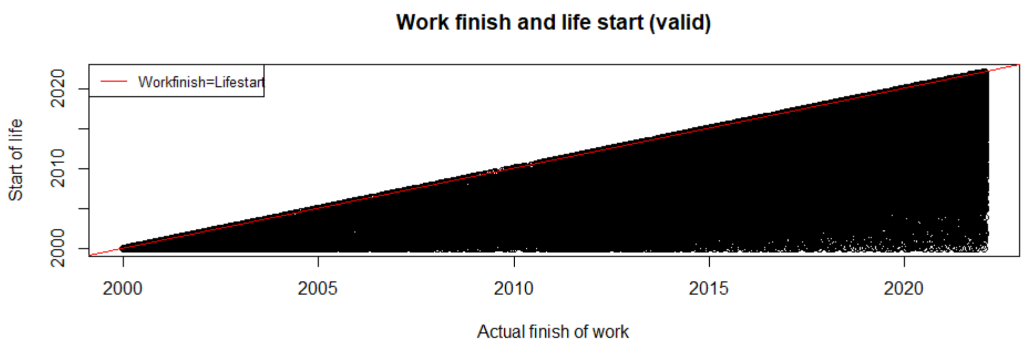


Figure 4: Data cleaning process of date-variables. Actual finish-date and Installation date of valid observations.

In order to check the validity of the date-variables, the start-of life variable and the date for the finish of the work were used. The latter variable was used since this variable had the least missing values compared to the other date-features. In Figure 1, 2, 3 and 4 four different stages of the data cleaning is displayed. The start of life variable comes from a different database where there are more valid start of life dates. In Figure 1 the original dates of the assets start of life and the date whenever there has been work done to the asset is displayed. In Figure 2 the dates where the start of life occurs after a work is done to the asset are changed to the *INSTALLDATE*. In Figure 3 all observations that live before 2000-01-01 are removed and the start of life and the finish of the work made to the asset are plotted. The red line represents whenever the start of life and the finish of the work is at the same day. Since a work-order cannot be completed before the start of life for the asset these observations are removed and displayed in Figure 4. With this date feature the days from start of life to any work done to the asset is derived. After removing or changing observations, depending on the work-type, date condition and work-order status, approximately 35 % of the original data is left.

4.2 Correlation Analysis

The data set used for the correlation analysis is built by summarising corrective- and preventive maintenance respectively for each of the assets.

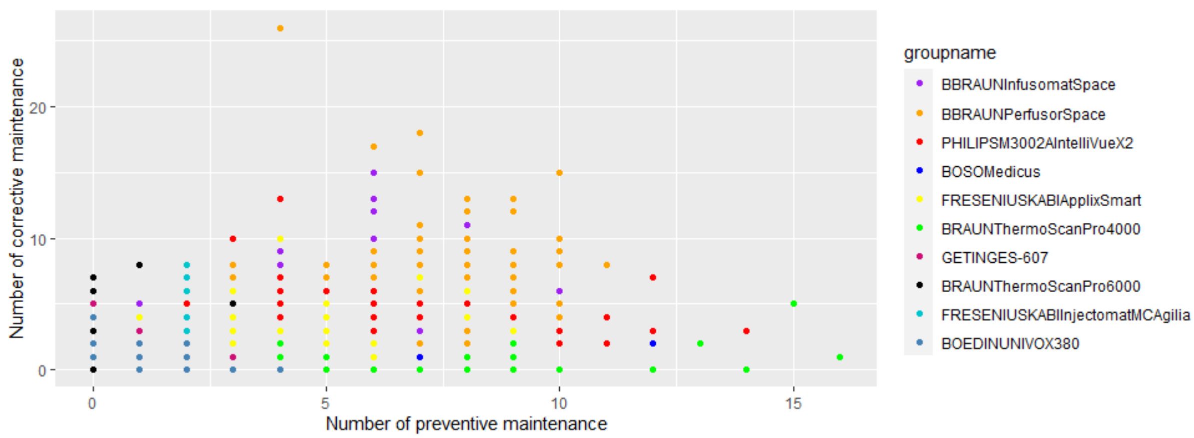


Figure 5: Preventive and corrective maintenance maintenance per group, color-coded by groups.

Figure 5 shows the total number of preventive- and corrective maintenance made for the ten biggest groups during a lifetime for an asset and is color-coded by group. The assets in this data started their life after 2000-01-01 and some of them have 'died' while others still 'live'. When we perform the correlation analysis over all groups combined, there does not seem to be a negative correlation. In other words, we did not find a direct support for the hypothesis that more preventive maintenance cause less corrective maintenance. When dividing into groups, some of them (like PHILIPS M3002AIntelliVueX2 shown in red and FRESENIUSKABI ApplixSmart shown in yellow) show a tendency to a negative correlation while other do not.

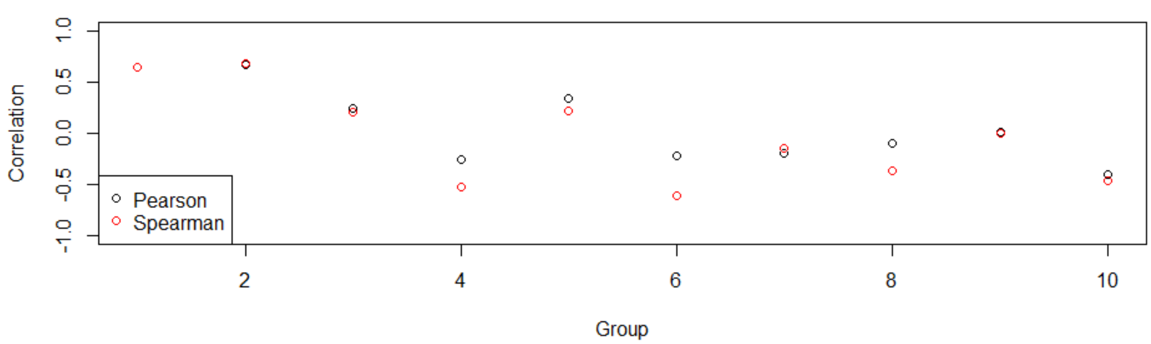


Figure 6: Pearson- and Spearman's correlation coefficients for the 10 biggest groups.

Figure 6 displays Pearson's and Spearman's correlation coefficients for the ten biggest groups(in terms of the number of assets). This correlation is based on the data set above in Figure 5. The correlation ranges from positive to negative. Groups 1,2,3 and 5 have a positive correlation between the preventive- and corrective maintenance. For these groups, the correlation suggests that the more number of preventive maintenance that is made to the asset, the more number of corrective maintenance has to be made on the asset. One could argue that a reasonable explanation for this can be how long the asset has been alive. Higher number of preventive maintenance can occur when the lifetime of the asset is longer, which could also lead to a high number of corrective maintenance, leading to a positive correlation.

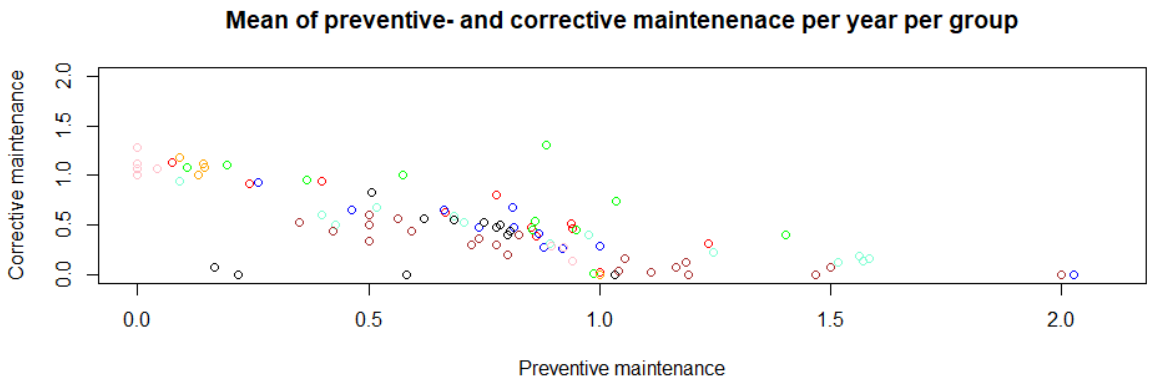


Figure 7: Mean of preventive- and corrective maintenance per year, divided into groups and colour-coded by groups.

The mean of number of corrective- and preventive maintenance within a group during a year is displayed in Figure 7. Here the maintenance is not represented per asset but per group and the mean per year is plotted. The colors in the plot correspond to each group. Here we can see that there might be a negative correlation between preventive and corrective maintenance compared to the results from the data in Figure 5.

Table 5: Pearson- and Spearman's correlation coefficient, divided into years, for the 10 biggest groups.

Group	Pearson correlation	Spearman correlation	Number of assets
1	-0.868	-0.938	523
2	-0.578	-0.602	518
3	-0.805	-0.911	323
4	-0.865	-0.896	318
5	-0.043	-0.317	303
6	-0.948	-0.912	282
7	-	-	266
8	-0.994	-0.700	253
9	-0.978	-0.752	210
10	-0.978	-0.963	208

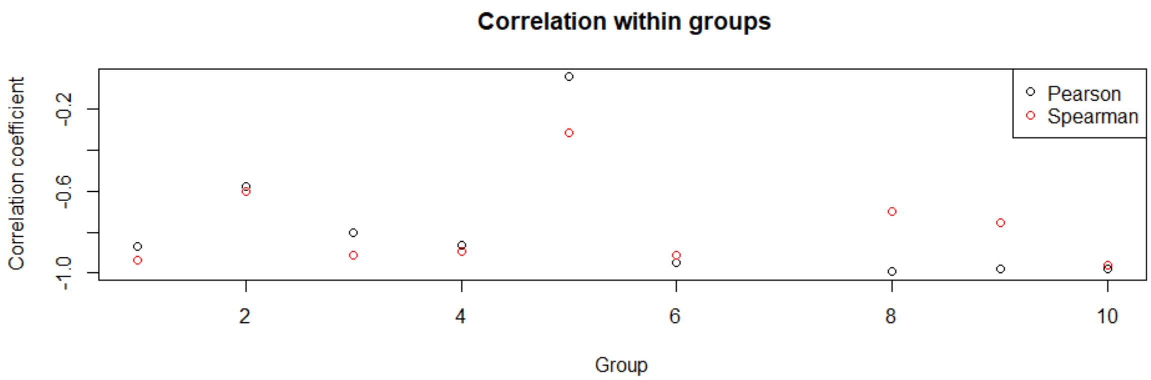


Figure 8: Pearson and Spearman's correlation coefficient of mean of preventive- and corrective maintenance per year.

The correlation between the mean of preventive- and corrective maintenance per year is shown in Table 5 and in Figure 8. After building the new dataset, a quality deficiency was discovered for group number 7. In Table 5 and in Figure 8 there is no value for the correlation fro group number 7. The information for this group was integrated in the data-system in 2020 but the assets lived before that. The historical data prior to 2020 was not added to NUS:s system. Therefore, there was only data for the year 2020 for group 7 and it is not possible to calculate the correlation coefficients.

4.2.1 Outliers

The normality and outliers in the data are identified and displayed in Table 6. The outliers are removed and the correlation is displayed in Figure 9.

Table 6: Outliers and normality for dataset with means of maintenance.

Group	**Outliers**	**Normality**
1	No	Yes
2	No	Yes
3	$PM_{2009} = 2.03$, $PM_{2016} = 0.26$	PM not normal
4	$PM_{2000} = 2$, $PM_{2005} = 1.5$	Yes
5	No	CM not normal
6	No	Yes
7	-	-
8	$PM_{2016} = 1$, $PM_{2020} = 0.09$, $CM_{2016} = 0$	CM not normal
9	No	CM, PM not normal
10	No	Yes

The correlation between preventive and corrective maintenance from the data displayed in Figure 7 without the outliers displayed in Table 6 is shown in Figure 9.

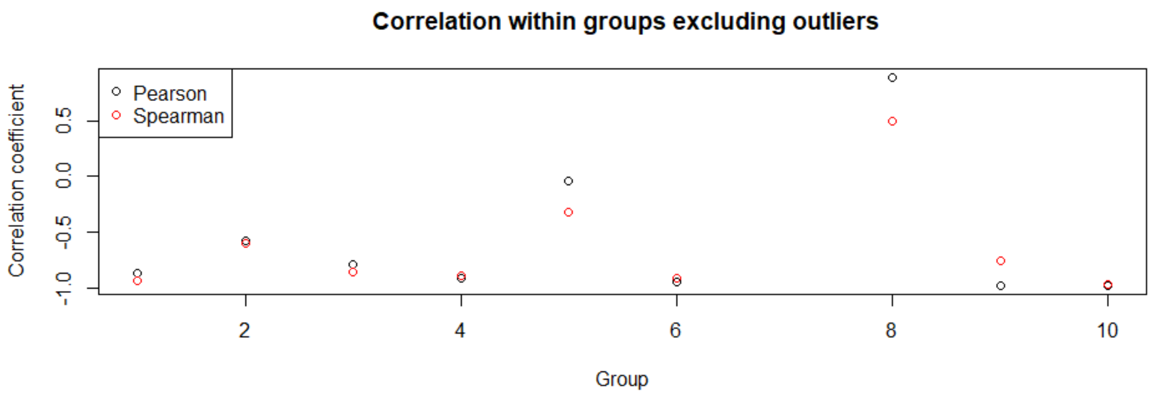


Figure 9: Pearson- and Spearman's correlation coefficient, divided into years, for the 10 biggest groups without outliers.

From Figure 8 and9 we can see that group 8 is highly affected by the outliers since the correlation differs a lot for this group. After evaluating the outliers, together with the supervisor at NUS, group number 8 was removed due to poor quality. A preventive maintenance schedule for many of the assets in this group was missing since the frequency of maintenance was manually changed to zero in 2016. The reason for this change in frequency is not clear and the supervisor suggests it is a human-error, hence the data for this group is excluded. Group 7 and 8 were removed and the next biggest groups were added.

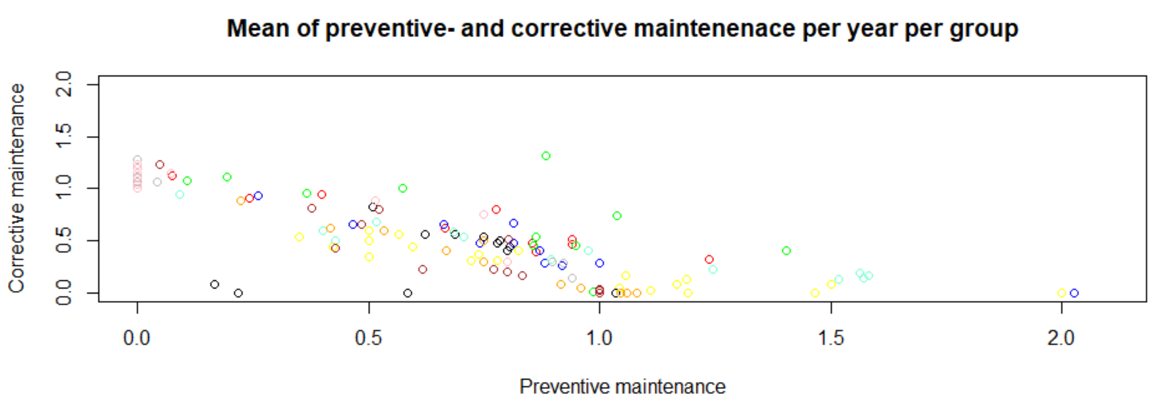


Figure 10: Preventive and corrective maintenance with valid group data

The mean number corrective and preventive maintenance within a group without group 7 and 8 and with group 11 and 12 is displyed in figure 10. The correlation of the data displayed in Figure 10 is shown in Figure 11. The correlation calculated without the outliers is displayed in Figure 12.

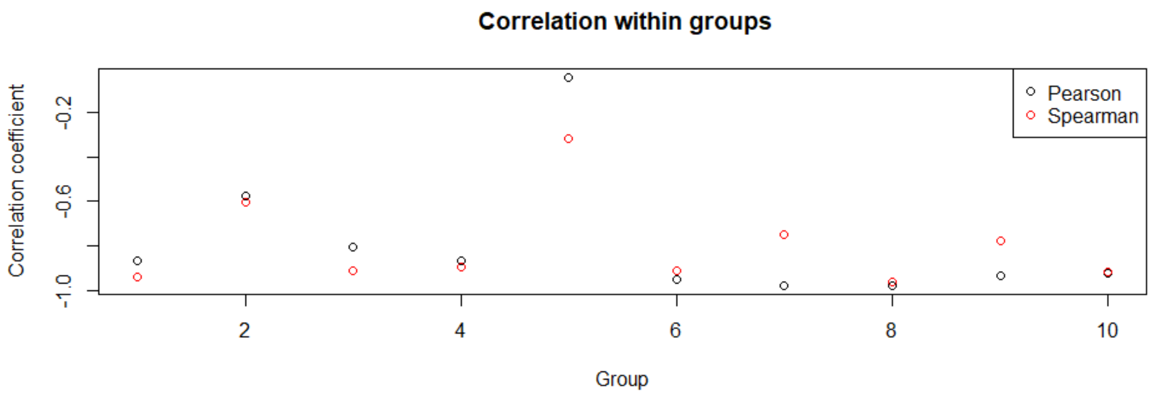


Figure 11: Correlation between preventive and corrective maintenance where group 7 and 8 is changed to the next biggest group.

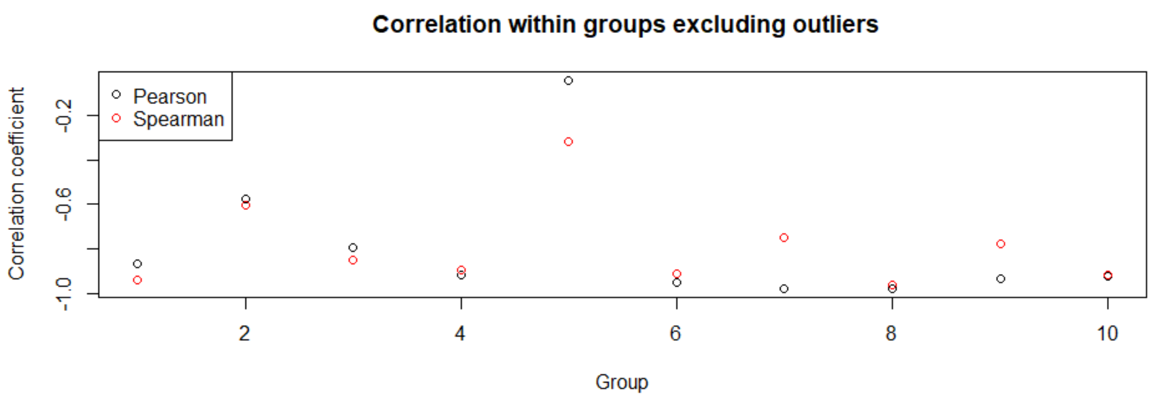


Figure 12: Correlation between preventive and corrective maintenance without outliers and where group 7 and 8 is changed to the next biggest group.

In Figure 11 and 12, we notice that the correlation is negative for both Spearman's and Pearson's correlation coefficients for all groups. There is not a big difference between the resulting two correlation coefficients, except for group number 5 and 7. Since Spearman's correlation coefficient is less restrictive compared to Pearson's, we focus on Spearman's correlation for group number 5 and 7. The correlation is not as strong for group number 5 but for the other groups the correlation is quite strong. The negative correlations shown in Figure 11 and 12 suggest that increasing number of preventive maintenance made on the asset will decrease the number of corrective maintenance that has to be done to the asset. The result shown in Figure 11 and 12 differs a lot from the result in Figure 6. An explanation to this can be how the data is represented. In Figure 11 and 12, the observations used to calculate the correlation include the mean number of maintenance made per year. In Figure 6 the observations include the maintenance made for one asset from start of life and up until now and in some cases the number of maintenance made during an entire lifetime. To some extent, this result answers our primary question but does not give any input to our secondary questions.

4.3 Regression Analysis

Next, we present the results for the regression analysis. In this part of the study, we removed the features not included in the model formulation by backward selection.

Model 1

Formulation after backward selection:

$$E[CM] = \beta_0 + \beta_1 PM + \beta_2 time\_alive + \sum_{i=2}^{10} \beta_{i+1} g_i$$

where g_i is equal to 1 if the observation belong to group number i otherwise it is equal to 0 and β_i is the corresponding coefficient for the respective group.

Table 7: Result of model 1.

Model 1	
β_0	-0.9
β_1	0.22
β_2	0.0005
β_3	1.1
β_4	1.6
β_5	-0.6
β_6	0.12
β_7	0.63
β_8	0.1
β_9	0.19
β_{10}	1.5
β_{11}	0.51
Adjusted R^2	0.3108
AIC	11875
RSE	1.774
R^2- *95% CI*	(0.2744, 0.3442)

In Table 7 the coefficient for the preventive maintenance in model 1 is positive. The bootstrap with its 95 % confidence interval of the variance explained is low at only 27.44 to 34.42 %. A rule-of-thumb is that the R^2 should be at least 50 % for the model to be relevant for explaining the observations.

Model 2

Formulation after backward selection:

$$E[CM\_cum] = \beta_0 + \beta_1 PM\_cum + \beta_2 dif\_pmfu + \beta_3 time + \sum_{i=2}^{10} (\beta_{i+2} g_i) + \beta_{13} PM\_cumul^2 + \beta_{14} dif\_pmfu^2 + \beta_{15} time^2$$

Table 8: Result of model 2.

Model 2	
β_0	-0.29
β_1	0.35
β_2	-0.57
β_3	0.00037
β_4	0.28
β_5	-1.3
β_6	-1.2
β_7	0.39
β_8	-0.88
β_9	-0.52
β_{10}	-0.69
β_{11}	-2.02
β_{12}	-0.88
β_{13}	-0.015
β_{14}	-0.006
β_{15}	0.000000035
Adjusted R^2	0.3469
AIC	18803
RSE	1.589
R^2 - 95% CI	(0,3188, 0,3750)

In Table 8 the coefficient for the preventive maintenance in model 2 is positive. The bootstrap with its 95 % confidence interval of the variance explained is low at only 31.88 to 37.5 %. Although the R^2 value is better for model 2 compared to model 1, the AIC value is bigger for model 2 indicating that model 1 is a better fit. The RSE is lower and hence better in model 2. Even if the models can be compared, the variance explained in the models are very low and the regression models does not explain the observed observations very well. This is an indication that there is not enough information from the observed observations to explain the response variable.

Model 3

Formulation after backward selection:

$$E[\bar{CM}] = \beta_0 + \beta_1 \bar{PM} + \beta_2 year + \sum_{i=2}^{10} \beta_{i+1} g_i + \beta_{12} \bar{PM}^2 + \beta_{13} year^2$$

Table 9: Result of model 3.

Model 3	
β_0	0.98
β_1	-0.97
β_2	0.034
β_3	0.15
β_4	-0.24
β_5	-0.15
β_6	-0.046
β_7	0.049
β_8	-0.31
β_9	-0.061
β_{10}	-0.098
β_{11}	-0.053
β_{12}	0.23
β_{13}	-0.0020
Adjusted R^2	0.7106543
AIC	-18.38655
RSE	0.2092
R^2 *95% CI*	(0.5265, 0.7949)

In Table 9 the coefficient for the preventive maintenance in model 3 is negative. The bootstrap with its 95 % confidence interval of the variance explained is between 52,65 and 79.49 %. The R^2 value is a lot better for model 3 compared to model 1 and 2. The AIC value is a lot smaller for model 3 indicating that the model is a better fit for the data. The RSE is lower and hence better in model 3. This model might explain a lot better but it estimates the mean number of corrective maintenance's during a year hence we do not get a prediction for one specific asset. This model estimates the mean number of corrective maintenance during a year within a group. This could indicate that there are outliers in the data that make it difficult to estimate the corrective maintenance. However, when removing the outliers in model 1 and 2, the coefficient corresponding to the preventive maintenance feature is still positive.

From model 3 we can see that the coefficient corresponding to the preventive maintenance is -0.97. This model suggests that the expected mean number of corrective maintenance that has to be done to the assets in the group during a year decrease with approximately one maintenance in average if there is an average increase of 1 in number of preventive maintenance made to the assets in the group. The difference in PMFU feature was not significant in model 3, hence there is no contribution from model 3 to answering the secondary question about the PMFU difference.

4.4 Survival Analysis

Finally, in this section, we present our results for the survival analysis. We will first present the results for Kaplan-Meier analysis, then we will discuss the Cox proportional hazards model.

4.4.1 Kaplan-Meier

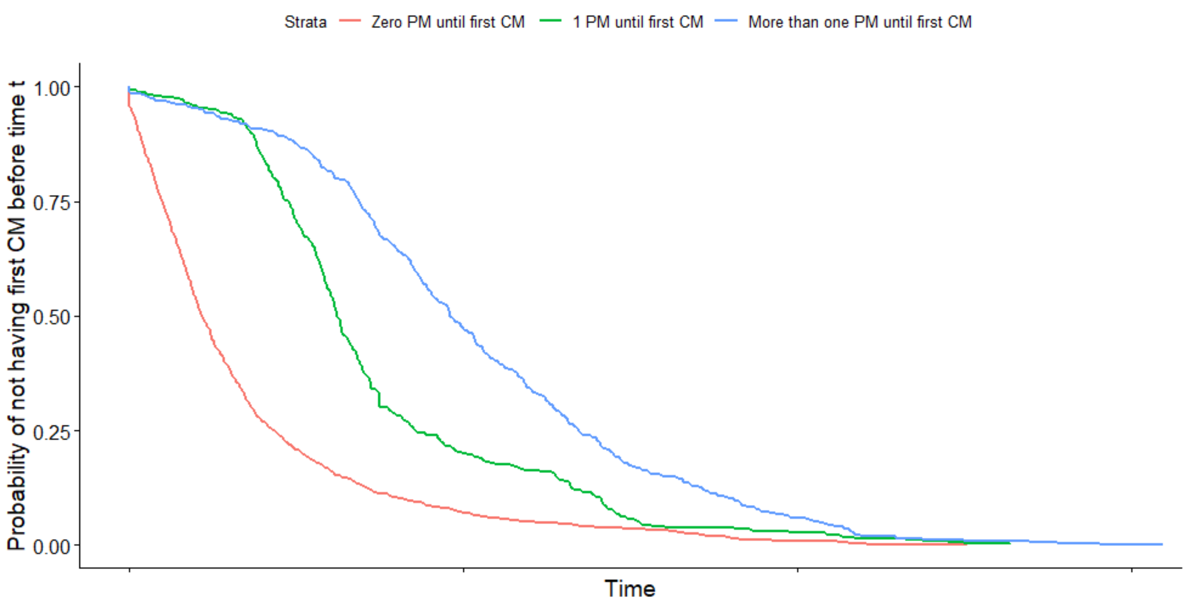


Figure 13: Kaplan-Meier, probability of not having first corrective maintenance made before time t, on three groups with none, one and more than 1 number of preventive maintenance made on the asset.

In Figure 13 the Kaplan-Meier analysis is visualised. The probability of an asset having its first corrective maintenance after time t is estimated using three different groups: One group includes observations where there have not been a preventive maintenance done to the asset; another one includes observations where there has been one preventive maintenance made on the asset; and the third group includes observations where there has been more than one preventive maintenance made on the asset. In Figure 13 we can see that the probability of an asset not needing corrective maintenance is higher when the more preventive maintenance is done to it. One conclusion from this figure is that the more preventive maintenance done to the asset, the more time will pass for the asset until it needs corrective maintenance.

4.4.2 Cox proportional hazards model

In this section, we present the result from the Cox proportional hazards model together with the validation of the assumptions. There are two Cox proportional hazards models, one model with total number of maintenance included as feature variables and one with maintenance ratio included as feature variables. The first obtained model is

Model formulation lifetime: $h(t) = h_0(t)e^{-0.195PM-0.431CM+0.0896dif\_pmfu}$

Since the proportional hazards assumption, that the hazard functions are proportional over time, is not fulfilled when creating a Cox model based on total number of maintenance during a lifetime, a second Cox model is created using the ratios instead. The proportional hazards assumption is checked using Schoenfeld residuals and is visualized in Figure 18. Since the ratios are divided by time, the hazard functions form this model might be proportional over time and the assumption can be fulfilled. We get the following formulation for this model: Model formulation Ratio: $h(t) = h_0(t)e^{-414.3PM\_ratio-1105CM\_ratio+0.07863dif\_pmfu}$

In both models, the coefficients for the preventive- and corrective maintenance is negative sugesting that the more maintenance made to the asset the probability of it living longer increase. The coefficient for the difference in PMFU value is positive for both models and hence the model suggest that the more the hospital deviate from the national prioritisation scale the probability of assets lifetime will decrease.

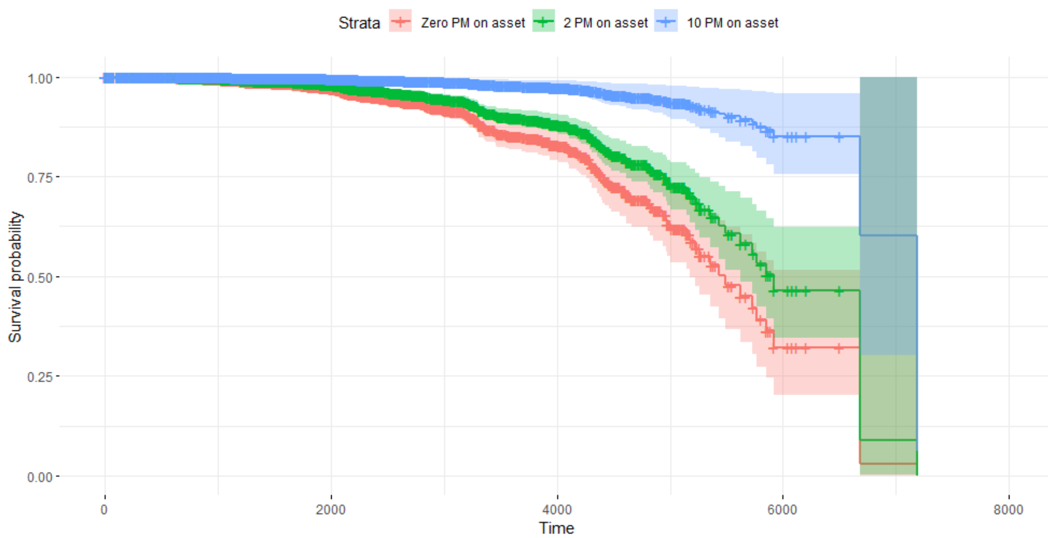


Figure 14: Cox proportional hazards model, probability of survival. Estimate of survival with three observations with different number of preventive maintenance during a lifetime for one asset

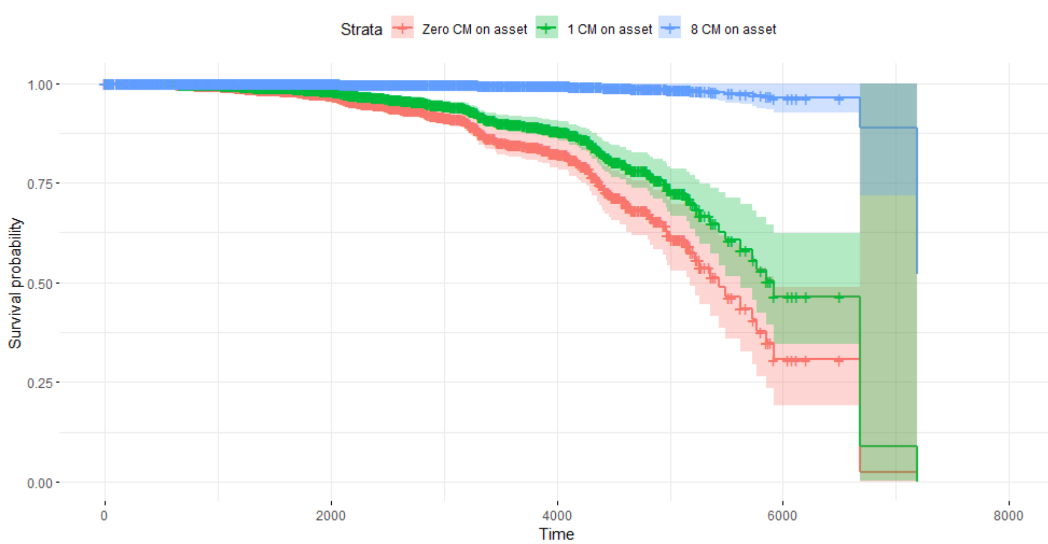


Figure 15: Cox proportional hazards model, probability of survival. Estimate of survival with three observations with different number of corrective maintenance during a lifetime for one asset.

In Figure 14 and 15 the estimate of the probability of surviving for an asset is displayed. In these Figures, the estimate comes from the Cox proportional hazards model with maintenance made during a lifetime. In Figure 14 an investigation of how the preventive maintenance affects the lifetime is made. This is made by using three different observations to make a prediction. The PMFU difference and the number of corrective maintenance are 1.3 and 1 respectively for all three observations. These values are selected since it is the mean of corrective maintenance and PMFU value. The preventive maintenance differs for the three observations where the first one has had no preventive maintenance made to it, the second has had 2 preventive maintenance's and the last has had 10 preventive maintenance's made to the asset during its lifetime. In Figure 15 an investigation of how the corrective maintenance affects the lifetime is made. This is made in the same way as for when investigating the preventive maintenance affect on the assets lifetime but with the mean of all the preventive maintenance as the number of preventive maintenance made to the asset. The three observations to make a prediction on has a varying number of corrective maintenance of 0, 1 and 8. Both plots in Figure 14 and 15 shows a confidence interval of 95 %.

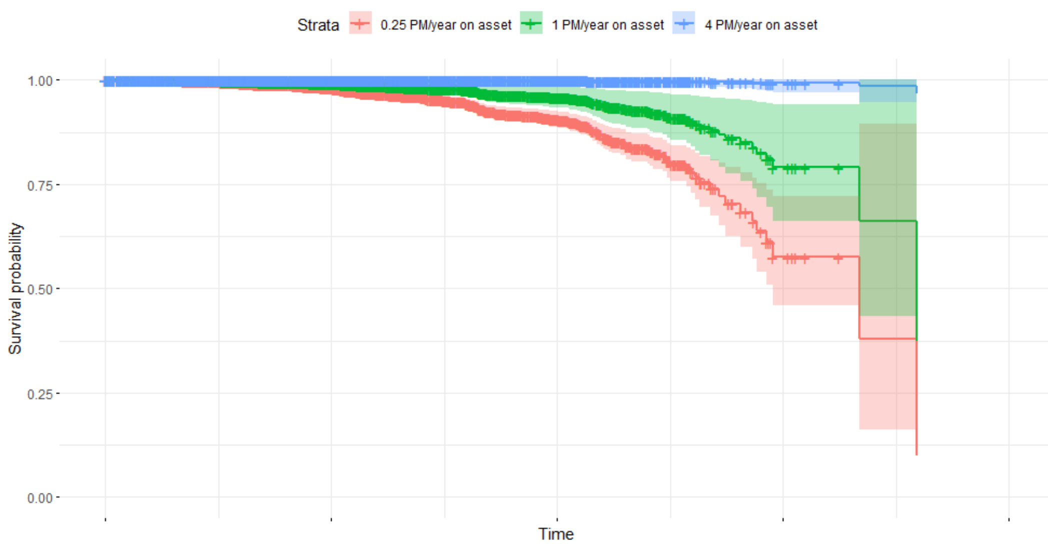


Figure 16: Cox proportional hazards model, probability of survival from ratio model formulation. Estimate of survival with three observations with different number of preventive maintenance during a lifetime for one asset.

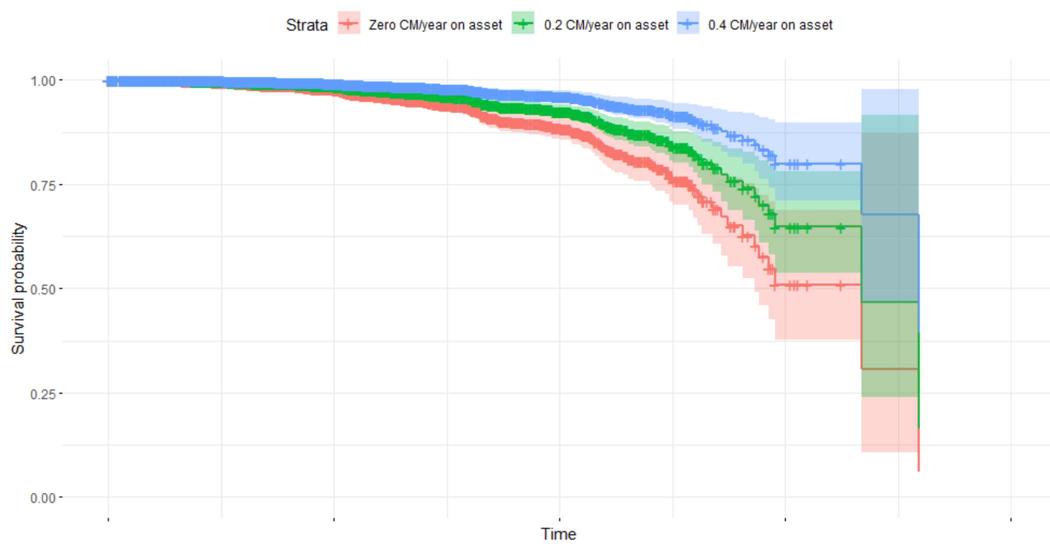


Figure 17: Cox proportional hazards model, probability of survival from ratio model formulation. Estimate of survival with three observations with different number of corrective maintenance during a lifetime for one asset.

In Figure 16 and 17 the estimate of the probability of surviving for an asset is displayed. In this Figure the estimate comes from the Cox proportional hazards model with average maintenance during one day. In Figure 16 an investigation of how the preventive maintenance affects the lifetime is made. This is made by using three different observations to make a prediction. The PMFU difference and the corrective maintenance ratio is 1.3 and 0.0003 respectively for all three observations. These values are selected since it is the mean of the corrective maintenance ratio and PMFU value. The preventive maintenance differs for the three observations where the first one get in average a preventive maintenance made to it every forth year, the second has get in average one preventive maintenance's made to it every year and the last has get 4 preventive maintenance made to the asset during its lifetime. In Figure 17 an investigation of how the corrective maintenance ratio affects the lifetime is made. This is made in the same way as for when investigating the preventive maintenance ratio affect on the assets lifetime but with the mean of the preventive maintenance ratio as the preventive maintenance ratio made to the asset. The three observations to make a prediction on has a varying corrective maintenance ratio of 0, 0.2/year and 0.4/year. As Figures 14 and 15, both plots in Figure 16 and 17 shows a confidence interval of 95 %. Since there are many censored observations in the data set the confidence interval grows with time. This is visible in Figures 14, 15, 16 and 17.

Validation of assumptions

Finally, we calculate the Schoenfeld residuals and the Martingale residuals to validate the assumptions for Cox proportional hazards model empirically. This is done on both models to verify the assumptions.

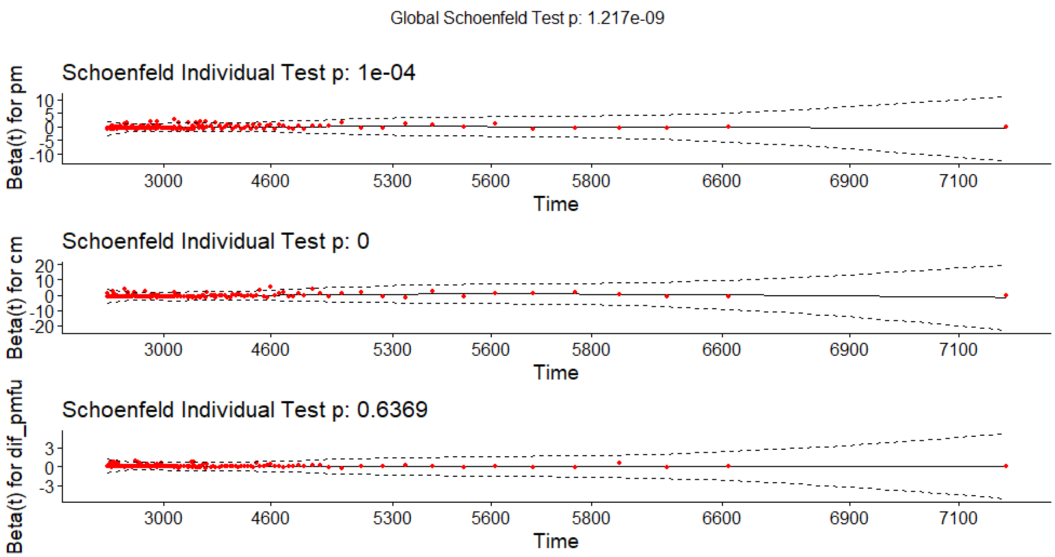
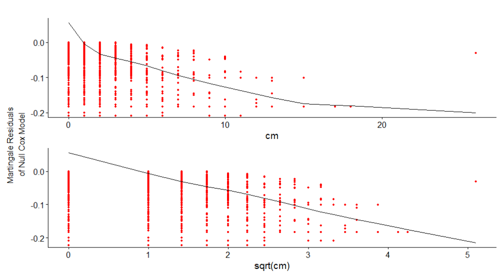
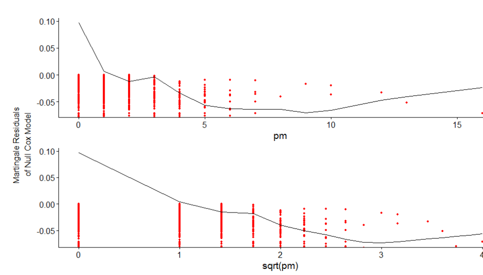


Figure 18: Schoenfeld residuals of corrective- and preventive maintenance and the difference in the PMFU value. The p-value for the test score is displayed for each variable ans well as the global test-score.

In Figure 18, the Schoenfeld residuals and the p-values are displayed for the model with number of maintenance made to the asset. Note that the p-value is smaller than 0.05 for the preventive- and corrective variable as well as on the model globally. This means that the proportional hazards assumption is violated and that the preventive- and corrective maintenance variable and the model globally do not have hazard functions that are proportional over time. Therefore the assumptions of a Cox proportional hazards model is not fulfilled.



(a) Martingale residuals of corrective maintenance



(b) Martingale residuals of preventive maintenance

Figure 19: Martingale residuals and square root martingale residuals of PM and CM.

Figure 20: Martingale residuals of the difference in PMFU value.

Figure 19 shows the martingale residuals and the square-root of the martingale residuals of the corrective- and preventive maintenance. Since the maintenance can only have one maintenance made at a time the variables has a discrete number of maintenance. Because the variables only take integers both variables might look like there is a pattern. The line in the plot suggest that the corrective maintenance fulfil the linearity assumption since the line is almost linear. When examining the preventive maintenance it does look slightly like the linearity assumption is violated. Figure 20 shows the martingale residuals of the PMFU value and it is clear from the line in the Figure that this variable violates the linearity assumption.

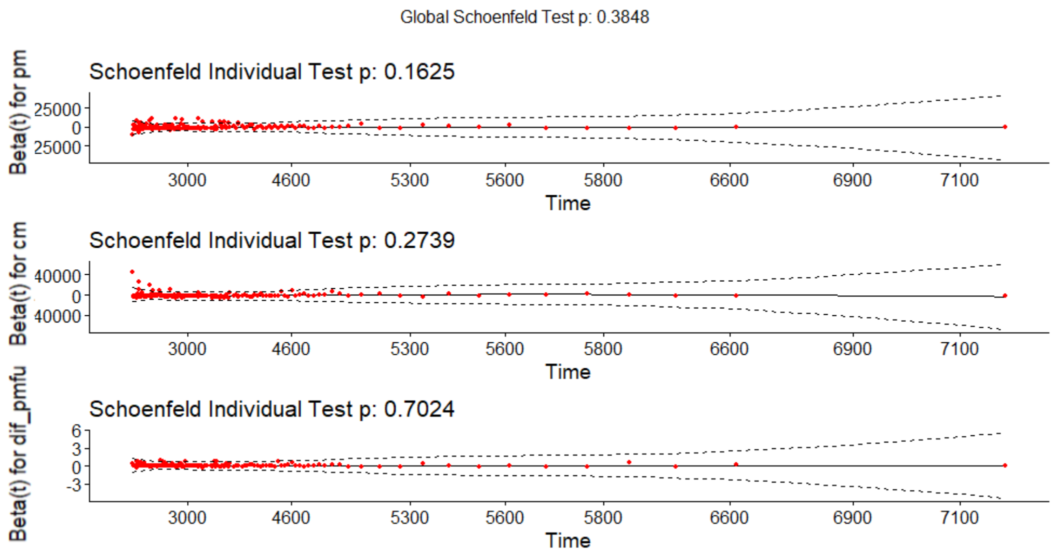
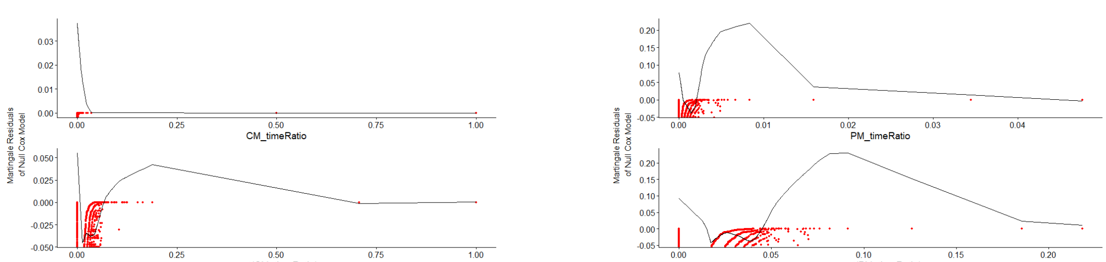


Figure 21: Schoenfeld residuals of corrective- and preventive maintenance ratios and the difference in the PMFU value. The p-value for the test score is displayed for each variable ans well as the global test-score.

In Figure 21 the Schoenfeld residuals and the p-values are displayed for the model with maintenance per day in average that is made to the asset. The p-value is greater than 0.05 for all variable as well as on the model globally. This means that the proportional hazards assumption is fulfilled and that all variables as well as the model globally have hazard functions that are proportional over time.



(a) Martingale residuals of corrective maintenance ratio.

(b) Martingale residuals of preventive maintenance ratio.

Figure 22: Martingale residuals and the square root martingale residuals of PM ratio and CM ratio.

Figure 22 shows the martingale residuals and the square-root of the martingale residuals of the corrective- and preventive maintenance ratios. In all four plots in the Figure, we can see that there seems to be a pattern for both preventive- and corrective maintenance which suggest that the variables violate the linearity assumption. It can be confirmed by the nonlinear curves in the plots. The martingale residuals of the PMFU difference is the same as in the first model, see Figure 22b. The model with maintenance ratios fulfil the proportional hazards assumption but not the linearity assumption.

5 Discussion and Conclusions

Correlation Analysis

From the correlation analysis, we can conclude that when looking at each asset individually the correlation varies a lot depending on which group the asset belongs to. The correlation has a wide spread and there is no apparent negative or positive correlation between the preventive- and corrective maintenance. When assessing the problem more generally and dividing the maintenance into years, the results are a bit more clear. Both Spearman's and Pearson's correlation coefficients are negative for all groups, suggesting that the more preventive maintenance made to the asset, the less corrective maintenance has to be done to that asset. However, the correlation between these two variables can have another explanation and does not necessarily mean that there is a relationship between the variables. The correlation can be found to be influenced by another variable. Also the fact that the results differ when calculating the correlation using each asset and divided into average per year suggests that the results might be unreliable. The regression analysis confirms these differing results between the methods.

Regression Analysis

Regression model 1 and 2 suggest that an increasing number of preventive maintenance made to the asset results in more corrective maintenance made to the asset. Although, both model 1 and 2 do not explain the data well since the variance explained by the models is approximately 30 % for both models. A rule-of-thumb is that a model should have a R^2 value of at least 50 % for the model to be considered explaining the data. When approaching the regression analysis in the same way as in the correlation analysis, we get similar results. The difference is that we can get a measure of how well the model explains the data. The third regression model explains the data much better than model 1 and 2, which might not be that surprising. Intuitively, it is easier to predict the average number of corrective maintenance per year and per group since this is a more general approach. From the correlation- and regression analysis, the primary question has been answered in a general manner. In this thesis, nothing suggests that the corrective maintenance will decrease with the more preventive maintenance made to a specific asset. However, when looking at the average maintenance in a group we can conclude, from the correlation and regression analysis, that there is a relationship between the corrective- and preventive maintenance. This relationship suggests that the more number of preventive maintenance made on average in a group, the less number of corrective maintenance has to be done to the assets in average for the group.

Survival Analysis

The Kaplan-Meier analysis suggests that the more preventive maintenance on the asset, the longer it will take until the asset needs any corrective maintenance. This finding is intuitive and confirms that there is a relationship between the preventive- and corrective maintenance. The resulting Cox models present that increasing the number of preventive maintenance made to the asset will increase the expected lifetime of the asset. An interesting result for the Cox models is that it suggests that increasing number of corrective maintenance made to the asset will increase the expected lifetime for the asset. This is intuitive since the corrective maintenance is still work made on the asset. However, none of the Cox models fulfilled the assumptions needed when fitting a Cox proportional hazards model. These violations suggest that the hazard functions are non-linear and that there is a variation in the maintenance made to the asset depending on how long time the asset has been alive. An improvement to this work would be to fit an extended Cox model instead. This might solve the problem with time-depending covariates. Due to the time constraint, a extended Cox model was never completed, but for future work this is a possible method to answer the questions more thoroughly. See Section 2.3.1 for using the extended Cox model to solve the problem. Even though the results from the survival analysis is in line with what was expected before the start of the thesis, the models do not explain the observations very well and there might be a better way to estimate the lifetime of the assets.

When examining how the deviation from the national prioritisation scale (PMFU) affects the number of corrective maintenance made to the asset, the variable is only significant for one of the regression models. This model does not explain the observations very well. The survival analysis shows that the expected lifetime is shortened if the medical technicians deviate from the national prioritisation scale and rate an asset lower than the national prioritisation scale. But as mentioned, the proportional Cox model does not fulfil its assumptions and does not have a reliable result.

Conclusions

The primary question, whether preventive maintenance results in less corrective maintenance, is largely answered affirmatively. The secondary question, whether the deviation from national prioritisation of preventive maintenance affects the number of corrective maintenance or deviations, has been examined. Still, the results are not reliable, and therefore this needs to be further investigated. The secondary question on whether recurring preventive maintenance results in less deviations in healthcare has not been investigated due to the time constraints. In conclusion, this work has shown significant results and suggests that there is a significant relationship between corrective- and preventive maintenance but further investigation is needed to confirm this. For future work, an extended Cox model is suggested as a possible approach to show how the PMFU value and maintenance affect the expected lifetime of the asset.

References

Che-Ani, A.I. and Ali, R. 2019. *Facility management demand theory: Impact of proactive maintenance on corrective maintenance.* Journal of Facilities Management, Vol. 17 No. 4, pp. 344-355. Available at: https://doi-org.proxy.ub.umu.se/10.1108/JFM-09-2018-0057.

Cox, D.R. 1972. *Regression Models Life-Tables.* Journal of the Royal Statistical Society. Series B (Methodological) , 1972, Vol. 34, No. 2 (1972), pp. 187-220. Available at: https://www.jstor.org/stable/2985181?seq=1. Accessed: 24 May 2022.

Everitt, B. S.., 2002 *The Cambridge Dictionary of Statistics*, Cambridge University Press. Avalible at: http://ebookcentral.proquest.com/lib/umeaub-ebooks/detail.action?docID=217871. Accesed: 2022-04-12

Fisher, Lloyd D. and Lin, D. Y. 1999. *TIME-DEPENDENT COVARIATES IN THE COX PROPORTIONAL-HAZARDS REGRESSION MODEL.* Annual Review of Public Health. Availible at: https://www.annualreviews.org/doi/full/10.1146/annurev.publhealth.20.1.145. Accesed 22 April 2022.

Frank E. Grubbs. *Sample Criteria for Testing Outlying Observations.* Ann. Math. Statist. 21 (1) 27 - 58, March, 1950. Available at:https://doi.org/10.1214/aoms/1177729885.

Gogtay, N and Thatte, U. 2017. *Principles of Correlation Analysis.* The Journal of the Association of Physicians of India. Available at:https://www.researchgate.net/publication/327475257_Principles _of_Correlation_Analysis/citation/download. Accessed: 23 March 2022.

Haitao Liao, Wenbiao Zhao and Huairui Guo. *Predicting Remaining Useful Life of an Individual Unit Using Proportional Hazards Model and Logistic Regression Model.* Annual Reliability and Maintainability Symposium, 2006., 2006, pp. 127-132. Available at: https://ieeexplore.ieee.org/document/1677362. Accessed: 26 January 2022.

Hashim, D and Weiderpass, E 2019. *Cancer Survival and Survivorship.* Encyclopedia of Cancer (third edition). Avalible at: https://doi.org/10.1016/B978-0-12-801238-3.65102-4.

Hartina Husain et al. 2018. *The Application of Extended Cox Proportional Hazard Method for Estimating Survival Time of Breast Cancer.* Journal of Physics: Conference Series. Availible at https://iopscience.iop.org/article/10.1088/1742-6596/979/1/012087/pdf. Accesed 18 april 2022.

Huang, HZ., Wang, HK., Li, YF. et al. 2015 *Support vector machine based estimation of remaining useful life: current research status and future trends.* J Mech Sci Technol 29, 151–163. Available at: https://doi.org/10.1007/s12206-014-1222-z.

J. Zhu, N. Chen and W. Peng,. 2019. *Estimation of Bearing Remaining Useful Life Based on Multiscale Convolutional Neural Network.* IEEE Transactions on Industrial Electronics, vol. 66, no. 4, pp. 3208-3216, Available at : 10.1109/TIE.2018.2844856. Accessed: 26 January 2022.

Kaplan, E. L., and Meier, P. 1958. *Nonparametric Estimation from Incomplete Observations.* Journal of the American Statistical Association 53, no. 282): 457–481. Available at: https://doi.org/10.2307/2281868.

Laerd, S., n.d. *Spearman's Rank-Order Correlation - A guide to when to use it, what it does and what the assumptions are.* Statistics.laerd.com. Available at: <https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php>. Accessed: 2022-03-03.

Link, Carol L. *Confidence Intervals for the Survival Function Using Cox's Proportional-Hazard Model with Covariates.* Biometrics, vol. 40, no. 3, 1984, pp. 601–09, https://doi.org/10.2307/2530904.

NCSS.com. n.d. *Cox Regression.* NCSS Statistical Software. Availible at:https://ncss-wpengine.netdna-ssl.com. Accessed: 8 June 2022.

Mohammed, S. 2019. *Introduction to Survival Analysis using R.* Workshop on Computational Biostatistics and Survival Analysis online.Tata Memorial Centre Advanced Centre for Treatment, Research and Education in Cancer, India. Availible at: https://shariq-mohammed.github.io/files/cbsa2019/1-intro-to-survival.html. Accessed: 5 of May 2022. Date lecture delivered: 23 of December 2019.

Shapiro, S. S., and M. B. Wilk. *An Analysis of Variance Test for Normality (Complete Samples).* Biometrika, vol. 52, no. 3/4, 1965, pp. 591–611. Avalible at:https://doi.org/10.2307/2333709. Accessed: 23 May 2022.

Upton, G. and Cook, I., 2014. *A dictionary of statistics.* 3rd ed. Oxford: Oxford University Press.

Appendix

Table A

Work type	Description
100	Planing
200	Acquisition
300	Delivery
400	Administrative administration
401	Service agreement
402	System management
500	Preventive Maintenance
501	Quality control
502	Random check
601	Training of healthcare staff
603	Corrective maintenance
604	Product-/system adaptation
605	Self-manufacturing
606	Rellocation
607	Deviation investigation in MTP healthcare
608	Material delivery
609	User support and consultation
611	Teknique round
700	Settlement
810	Research
820	Development
850	Clinical routine
860	Clinical development
861	Optimization X-ray physics
900	Internal activities
901	Competence development
902	Work environment and health-promoting workplace
903	Quality work
907	Deviation and customer complaints
909	Own equipment
910	Projects and assignments according to routine