Stockholms universitet

Comparing Time to Event Models for Sparse High Dimensional Medical Questionnaire Data of Rheumatoid Arthritis.

Gustaf Randén

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2026:7
Matematisk statistik
Juni 2026

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematiska institutionen

Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2026:7**
http://www.math.su.se

Comparing Time to Event Models for Sparse High Dimensional Medical Questionnaire Data of Rheumatoid Arthritis.

Gustaf Randén*

June 2026

Abstract

In this study, three survival models were used to predict the likelihood to develop Rheumatoid Arthritis. The methods compared were the Cox regression, Weibull regression and Random Survival Forest. They were tested on a Questionnaire study answered by 79 individuals. The answers were stored in a matrix with 475 variables for each individual. Pre-processing the data included usage of principal component analysis. The prediction models were validated by the Concordance index, Brier's score, a Confusion matrix and the Receiver Operating Characteristic curve. The results indicated that Cox and Weibull models could more accurately predict if a specific person would develop rheumatoid arthritis. However the Random Forest had the highest specificity, which means that it could predict which patient would not develop arthritis. The Random Forest model could more accurately rank who is more likely to develop the disease in three years even if their probability estimates were worse.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: gus.randen@gmail.com. Supervisor: Chun-Biu Li.

Acknowledgments

I would like to thank my supervisors for aiding me throughout the process of writing my thesis. I thank Associate Professor Chun-Biu Li from the department of Mathematics at Stockholm university. I would also like to thank Associate Professor Helga Westerlind and PhD and project coordinator Marina Dehara from the division of clinical epidemiology at Karolinska Institute.

List of notations.

$B(t)$: Briers score at time t

C_X : Covariance matrix of centred covariates X

C_Y : Covariance matrix of the latent lower dimensional representation Y

c_i : An estimator associated with region SR_i given by a decision tree

d_k : Number of events occurring at survival time k.

E : The expected value of a random variable.

$G(t)$: A Kaplan Meier estimator for the probability for an observation to not have been censored before time t

$dH(t)$: An increment of the Nelson-Aalen estimator.

$H(t)$: The cumulative hazard. It is the integrated hazard rate.

$h(t)$: The hazard rate. The instantaneous probability for an event to occur, given that it has not occurred yet.

$I\{x \in SR_i\}$: An indicator variable which equals 1 if a vector of covariates x is in subregion SR_i

IBS : the Integrated briers score.

$L(\beta)$: A likelihood function.

N_{tree} : The number of trees the Random Forest Model creates

SR_i : Region i of the covariates used by a decision tree

$S(t)$: The Survival Function. The probability for an observation to not register the event by time t.

T_i : The time to event or survival time for observation i.

$T_{(k)}$: The k:the smallest survival time.

t : Notation for arbitrary timepoint

W : A random variable following the standardized extreme value distribution, se equation (28).

w : an variable representing the logarithm of a specified time point

$w(t)$: A weight of the inverse probability to have been censored for each observation with known outcome.

Y : The latent lower dimensional representation of covariates X assumed by PCA

$Y_i(t)$: The outcome states for observation i at time t. It equals 1 if the event has yet to occur and 0 otherwise.

$Y_*(t)$: The number of observation which still has to experience the event at time t.

X : A matrix of covariates/dependent variables.

x_i : The covariates for an observation i.

x : A variable for an observation with unspecified values for the covariates. The variable can be assigned later.

β : Coefficients in Cox and Weibull regressions.

δ_i : An indicator variable indicating if the event for observation i was observed. If observed, it equals 1, otherwise 0.

$\pi(i;t)$: Conditional probability of observing an event at time t for an individual i given the past and that an event occurs at that time.

ρ_{tree} : The correlation between identically distributed decision trees

σ^2_{tree} : The variance of a single decision tree

1 Introduction

In statistical studies, it is often of interest to study data concerning the time until an event occurs (Aalen et al., 2008). The event could be anything of interest. It could be time from symptom to diagnosis, time until development of disease or time until death. Data from such situations contain mixtures of observations where the event of interest did happen and others where it did not occur. If it did not occur, the observation is said to have been right censored. Ordinary statistical models cannot handle this type of data. For instance, calculating a mean survival time is not possible, because in many observations, the event occurs after the end of the study. Ignoring the censored observations will cause a multitude of issues. For example, the mean time will be underestimated since censoring is biased towards the longer event times.

To handle these issues, statistical models, often called survival models, have been developed. The name comes from their association with epidemiology and time until a patient dies. *Cox regression* is one of the most common survival models in medical research, as it can be used to examine how risk factors influence the time until an event of interest occurs (Cox, 1972). In contrast to the *Weibull regression* (Kalbfleisch & Prentice, 2011, 41-42), it does not require knowledge of the distribution of the event times. Otherwise they are the same. *Random Survival Forest regression* (Ishwaran et al., 2008) predicts the time until the event by choosing the average of a group of decision trees. In this thesis, Cox regression, Weibull regression and Random Survival Forest are compared. Four evaluation techniques for them are specified. They are the Concordance index (Harrell Jr et al., 1982), Brier scores (Gerds & Schumacher, 2006), Confusion matrix and the receiver operating characteristic (ROC) curve (Agresti, 2013, 223-225).

The models were applied to data from a questionnaire study regarding patients with symptoms of rheumatoid arthritis (RA)

(Knevel et al., 2022). RA is a chronic autoimmune disease with symptoms like pain, swelling, stiffness and skin conditions, specifically around the large joints (Smolen et al., 2018) The disease can lead to severe reduction in physical functions and, while rarely fatal, it can negatively impact quality of life. RA affects approximately 1% of the Swedish population and is more in women than men. The onset of the disease in Sweden is typically between ages 45 and 65. Both genetic predisposition and smoking are well-established risk factor for developing the disease. Although RA often progresses slowly, early treatment is important to prevent permanent joint damage. The main treatment options include disease modifying anti-rheumatic drugs (DMARDs) corticosteroids are classical medical treatments

As mentioned earlier, this study is based on a previous study by (Knevel et al., 2022). They developed a questionnaire and based on the answers they received in the questionnaire, they gave each individual a risk score deciding if they should a) seek a general practitioner or b) a rheumatologist. The results were validated in three ways. 1) They conducted a Wilcoxon rank test between the scores of the patient diagnosed with RA and those who were not. It gave a statistically significant difference, $p < 0.0001$. 2) A confusion matrix of which patients' risk scores reached two different thresholds or not was studied. 3) They also created a ROC curve for the scores. The area under the curve (AUC) was 75.3%. (Knevel et al., 2022) concluded that the scoring method needed to be further improved.

The aim of this master thesis is to evaluate the three mentioned survival models, Cox, Weibull, and Random Survival Forest and their application to the RA study.

The second section of the thesis contains theory behind the survival models and their evaluation techniques. The third section describes the dataset and how it was processed. In the fourth section the results are presented. Lastly, the result of the study is discussed.

2 Theory

In this section, Cox regression (Cox, 1972), Weibull regression (Kalbfleisch & Prentice, 2011) and Random Survival Forest (Ishwaran et al., 2008) are defined and explained. The validation methods for the survival models, the Concordance index (Harrell Jr et al., 1982), Brier scores (Gerds & Schumacher, 2006), Confusion matrix and the receiver operating characteristic (ROC) curve (Agresti, 2013, 223-225). are also discussed. First notations and concepts of time-to-event data are defined and described.

2.1 Basics on survival models

The following example illustrates properties of time-to-event data. Imagine a lightbulb company which wants to study the lifetime of their lightbulbs. The idea is to collect data from five new bulbs over a period of ten years. They are not turned on at the same time. The example data is visualized in Figure 1 and 2.

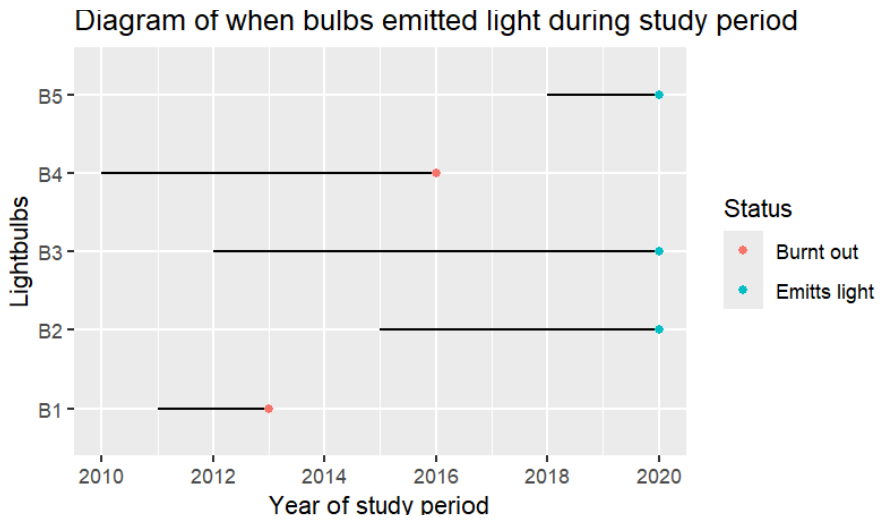

Figure 1: This diagram shows when the bulbs emitted light. The x-axis represents the years of the study period. The y-axis represents each lightbulb.

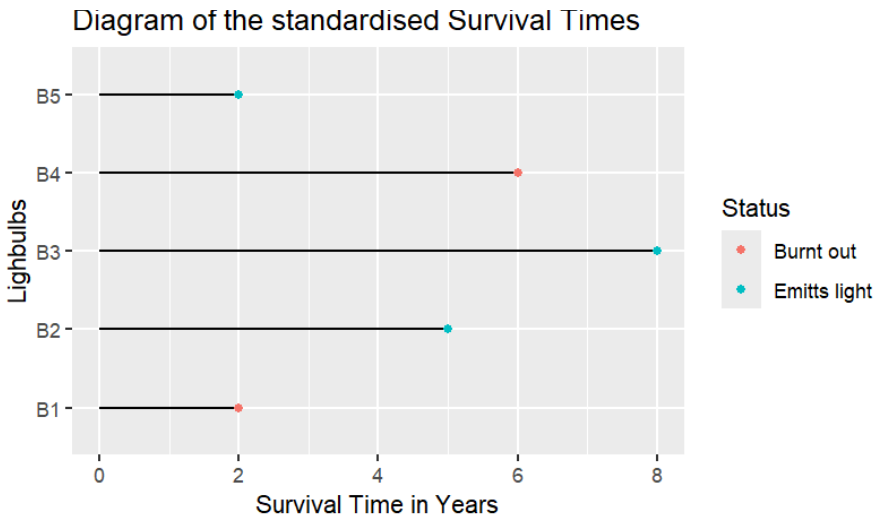

Figure 2: This diagram shows the observed life-lengths of each light bulb. The X-axis represents the survival times in years. The y-axis represents each lightbulb.

Let the lifetime of a bulb i be denoted by the survival time T_i. From the plots, the survival time for each light bulb is

$$T_1 = 2, T_2 = 5^*, T_3 = 8^*, T_4 = 6, T_5 = 2^*.$$

The notation * denotes a censored observation. It is often useful to sort the survival times in increasing order. Let (i) be the i:th shortest time, then the order of lifetimes of the light bulbs is

$$T_{(1)} = 2, T_{(2)} = 2^*, T_{(3)} = 5^*, T_{(4)} = 6, T_{(5)} = 8^*.$$

There are bulbs which still shine when the study period has ended. They could keep working for many more years. In survival analysis, this type of missing information is called right censoring or just censoring. Define δ_i to be an indicator variable representing if an observation was censored. It equals 1 if the event of interest was seen and 0 otherwise. In the example δ_1 and $\delta_4 = 1$ since the failures were observed. Likewise, $\delta_2, \delta_3, \delta_5 = 0$ since the bulbs still worked at the end of the study period. Another useful property is the number of light

bulbs which are still functioning at a specific timepoint of interest t. Let $Y_i(t)$ be an indicator variable for a lightbulb i. It indicates whether the bulb is known to still emit light after t years. In the diagram we see that for time $t = 4$,

$$Y_1(4) = Y_5(4) = 0$$

and

$$Y_2(4) = Y_3(4) = Y_4(4) = 1.$$

These variables can be used to define the total number of bulbs known still working as the sum of the indicator variables.

$$Y_*(t) = \sum_{i=1}^{n} Y_i(t) \quad (1)$$

For an example, when $t = 4$, $Y_*(4) = 3$.

It is usually of interest to estimate the probability for an event to happen after a specified time point. This is done by estimating the Survival function $S(t)$. It is defined to be the complement of the cumulative density function, $F(t)$, for an event to occur by time t.

$$S(t) = P(T > t) = 1 - F(t) \quad (2)$$

In some applications, the event is known to not have occurred by time s. Then the conditional survival function for t given s is described by

$$S(t|s) = P(T > t \mid T > s) = \frac{P(T>t)}{P(T>s)} = \frac{S(t)}{S(s)} \quad (3) \quad (3)$$

The survival function can be described by a product of conditional survival functions. Note that the survival function covers the probability of an event not happening in the time interval $[0, t]$. This interval can be partitioned into smaller time intervals $0 = t_0 < t_1 < \ldots < t_K = t$. For each interval, conditional survival functions $S(t_k|t_{k-1})$ can be created. Multiplying these together yields

$$\prod_{k=1}^{n} S(t_k|t_{k-1}) = \frac{S(t_1)}{S(t_0)}\frac{S(t_2)}{S(t_1)}\cdots\frac{S(t_k)}{S(t_{k-1})} = S(t) \quad (4)$$

This relation is utilized by the Kaplan Meier estimator (Kaplan & Meier, 1958). It assumes that if no event occurred between two timepoints $t_{k-1}$and t_k, then the estimated survival probability $S(t_k|t_{k-1})$ equals 1. A different estimate is needed if instead one or more events occur at time point t_k. The probability of d_k events occurring at time t_k can be estimated by d_k divided by the number of observations still at risk $Y_*(t_{k-1})$. The conditional survival function is then estimated by the probability of the d_k events not occurring. This yields the estimate

$$S(t_k|t_{k-1}) = 1 - \frac{d_k}{Y_*(t_{k-1})} \quad (5)$$

This estimate is substituted into (4). Kaplan Meier assign t_{k-1} to be the k:th shortest unique survival time $T_{(k)}$. Thus, the Kaplan Meier estimator is defined for all $T_k < t$ as below.

$$\hat{S}(t) = \prod_{T_k<t}\left(1 - \frac{d_k}{Y_*(T_k)}\right) \quad (6)$$

Another useful property of survival data is the hazard rate (Aalen et al., 2008, 6). It represents the instantaneous probability for the event to happen, given that it has not occurred earlier. The hazard rate $h(t)$ is related to the survival function and is defined by

$$h(t) = \lim_{\Delta t\to 0}\frac{1}{\Delta t}\frac{S(t)-S(t+\Delta t)}{S(t)} = -\frac{S'(t)}{S(t)}. \quad (7)$$

Many regression models use it instead of the survival function. The cumulative hazard is defined by integrating the hazard rate until a specified time t

$$H(t) = \int_0^t h(s)ds = -log(S(t)) \quad (8)$$

The cumulative hazard has many useful properties. For instance, it can be used to validate parametric models and be used to calculate the Kaplan-Meier estimate. The cumulative hazard can be estimated by the Nelson-Aalen estimator (Nelson, 1969 as cited by Aalen et al., 2008, 70) (Aahlen, 1978)

$$\hat{H}(t) = \sum_{T_k<t} \frac{d_k}{Y_*(T_k)} \quad (9)$$

Each increment $dH(T_k) = \frac{d_k}{Y_*(T_k)}$ in the sum has a corresponding term in the Kaplan-Meier estimator. It can be described as

$$\bar{S}(t) = \prod_{T_k<t} (1 - dH(T_k)) \quad (10)$$

More details and proper proofs of the estimators can be found from standard textbooks on survival analysis such as (Aalen et al., 2010).

The estimators which have been described do not take into account the effect of any predictive variables or risk factors. In this thesis such variables are called covariates. Let $\boldsymbol{x} = (x_1, x_2, ..., x_p)$ denote a vector of p number of covariates. An observation i in a sample of size n is defined by the triplet $(T_i, \delta_i, \boldsymbol{x}_i)$. Here T_i is the observed survival time for observation i, δ_i is the indicator variable for whether the observation was censored and $\boldsymbol{x}_i$ is the observed covariates for the observation. The covariates for all observations is stored in the $n \times p$ covariate matrix X. Row i of the matrix corresponds to the observed covariates $\boldsymbol{x}_i$ for observation i.

2.2 Cox regression model

Here below Cox model is described and derived. This model like the other two mentioned models will later be used on data from the

RA study. The common purpose of using all three survival models in this thesis is to model the relationship between the time of onset of the disease and the answers of the questionnaire.

Cox regression is one of the most used statistical models to analyse time to event data (Cox, 1972). It is used for estimating the effects that the covariates have on the hazard rate. To simplify the interpretation of the effects of the covariates, the model makes some assumptions on the hazard rate. One is that the hazard rate has the form of a baseline hazard scaled by an exponential linear combination of the covariates $\boldsymbol{x}$,

$$h(t; x) = h(t)\exp(\beta^{\top}\boldsymbol{x}). \tag{11}$$

$h(t)$ is the base hazard rate which only depends on the time t. The effect of the covariates is through the scaling of $\exp(\beta^{\top}\boldsymbol{x})$. β is a parameter vector associated with the covariates and does not contain an intercept. Since $\exp(\beta^{\top}\boldsymbol{x})$ does not depend on time, the effect of the covariates is proportional to the base hazard. It is therefore common to refer to the Cox model as a proportional hazards model. There are many reasons why a log-linear model is chosen. One reason is because the exponential function makes sure the hazard rate is positive. Another reason is because the exponential function makes the Cox model easily interpretable. For instance, assume a model with only one covariate x, which means $p = 1$. The sample consists of two observations with covariates x_1 and x_2. The covariates are related through $x_2 = x_1 + 1$. The ratio of their hazard rates is.

$$\frac{h(t)\exp(\beta x_2)}{h(t)\exp(\beta x_1)} = \exp(\beta(x_2 - x_1)) = \exp(\beta) \tag{12}$$

Here $\exp(\beta)$ can be interpreted as a scaling factor for the ratio by a unit increase of covariate x.

To fit the parameters β, regular likelihood methods do not suffice due to the unspecified base hazard. To handle these issues, one can estimate the parameters using the partial likelihood. The

partial likelihood uses the conditional probabilities $\pi(i;t)$ of observing the event at time t for an individual i given the past and that an event occurs at time t.

$$\pi(i;t) = \frac{Y_i(t)\exp(\beta^\top \boldsymbol{x}_i)}{\sum_{l=1}^{n} Y_l(t)\exp(\beta^\top \boldsymbol{x}_l)} \tag{13}$$

The partial likelihood $L(\beta)$ is calculated by each observation i and assigning t the corresponding survival time T_i. It then takes the form

$$L(\beta) = \prod_{i=1}^{n} \pi(i;T_i) = \prod_{i=1}^{n} \frac{Y_i(T_i)\exp(\beta^\top \boldsymbol{x}_i)}{\sum_{l=1}^{n} Y_l(T_i)\exp(\beta^\top \boldsymbol{x}_l)} \tag{14}$$

The parameters are then estimated by finding which β maximises the likelihood. To distinguish these estimated values from other possible values for β, let the maximum likelihood estimate be denoted $\hat{\beta}$. To test if $\hat{\beta}$ is significantly different from a null model β_0 for which $\beta = 0$, the log likelihood ratio statistic can be calculated.

$$\chi^2_{LR} = 2(\log(L(\hat{\beta})) - \log(L(\beta_0))) \tag{15}$$

This statistic is χ^2 distributed with p degrees of freedom.
To compare Cox regression with other likelihood-based models, Akaike's information criterion (AIC) (Hastie et al., 2009, 231) is often used.

$$AIC = -2\log(L(\hat{\beta})) + 2p \tag{16}$$

The model with the lowest AIC is the model with the closest fit.

Even though having estimated the parameters $\hat{\beta}$, it is not possible to assign survival probabilities. This is because we have not specified the base hazard $h(t)$. In Cox regression, it is standard to estimate the cumulative base hazard using the Nelson-Aalen estimator (equation (9)) (Aalen et al., 2008, 141). To include the effect of the covariates, one substitutes $Y_*(T_j) = \sum_{l=1}^{n} Y_l(T_j)$ in

equation (9) with $\sum_{l=1}^{n} Y_l(T_j)\exp(\hat{\beta}\boldsymbol{x}_l)$, yielding the base hazard estimator

$$\hat{H}_0(t) = \sum_{T_j<t} \frac{d_k}{\sum_{l=1}^{n} Y(T_j)\exp(\hat{\beta}\boldsymbol{x}_l)} \tag{17}$$

The full cumulative hazard is achieved by multiplying with the proportional hazards function

$$\hat{H}(t;\boldsymbol{x}) = \hat{H}_0(t)\exp(\beta^{\top}\boldsymbol{x}). \tag{18}$$

With the Nelson-Aalen estimator defined, the survival function can be estimated using the Kaplan-Meier estimator in equation. (6),

$$\hat{S}(t;\boldsymbol{x}) = \prod_{T_j<t} (1 - d\hat{H}(t;\boldsymbol{x})) \tag{19}$$

There are however many disadvantages to estimating the survival function using the Cox model. A major problem is that the hazard rate may not have a proportional base hazard. A second disadvantage is that the true effect of the covariates may not be linear. The Cox model cannot take into account these issues and as a result it may give a poor fit. Thirdly, there can be a problem caused by how the base hazard was estimated. If the model is extrapolated for timepoints outside the range of the observed survival times, the model may not give accurate cumulative hazards and survival probabilities. When predicting for new time points t, the Cox model will use the Nelson-Aalen estimator up to the largest event time trained on less than t. If t is smaller than all the event times, the predicted survival probability will be 1. Also, for *t* being much larger than the largest event time, it will still have the same base hazard as cox regression at the largest event time.

2.3 Weibull regression model

To handle the problems with estimations of the base hazard in the Cox model, parametric models are an alternative. These models

assume the survival function follows a known probability distribution. One commonly used is the Weibull distribution and associated model (Kalbfleisch & Prentice, 2011). The distribution is defined by the following cumulative density function

$$F(T < t) = 1 - \exp(-(\lambda t)^{\gamma}) \tag{20}$$

Equivalently, the survival function is defined by

$$S(t) = \exp(-(\lambda t)^{\gamma}) \tag{21}$$

It is a generalization of the exponential distribution with parameters λ denoted the scale and γ denoted the shape. These parameters can be understood when looking at the hazard rate of the distribution,

$$h(t) = \lambda\gamma(\lambda t)^{\gamma-1} \tag{22}$$

When $\gamma < 1$, the hazard is monotonously decreasing for increasing times t. If $\gamma > 1$ it is instead monotonously increasing. If $\gamma = 1$, the hazard is constant, leaving only λ. The only distribution with a constant hazard is the exponential. This case implies the survival function follows the exponential distribution. λ can thus be interpreted as the equivalent to the rate of an exponential distribution. γ can be understood as a parameter describing how fast the hazard increases or decreases. To model the effect of covariates on Weibull-models, like Cox-regression, the hazard is assumed to be scaled by an exponential linear combination of the covariates.

$$h(t; \boldsymbol{x}) = \lambda\gamma(\lambda t)^{\gamma-1}\exp(\beta^{\top}\boldsymbol{x}) \tag{23}$$

This assumption causes the survival function to take the following form

$$S(t; \boldsymbol{x}) = \exp(-(\lambda t)^{\gamma}\exp(\beta^{\top}\boldsymbol{x})) \tag{24}$$

The parameters β, λ and γ, are more easily estimated from the logarithm of the Weibull distributed event times T. The probability of the logarithmised event times being larger than a log time w is

given by

$$P(\log(T) \geq w; \boldsymbol{x}) = P(T \geq \exp(w); \boldsymbol{x}) = \exp(-(\lambda \exp(w))^{\gamma} \exp(\beta^{\mathsf{T}} \boldsymbol{x})) \quad (25)$$

Substituting the variables to $z = -\log(\lambda \exp(\beta^{\mathsf{T}} \boldsymbol{x} / \gamma))$ and $\sigma = \gamma^{-1}$, we have the survival function to be equal to

$$S(\exp(w); \boldsymbol{x}) = \exp(-(\lambda \exp(\beta^{\mathsf{T}} \boldsymbol{x} / \gamma) \exp(w))^{\gamma}) =$$

$$= \exp(-\exp((w - z)/\sigma)) \quad (26)$$

This distribution is called an extreme value (minimum) distribution. (Kalbfleisch & Prentice, 2011, 33) with parameters z and σ. It has the property of being able to be written as a linear combination of z and σ times a random variable W.

$$\log(T) = z + \sigma W \quad (27)$$

Therefore, the observed log survival times w can be seen as a realisation of W and their linear transformation. The random variable W has the previously mentioned extreme value distribution, but with parameters $z = 0$ and $\sigma = 1$. W has the probability density function

$$f(w) = \exp(w - \exp(w)) \quad (28)$$

reverting the substitution of z, it is possible to write the logarithm of a Weibull distributed random variable T as

$$\log(T) = z + \sigma W = -\log(\lambda) - \boldsymbol{x}\sigma\beta + \sigma W \quad (29)$$

This is similar to a linear regression model. It is possible to treat $-\log(\lambda)$ as an intercept for the model. Furthermore, substitute $\sigma\beta = \gamma^{-1}\beta = \beta^{*}$. Include the intercept among β^{*} and an extra element of value one in $\boldsymbol{x}$.

$$\log(T) = \boldsymbol{x}\beta^{*} + \sigma W \text{ (28)} \quad (30)$$

To fit parameters β^{*} and σ , treat the log Weibull survival time as a realization of the standard extreme value random variable.

$$W = \frac{\log(T) - \boldsymbol{x}\beta^*}{\sigma} \tag{31}$$

Now the parameters can be estimated with the maximum likelihood method using the extreme value distribution. The likelihood needs to be slightly modified to take into account censoring. If the observation was not censored, the probability density function $f(w)$ is used. If not, the cumulative density function $F(w) = \int_{-\infty}^{w} f(u)du$ is used. Utilizing the censoring indicator variable δ_i defined earlier in subsection 2.1, the likelihood is

$$L(\beta^*, \sigma) = \prod_{i=1}^{n} \sigma^{-1} f(w_i)^{\delta_i} F(w_i)^{1-\delta_i} \tag{32}$$

With

$$w_i = \frac{\log(T_i) - \boldsymbol{x}_i\beta^*}{\sigma} \tag{33}$$

When the parameters β^* and σ have been estimated, they can be converted to the original parameterisation of the Weibull model.

Since the Weibull model also uses the maximum likelihood technique, one can, like the Cox model, calculate the likelihood ratio statistic and AIC for it.

The Weibull model shares many of the same disadvantages as the Cox model. Both may fit poorly if the hazard rate cannot be factored in a base hazard and a relative risk function. Both also share the problem that the effect of the covariates is assumed to be linear. The main difference is that the Weibull model defines the base hazard and could in theory extrapolate for unobserved time points. On the other hand, if the base hazard is not monotonously increasing/decreasing with time, then the Weibull model will not be able to detect the change correctly in contrast to the Cox model.

2.4 Random Survival Forest

Random Survival Forest is an adapted Random Forest model for survival data. Random Forest was originally developed as an improvement of single regression and classification decision tree models (Breiman, 2001). An introduction to the general theory of Random Forest is given by (Hastie et al., 2009, 587-604). The advantage of decision trees over linear models is that they can intrinsically handle nonlinear effects of the covariates. A disadvantage is that they often have high variance which contributes to higher prediction error. For building tree models in general, see (Breiman, 2001) (Hastie et al., 2009, 305-317). However, they do not describe survival trees. This can be found in (Ishwaran et al., 2008). The goal of tree-based models is to divide the space given by the covariate matrix $X,$ into subregions SR with observations whose covariates are similar. At the same time, the difference between the regions should be large. To find such subregions, tree models try to create a binary split on one or more covariates. This split should maximize the difference and create two sub samples called leaves or nodes. The splitting is then done recursively until a new split would need to create a leaf with less members than a specified minimum amount required. Then for each terminal node, an estimator of interest c_i is calculated. This estimator is different depending on if the model is a regression, classification or a survival model. The estimator c_i gives the prediction for the subregion SR_i. To denote if the covariates $\boldsymbol{x}$ of an observation belongs to a subregion, use the indicator function $I\{\boldsymbol{x} \in SR_i\}$. A decision tree model can then be described by

$$f(\boldsymbol{x}) = \sum_{i=1}^{k} c_i I\{\boldsymbol{x} \in SR_i\} \tag{34}$$

To model the survival and cumulative hazard rates for region SR_i, the estimated c_i can be assigned the associated Kaplan-Meier and Nelson-Aalen estimator (equation (6) and (9) respectively).

It remains to decide how to define the split criterion. One statistic

commonly used when comparing two survival curves is the log-rank statistic (Aalen et al., 2008, 106-109). The statistic's original application is in hypothesis testing of different effects of covariates between two or more populations. This statistic can be re-purposed to create a split for which covariate of $\boldsymbol{x}$ and threshold τ that maximizes the log-rank statistic. The covariates will split in a left and right node $L = \{\boldsymbol{x}_i \leq \tau\}$ and $R = \{\boldsymbol{x}_i > \tau\}$. Before defining the log-rank statistic some notation is needed. Let m denote the number of unique observed event times. Assign $Y_L(T_j), Y_R(T_j)$ and $Y_*(T_j)$ to equal the number at risk in the left split, right split and in total at time point T_j. Let $d_{j,L}, d_{j,R}$ and $d_{j,*}$ equal the number of events in the left split, right split, and in total at the same time point T_j. Then the log-rank statistic is defined by

$$L(X,\tau) = \frac{\sum_{j=1}^{m}\left(d_{j,L} - \frac{Y_L(T_j)}{Y_*(T_j)}d_{j,*}\right)}{\sqrt{\sum_{j=1}^{m}\frac{Y_L(T_j)Y_R(T_j)}{Y_*(T_j)^2}\left(\frac{Y_*(T_j)-d_{j,*}}{Y_*(T_j)-1}\right)d_{j,*}}} \tag{36}$$

The main improvement Random Forest makes is to utilize bootstrap techniques to reduce the variance. Assume that a number N_{tree} of identically distributed correlated decision trees can be created. Then the prediction error can be reduced by taking the average of the trees. This is because the mean would be the same as for a single tree, but the variance would equal

$$\rho_{tree}\sigma^2_{tree} + \frac{1-\rho_{tree}}{N_{tree}}\sigma^2_{tree} \tag{37}$$

Where σ^2_{tree} is the variance of a single decision tree. ρ_{tree} is the correlation between trees. The variance will decrease if many weakly correlated models are created. To build such models, is the motivation to use the bootstrap technique. It works by taking a random subset of size m of the p covariates. With this sample the decision tree is created. This methodology is repeated N_{tree} times. To create the final model, take the average of all bootstrap

models. Since the subsets may overlap, the models will be correlated. Choosing a high number of bootstrap covariates m in relation to the original number of covariates p would create highly correlated models. This is because the bootstrap samples are more likely to overlap. On the other hand, fewer covariates would reduce the correlation. To build a Random Survival Forest model, let $\hat{S}_i(t; \boldsymbol{x})$ and $\hat{H}_i(t; \boldsymbol{x})$ denote the estimated survival function and cumulative hazard rate for a decision tree created using bootstrap sample i. Then let

$$\bar{S}(t; x) = \frac{1}{N_{tree}} \sum_{i=1}^{N_{tree}} \hat{S}_i(t; \boldsymbol{x}) \qquad \bar{H}(t; x) = \frac{1}{N_{tree}} \sum_{i=1}^{N_{tree}} \hat{H}_i(t; \boldsymbol{x}) \tag{38}$$

be the average survival and cumulative hazard of the trees. These objects need to be created separately since the standard relation (equation (8)) between them does not hold for model averaging. This can be proven using Jensen's Inequality (Ishwaran et al., 2008). Let E denote the expectation, then the proof is as follows.

$$-\log(\bar{S}(t; \boldsymbol{x})) = -\log(E[S(t; \boldsymbol{x})]) \leq$$

$$\leq E[-\log(S(t; \boldsymbol{x}))] = E[H(t; \boldsymbol{x})] = \bar{H}(t; \boldsymbol{x}) \tag{39}$$

2.5 Validation methods

Comparing the three survival models is not straightforward since they are using different optimization procedures. Cox and Weibull regression are two types of linear models that maximize a likelihood function. Validation methods using the Chi-squared statistic and Akaike's information criterion (AIC) are used for them. In contrast, Random Survival Forest does not utilize a likelihood function and therefore the previously mentioned validation methods cannot be used. To compare the models, shared performance metrics have been developed. In this theses, the following four validation methods and performance metrics are explained and applied: the Concordance index (Harrell, Jr et al., 1982), Brier's score (Gerds &

Schumacher, 2006) and Confusion matrices (Agresti, 2013,223-225) together with the Receiver Operating Characteristic (ROC) curve at a specified timepoint.

2.5.1 Concordance index

The concordance index measures how well the survival models can rank who is more likely to develop the disease. It is also called Harrell's C-index (Harrell, Jr et al., 1982). The index will range between 0 and 1 since it is an adjusted ratio between concordant and permissible. 1 is an identical match, and 0.5 equivalent to random guessing. If the C-index is close to 0, it indicates that the model wrongly ranks the most likely observations as least at risk.

More specifically, the concordance is defined by comparing pairwise ordering of observed survival times against the ordering of the estimated hazards. For observations i and j, we can order their survival times T_i and T_j except in the following cases:

1. if both T_i and T_j are censored.
2. if T_j is censored and $T_j < T_i$

Let the set of pairs which can be ordered be denoted D. Call the number of pairs $|D|$ permissible. Using this set, scores for the following types of situations can be assigned. For all combinations D and their estimated hazards $\hat{H}(T_i)$, $\hat{H}(T_j)$, the following scores are given.

If $T_i < T_j$ or $T_j < T_i$ and the associated $\hat{H}(T_i)$ and $\hat{H}(T_j)$ keeps the ordering, assign 1 point. In the case the hazards are equal, $\hat{H}(T_i) = \hat{H}(T_j)$, assign 0.5 points.

If the event times are equal, $T_i = T_j$ and the hazards are also equal, $\hat{H}(T_i) = \hat{H}(T_j)$, assign 1 point. In the case when the hazards are not equal, $\hat{H}(T_i) \neq \hat{H}(T_j)$ assign 0.5 points.

All the scored combinations are called concordant. The ones not fulfilling any of the conditions are called discordant. Harrell's C-index is calculated through the sum of the scores divided by the scores plus the number of discordant.

$$C\text{-}index = \frac{Total\ Score\ of\ Concordant}{Total\ Score\ of\ Concordant + Number\ of\ Discordant} \quad (40)$$

It is of interest to estimate the standard error of the index. For the Random Forest model, this is the standard deviation of the concordance index for the trees. For the Cox and Weibull models, one could use the Jackknife technique (Therneau, 2026). The main idea of the Jackknife technique (Cameron & Trivedi, 2005, 375-376) is to leave out one observation i for the estimation of the concordance index $C\text{-}index_{-i}$. This is done once for every observation in the sample. The Jackknife also estimates a mean of these estimates of the concordance index, which can be considered the Jackknife estimate of the index,

$$C\text{-}index_{Jack} = \frac{1}{n}\sum_{i=1}^{n} C\text{-}index_{-i} \quad (41)$$

From these estimates one can calculate the Jackknife estimate of the standard error by

$$se_{jack} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(C\text{-}index_{-i} - C\text{-}index_{Jack})^2} \quad (42)$$

2.5.2 Briers score

Brier score provides a good evaluation measure alternative that focuses on closeness of fit rather than ranking of individuals. It is a measurement of the accuracy of the probability estimates. It will be in the interval between 0 and 1. Lower values indicate a more accurate prediction.

The Brier score is defined as a mean squared error, MSE. Let once again E denote the expectation. Then, since the purpose of survival analysis is to estimate the survival function, the MSE equals

$$MSE = E[(S(t;x) - \hat{S}(t;\boldsymbol{x}))^2] \quad (43)$$

This loss function cannot be directly used since the true survival function $S(t;\boldsymbol{x})$ is not known. The function can be approximated by the indicator variable $Y_i(t)$, from section 2.1. $Y_i(t)$ represents if an event has yet to occur for observation i. This leads to the loss function

$$B(t) = \frac{1}{n}\sum_{i=1}^{n}(Y_i(t) - \hat{S}(t;\boldsymbol{x}))^2. \quad (44)$$

This function still has issues, since the survival status for censored individuals cannot be decided. To control for this issue, the principle of inverse probability of censoring weights, (IPCW) could be used (Gerds & Schumacher, 2006). The idea is to discard the censored observations $\delta_i = 0$ and weigh the remaining at time points of interest t by the inverse probability of them being censored $G_i(t)$. The censoring distribution can be estimated by the Kaplan-Meier estimator. Instead of modelling for the original event, the estimator models the time to censoring. The weight has the form of

$$w_i(t) = \frac{(1-Y_i(t))\delta_i}{G_i(T_i^-)} + \frac{Y_i(t)}{G_i(t)} \quad (45)$$

The terms in the weight are exclusionary. The first term gives the weight for observations which experienced the event of interest before t. In this case, the weight is inverse probability of being censored at time, T_i^-, which is just before they experienced the event at time T_i. The second term is if by time t, the observation has still not experienced the event. In this case, the weight is the inverse probability of the observation to have been censored at time t.

Weighing the Brier score by IPCW, yields the weighted score of

$$B(t;\boldsymbol{x}) \;=\; \frac{1}{n}\sum_{i=1}^{n} w_i(t)\left(Y_i(t) \;-\; \hat{S}(t;\boldsymbol{x}_i)\right)^2 \;= \tag{46}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{S}(t;\boldsymbol{x}_i)^2(1-Y_i(t))\delta_i}{G_i(T_i^-)} + \frac{(1-S(t;\boldsymbol{x}_i))^2 Y_i(t)}{\hat{G}(t)}\right) \tag{47}$$

Due to the Brier score depending on time t, it will have different values at different time points. One way to visualize the score is to create a plot of the score against time. Such a plot is displayed in Figure 7 in chapter 4. If one instead want to compare the different models by a statistic, one could calculate the integrated Brier score

$$IBS(t) \;=\; \int_0^t B(u)du \tag{48}$$

To account for that larger timepoints t will yield a higher Integrated Briers score, one can divide it by t. This quantity is referred to as the continuous rank probability score, CRPS, in the Random Forest implementation used.

2.5.3 Confusion matrix for survival at specified time point

It is often of interest to study the survival probability at a specified time point. An example is the end point of a study. At a specified time point, the survival curve can be used for a binary classification model. In that case common evaluation methods for classification can be used. The following methods have been discussed by (Agresti, 2013). A confusion matrix is a 2×2 matrix between the trained classification and the observed. It results in 4 cells, True negative, where both the predicted and observed outcome is negative. False negative, here the predicted is negative but observed is positive. False positive where the predicted is positive and observed negative. True positives where both the prediction and the observed are positive. The trained classification is defined by the survival probability by the specified time point exceeding a threshold π_0. The threshold is often set to $\pi_0 = 0.5$, but can be any value between 0 and 1.

Two values commonly used in medical studies are sensitivity and specificity (Agresti, 2013, 223). Sensitivity is the probability for a model to predict a positive result given it was actually positive. Likewise, specificity is the probability for a model to predict a negative result given that it actually was negative. Sensitivity and specificity are usually presented in percentage form.

2.5.4 Receiver Operating Characteristic curve for survival at specified time point

Another validation method for a specified time point (Agresti, 2013) is the Receiver Operating Characteristic (ROC) curve. It is a curve defined by the pairs of sensitivity and the rate of false positives, $(1 -$ specificity), for all thresholds π_0 of interest. Due to the ranges of the pair, the curve is defined in a two-dimensional space with axes between 0 and 1. A well performing classification model has a high sensitivity and high specificity. This means an ideal ROC curve is one close to the left and upper edges of the diagram. A curve close to being a straight line between lower left corner and the upper right corresponds to a model guessing at random. See Figure 8 in chapter 4 as an example. A numeric value indicating how well the ROC curve performs can be retrieved by the area under the curve (AUC.) A well performing model has thus an AUC over 0.5.

2.5.5 Problems with Confusion matrix and Receiver Operating Characteristic curve in survival analysis.

The problem with the confusion matrix and ROC curves when applied in a survival analysis setting, is that they depend on a specified timepoint of the survival function. Early timepoints are likely to have a high prediction accuracy. This is because the events would not have had enough time to occur. Late timepoints can also be problematic because some types of events are almost guaranteed to have happened by then. To use these evaluation methods, the time point needs to be well motivated.

3 Materials and method

3.1 Description of data

To compare Cox regression, Weibull regression and Random Survival Forest, the models were evaluated on a medical dataset. concerning patients at risk of developing rheumatoid arthritis (RA). The data comes from a prospective research programme at Karolinska Institute called Risk-RA. It recruits patients at risk of developing RA from the following three rheumatology clinics in the Stockholm region: Karolinska University Hospital, Academic Specialist Center and Danderyd Hospital. To be included in the programme, the patients must fulfill the following five criteria (Cîrciumaru et al., 2024). 1) Have musculoskeletal complaints which after assessment by rheumatologists and other specialties are suspected to be caused by a rheumatic disease. 2) They are screened for no current or prior rheumatic diseases. 3) Have a positive antibody test against CCP2 (second-generation anti-cyclic citrullinated peptide). 4) lack signs of arthritis when assessed on ultrasound. Included patients are followed for three years, to see if they developed the disease.

Patients newly included in the Risk-Ra programme were asked to state their symptoms through a questionnaire on a website called “Rheumatic?” (Knevel et al. 2022). The questions regarded symptoms of rheumatic diseases in general, and as such, some questions had more relevance to RA than others. The purpose of the questionnaire was to advise the person answering whether they should seek medical care or not.

The number of participating patients was 79 and the questionnaire had 76 questions. The questions were of three types, multiple choice, ordinal and a continuous scale and encoded as 475 variables for each patient. Each variable was binary and indicated whether a person had a specific symptom or other health and lifestyle factors. For ordinal and continuous scale questions, the variables also represented the extent of symptoms or other health and lifestyle factors. The topics of the questions could be

subdivided into five categories. Four of them concerned symptoms regarding pain, swelling, stiffness and skin conditions. The fifth contained questions regarding lifestyle and family history of rheumatic and related diseases.

Several questions were hierarchical. If one answered “no” to a question, no follow up questions were asked concerning the specific topic. If the participants were unsure what they should answer, there was an option to answer ”I don’t know”.

In a previous pilot study (Knevel et al., 2022) scores were given for each variable of the patients. The authors assigned some variables with higher scores than others. The variables that represented the most important symptoms for developing RA had scores of 30. Vague and only possibly related symptoms had a score of one or two. For variables representing unrelated symptoms, a score of zero was assigned.

The scores for the continuous scale questions were created by dividing the scale in five levels. For each higher level, a score of two was added, with the maximum being ten. These questions are from now on treated as ordinal in this thesis. A total score was calculated for each patient. When answering the Rheumatic? questionnaire, the maximum possible score was 760. The authors of the pilot study defined two threshold levels to decide if advice to seek care should be given. Between 250 and 400 points, the patient was advised to seek a general physician. If the score was over 400, a rheumatologist should be consulted. The data set in the pilot study contained 50 individuals, who were either diagnosed with the disease during the three-year study period or not. Since then, more data has been collected. At the time of writing, 79 individuals have been observed. Of these patients, 65 could at the end of the study period be determined if they were diagnosed with RA or not. There were 14 individuals which were censored because they entered the study late and had not been observed for three years. In this study the scores were reused. The symptoms which had no score in the previous study were assigned 0.5 in this study. The reason for the change is to avoid excluding

information.

The original authors created confusion matrices for each threshold. The sensitivity and specificity of these matrices was also reported. To validate how well the scoring worked for arbitrary thresholds, the authors created a ROC curve. The properties of how the data was collected and its purpose, aligns with the regression methods discussed in section 2.

It is of interest to compare the performance of the survival model against the previous scoring method. Therefore, the total score for each patient was re-calculated and compared with the updated diagnosis status of the patients.

Special treatment is needed to handle the cases where a patient gave the response “I don’t know”. If the answer belongs to a question with scored alternatives in the previous study, the patient could have given important information, if they knew what they should answer. In this thesis “I don’t know” answers belonging to such questions are imputed. Other “I don’t know” answers are treated as not having occurred. The idea behind imputation is to create predictive models of only the data where no one answered “I don’t know”. The target variables for each model are a question where observations gave the problematic answer. Then the models are used to predict which answer it most likely should have been. The “I don’t know” answers belonging to questions with scored alternatives in the previous study are concentrated in three questions see, Figure 3-5. Only seven individuals gave the answer for these questions. For them imputation is performed by K-nearest neighbours classification, (KNN) (Hastie et al., 2009, 14) for the questions they answered “I don’t know”. The classifications are done by comparing these patient with the k most similar patients which knew their answers. The similarity measure used between patients is the euclidean distance of the scored answers excluding those belonging questions where some answered, “I don’t know”. The euclidean distance can be problematic in this context though, since it assumes continuous coordinates and will

not give a distance corresponding to a value achievable by any variable. To validate how well the KNN performed, the mean squared error between the prediction and answers of people who knew what they should answer is calculated. For multiple choice questions, the sum of the mean squared errors for each alternative are used. To choose the number of *k* neighbours to apply, ten-fold cross validation for k between 2 and 30 is performed. From the cross validation, the number of neighbours minimizing the error is to be used when performing imputation afterwards.

3.2 Principal Component Analysis

Since the number of symptoms far exceeds the number of participants, fitting the data directly to the covariates would lead to an unreliable result. This is due to the curse of dimensionality. As an attempt to solve this issue, principal component analysis, PCA, was used to find a lower dimensional representation. A short summary of principal component analysis (PCA) is presented here based on how the method is described by (Lee & Verleysen, 2007).

PCA assumes that the observed p dimensional covariate matrix X originally came from a latent lower q dimensional representation Y. The columns of Y are assumed to be independent. Y is also assumed to have been transformed by an orthogonal matrix W to a higher dimension giving X

$$X = WY \tag{49}$$

Hence it is implied that $\hat{Y} = W^{\top}X$. Another important assumption of PCA is that both X and Y have column means of zero. If the mean of the columns of the covariate matrix is not zero, then they are centred to zero by subtracting the mean of each column. The covariance matrices of X and Y can be denoted as $C_X = XX^{\top}$ and $C_Y = YY^{\top}$. The latter is also a diagonal matrix. PCA estimates the latent variables by using the eigenvalue decomposition of the covariance matrix of

$$C_X = V\Lambda V^{\top}. \tag{50}$$

V is the matrix of eigenvectors of C_X and Λ is a diagonal matrix of the eigenvalues of C_X in descending order. PCA estimates W with the p-first column vectors of V. Let $I_{p\times q}$ be a truncated $p \times p$ identity matrix consisting of the q first columns. Then $W = VI_{p\times q}$ and $\hat{Y}$ is estimated as

$$\hat{Y} = I^{\top}_{p\times q} V^{\top} X \tag{51}$$

The covariance of C_Y can also be formulated by

$$C_Y = I^{\top}_{p\times q} \Lambda I_{p\times q} \tag{52}$$

An important consequence of this result is that the eigenvalues of C_X equal the variances of the latent variables. In theory, the number of nonzero eigenvalues should equal the number of latent variables. Due to the presence of noise when collecting X, the number values are the same as observations. A common assumption is that the noise has less variance than the latent variables. Thus, the problem becomes finding the q latent variables which explains most of the variance. To do that one often plots the eigenvalues and tries to keep the ones which together explains most of the variance. For an example, see Figure 6.

All the conditions which PCA assumes are not met by the weighted questionnaire data. For instance, the euclidean distance is problematic in this context since it assumes continuous coordinates and will not give a distance corresponding to a value achievable by the variables. A consequence of this is that PCA may not correctly retain the variance it seeks when reducing the dimensionality. This will reduce the quality of the predictors for the survival models and could negatively affect their performance. A different dimensional reduction technique satisfying all the properties of the data could give better results.

3.3 Implementation, Programming languages and Packages

To implement and analyse the survival models, the programming language R (R Foundation for Statistical Computing, 2026) and programming environment Rstudio was used. The majority of the models did not exist in the base version and had to be installed. The following packages were used.

The *Survival* package (Therneau et al., 2026) for Cox and Weibull

regressions

RandomForestSRC package (Ishwaran & Kogalur, 2026) for random survival forest.

The *SurvMetrics* package (Zhou et al., 2025) was used to calculate the Briers score for the Weibull model which didn't have a native implementation in Survival.

The *ROCR* package (Sing et al., 2026) to calculate the ROC curve and AUC statistic.

To implement PCA, the function prcomp in the base package *Stats* in R was used.

To perform the Knn imputation, the packages *modelr* and *Directional* (Tsagris et al., 2025) was used

For data wrangling and plotting tools in the base R package, the *Tidyverse* packages were used.

When implementing the methods a couple of parameters needs to be specified. For the Random Survival Forest model, it was decided to let the minimum node size to consist of 5 observations. The number of trees created were 1000. Due to the limited sample size, it was decided to only evaluate the three models on a training set. This is problematic because evaluating the survival models on the training set could cause them to overfit. Hence the result should be interpreted with caution. The ROC curve and confusion matrix used the survival probability after three years. This timepoint was chosen, since it was the end of the study period.

4 Results

4.1 Result of pre-processing

To handle the Scored questions where patients answered “I don’t know” imputation was performed. The imputation method used was k-nearest neighbours. To choose how many neighbours it should use, cross validation for between 2 and 30 neighbours were done. The results are presented in Figure 3-5. In these Figures the black points represent the mean squared error between prediction and observations. The bars are their standard deviation. For all questions, the mean squared error was the lowest for two neighbours. The large error bars shows that the error was not significantly different compared to other choices of number of neighbours. From this result, it was decided to use two neighbours when imputing the answers.

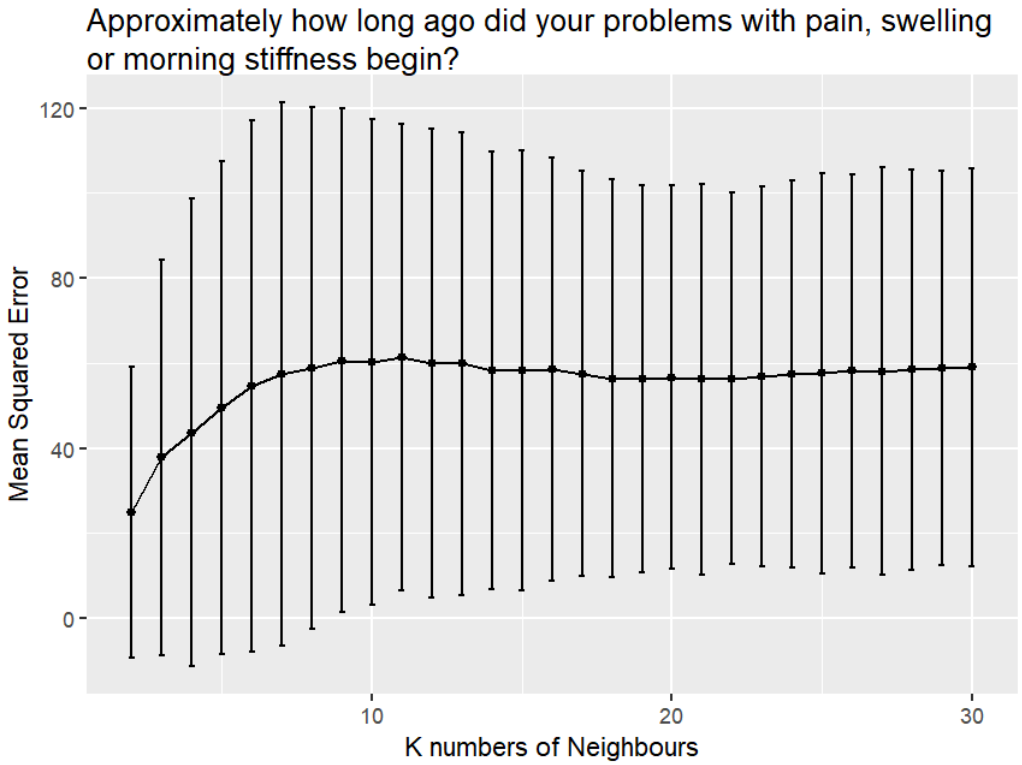

Figure 3: Cross validation plot to decide how many neighbours KNN should use for imputation of the question above. The number of K neighbours tested were between 2 and 30. The black dots are the mean squared error of cross-validation for each K. The bars represent one standard error. Two neighbours minimize the error.

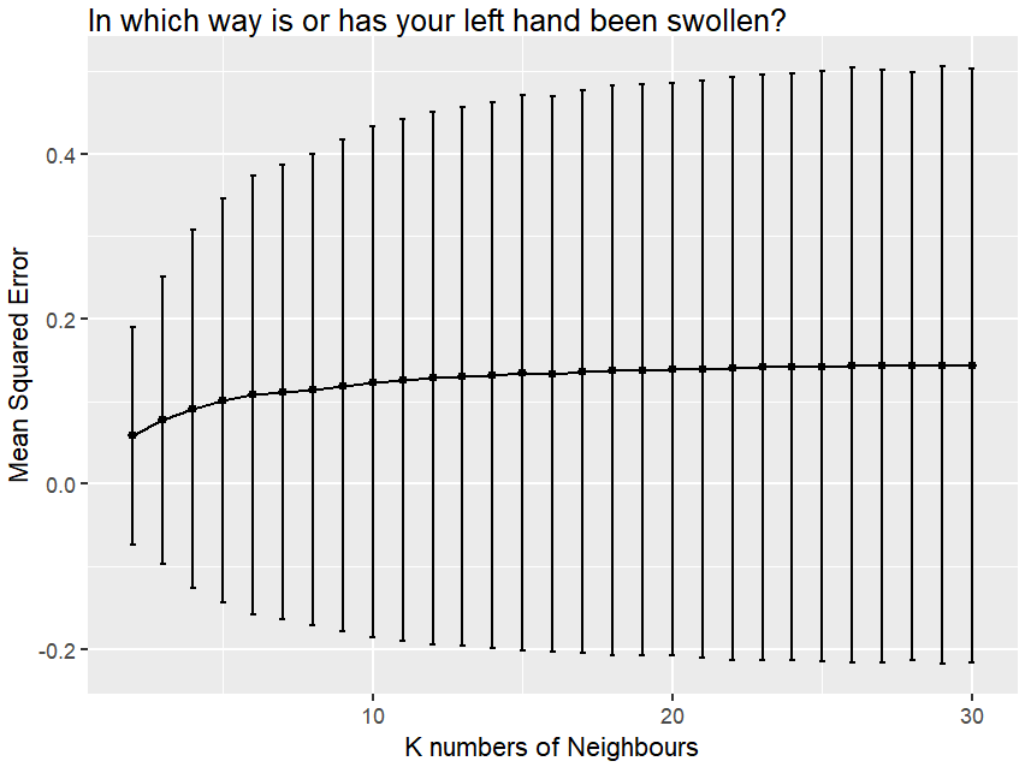

Figure 4: Cross validation plot to decide how many neighbours KNN should use for imputation of the question above. The number of K neighbours tested were between 2 and 30. The black dots are the mean squared error of cross-validation for each K. The bars represent one standard error. Two neighbours minimize the error.

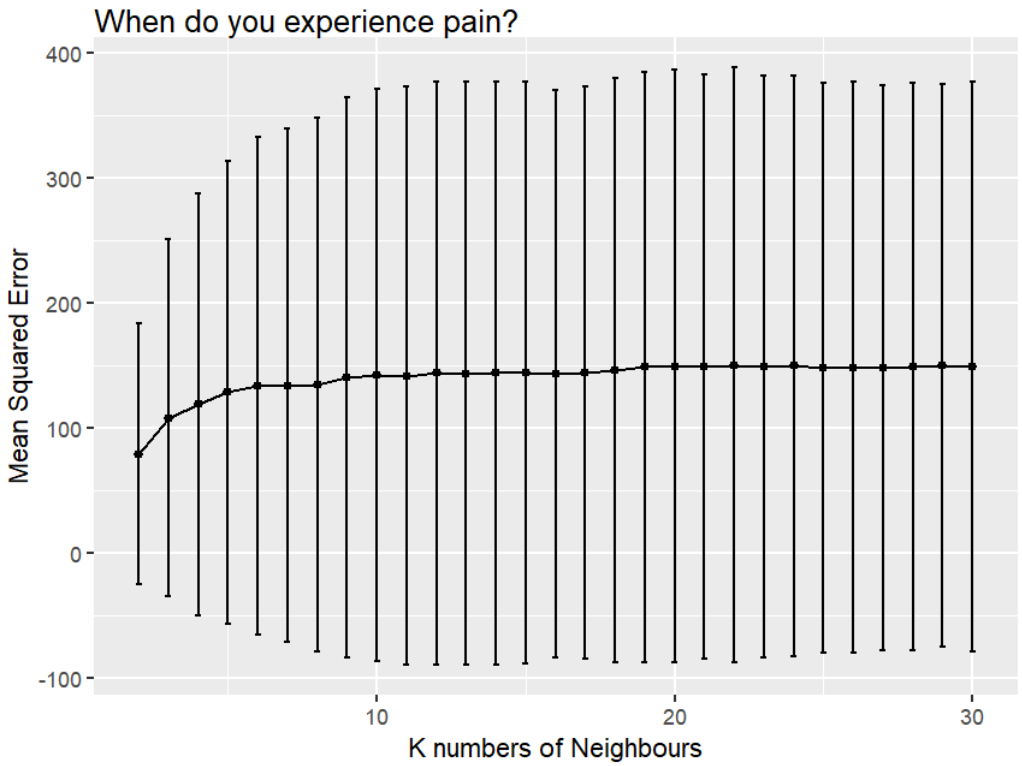

Figure 5: Cross validation plot to decide how many neighbours KNN should use for imputation of the question above. The number of K neighbours tested were between 2 and 30. The black dots are the mean squared error of cross-validation for each K. The bars represent one standard error. Two neighbours minimize the error.

To reduce the dimensions of the RA data, PCA was applied. To decide how many principal components to use, a bar plot of the eigenvalues and variance of components are presented in Figure 6. The first eigenvalue explains much of the variance. The rate of the decline of the eigenvalues seems to go down after the seventh eigenvalue. To keep most of the variance of the original data, it was decided to use the first seven principal components. With these components the predictive models were created.

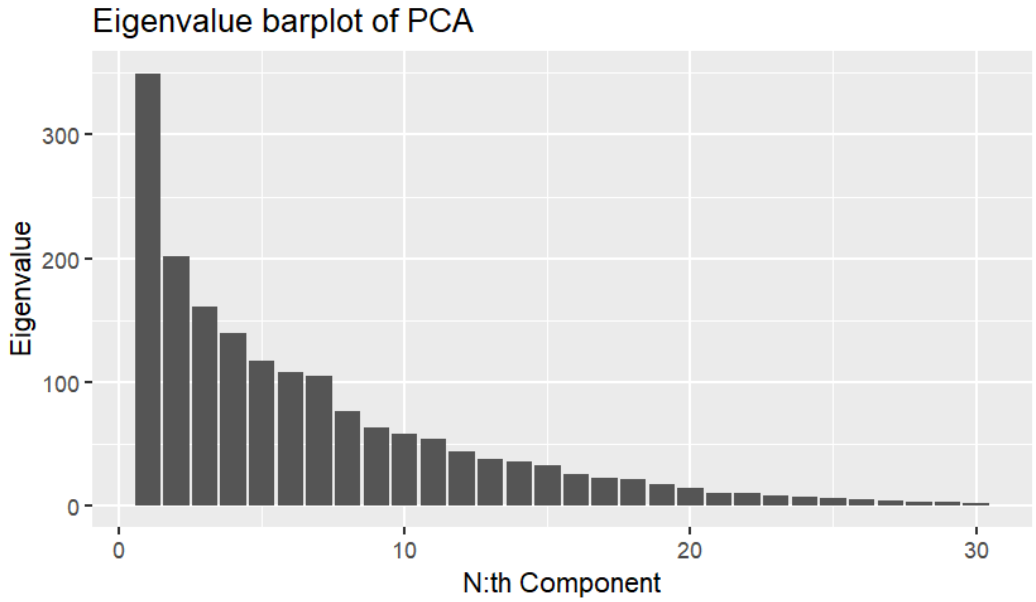

Figure 6: A barplot of the eigenvalues of PCA in descending order. Due to the gap between the seventh and eight eigenvalues the first seven components were chosen.

4.2 Statistics of model performance

In the tables below, key statistics for the models performances are presented. Standard statistics for the Cox and Weibull models are presented in Table 1. Their AIC, maximum and null model of their log likelihood, chi-squared statistic, and p-value are presented. The scale of the Weibull model was estimated as $\hat{\lambda} = 0.97$. The AIC was lower for the Cox model. The maximum and null log-likelihood were doubled in the Weibull model. The chi-square statistic and the p-value were similar. The p-value was significant for both, meaning the covariates had a significant effect of predicting RA.

Table 1: Key statistics for Cox and Weibull models

	AIC	Max log likelihood	Null-log likelihood	Chi Squared	P-value
Cox regression	251	-127.9	-118.8	18.2	0.01
Weibull regression	541	-261.8	-271.2	18.8	<0.01

The key statistics for the Random Survival Forest model are described in Table 2. They are, Terminal node size, Average number of terminal nodes of the trees, integrated Briers score (CRPS) and Performance error. The Performance error equals one minus the concordance index.

Table 2: Key statistics for Random Forest model

Terminal node size	Average number of terminal nodes	Integrated Briers score (CRPS)	Performance error
5	10.1	180	0.42

4.3 Validation result

4.3.1 Concordance index

In Table 3, the concordance index and associated standard error for each model are shown. The index indicates how well the models rank, who is more likely to develop the disease. A higher index indicates better ranking and an index of 0.5 is close to random guessing. The Cox and Weibull models gave similar results. Their Concordance index was 0.734 and 0.735 with both having a standard error 0.047. Random Forest model gave a lower index of 0.579 and a standard error of 0.007. This result indicates that the Cox and Weibull models could rank who is more likely to develop the disease more accurately than Random Survival Forest.

Table 3: Concordance index for the three survival models.

	Cox regression	Weibull	Random Forest
C-index	0.734	0.735	0.579
Standard error	0.047	0.047	0.007

4.3.2 Briers Score

In Figure 7, curves of the Brier scores have been plotted against time. A lower brier score means that the estimated survival probabilities are closer to the observed disease status. In general, the score increases with time. Both the Random Forest and the Cox model start at a score of 0. The Weibull model differs and instead starts from around 0.14. The Random Forest model increases the most and diverges from the Cox model. For the later time points it has a higher score than the Weibull model. The Cox model increases but not to the same degree. This model has the lowest Briers score for almost all time points. The Weibull model has the smallest increase in Brier score but remains higher than the score obtained by the Cox model. This result indicates that the

survival probabilities of the Cox model were more accurate than the other two. The survival probabilities were more accurate for Random Survival Forest than Weibull for early time points. This result was reversed for later time points.

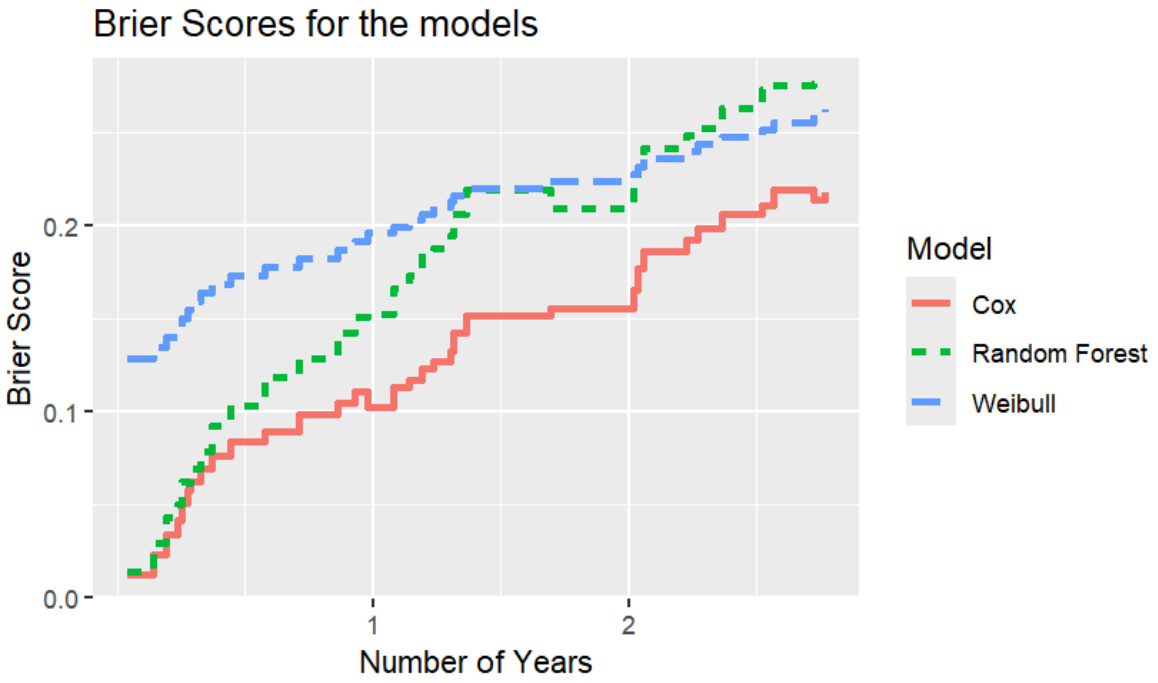

Figure 7: Briers scores for each model over most of the study period. The score is calculated from when the first patient is diagnosed to when the last patient is. Brier's score is a weighted mean square error between the survival probability and the observed outcome at the time. A lower score indicates better model performance.

4.3.3 Confusion matrix and ROC curve

In Table 4, a confusion matrix is created. It describes how well the models predict diagnosis at the end of the study period. The table compares the predictions with the observed diagnosis. As mentioned earlier, 65 individuals could be evaluated in the study, of which 32 developed RA. The Cox and Weibull models gave identical predictions. The Random Forest model predicted that 56 of 65 persons would not develop the disease and it had no false positive prediction.

Table 4: Confusion matrices of the classification by the survival models and their specificity and sensitivity.

Model:	Cox		Weibull		Random Forest	
	With no RA	RA	With no RA	RA	With no RA	RA
Predicted with no RA	23	14	23	14	32	25
Predicted with RA	9	19	9	19	0	8
Total Number of Patients	32	33	32	33	32	33

In Table 5, the Sensitivity and Specificity of the models are presented. The models often classified those who would develop the disease as not likely. This can be seen by the sensitivity of the models being between 24% and 58%. In contrast, when predicting the status for individuals who did not develop RA, the models were frequently correct. This can be seen by the specificity being higher than 72%. The sensitivity of the Random Forest model was lower than the Cox and Weibull models. However, the specificity was higher for the former. The result indicates that the Random Forest model was more conservative than the other models at predicting if a person was likely to develop RA.

Table 5: Sensitivity and specificity of the classification by the survival models.

Model	Cox	Weibull	Random Forest
Sensitivity	58%	58%	24%
Specificity	72%	72%	100%

In Figure 8, the receiver operating characteristic (ROC) curves for each model and the original scoring method is presented. True and false positive prediction rates for the models are calculated for different thresholds of the Survival probability as mentioned in section 2.5.4. The area under the curve, AUC, statistic for Cox was 0.73, for Weibull 0.68 and Random Survival Forest 0.90. AUC statistics closer to 1.00 indicate a model which can determine which observation is more likely to develop the disease before 3 years. This result indicates that Random Survival Forest was noticeably better at the task than the other models.

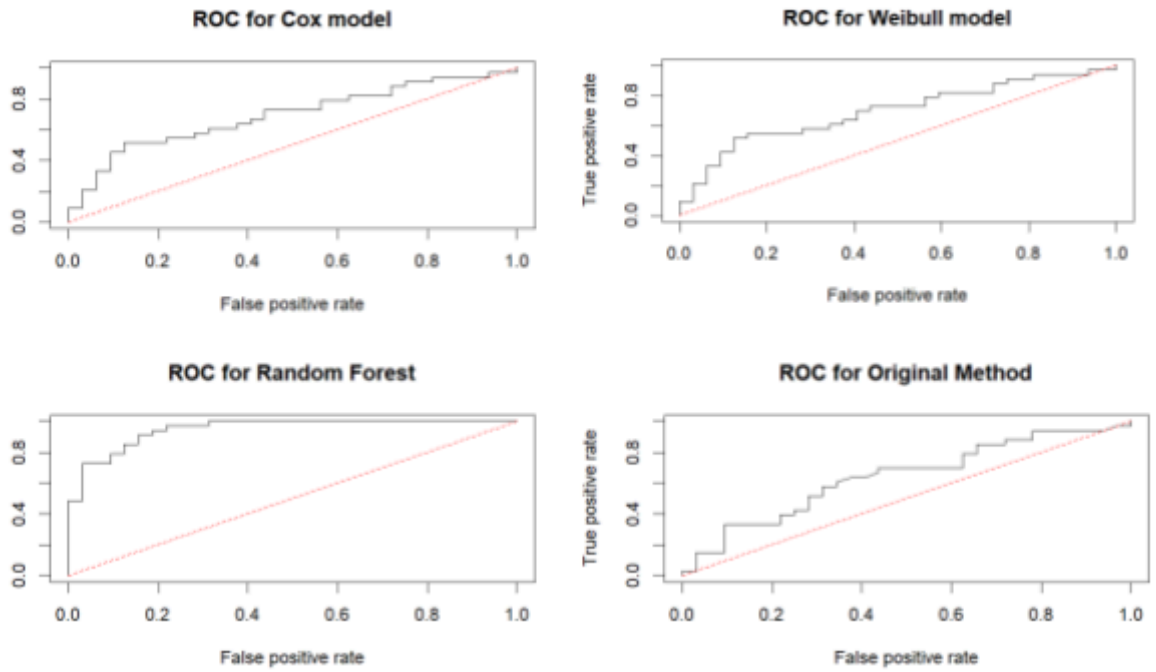

Figure 8: The Receiver Operating Characteristic curve for the Cox model, Weibull model, Random Forest model and Original method. The area under the curve (AUC) was 0.73 for Cox, .0.68, for Weibull, 0.90 for Random Survival Forest and 0.64 for the original scoring method.

4.4 Result of reproducing original scoring model

To compare the performances of the survival models against the original methodology, the latter was reproduced. The reproduction of the scoring method used the newer version of the RA data set mentioned in this thesis and which the survival models utilized. The total score for each individual was calculated. A comparison between the number of individuals whose score reached the 250 and 400 score thresholds against the diagnosis status after 3 years are done in the confusion matrices in Table 6.

Table 6: Predictive performance of the reproduced scoring method against their two thresholds including sensitivity and specificity.

Model	Threshold score 250		Threshold score 400	
n = 65	No RA	RA	No RA	RA
Individuals with scores not above threshold	29	24	32	33
Individuals with scores above threshold	3	9	0	0
Sensitivity	27%		0 %	
Specificity	91%		100%	

The predictions of people to not develop rheumatoid arthritis was generally correct. The specificity for 250 and 400 points thresholds were 91% and 100% respectively. However, for both thresholds, the original method had low sensitivities of 27% and 0%. This means that the scoring method had many false negatives. In the lower left quadrant of Figure 8 is the ROC curve for the original scoring method presented. True and false positive prediction rates for the method are calculated for different thresholds of the total score. The AUC statistic of the ROC was 0.64. The statistic is above 0.5 which corresponds to better than random guessing. The AUC is also with a noticeable margin below 1.00. This means that the original model can to some degree order which individual is more likely to develop rheumatoid arthritis.

5 Discussion and Conclusion

5.1 Discussion

The aim of this study was to compare three survival models, Cox regression, Weibull regression and Random Survival Forest, on their ability to predict when and if a person will develop rheumatoid arthritis. A secondary goal was to examine their applicability to the previous RiskRa study. The performance of the models varied between the validation methods. The importance of the validation methods depends on which quality of the models are of interest. For investigating how well the models predict when a person will develop the disease, the Concordance index and Brier score is more important. For the application to the RA study, the previously used confusion matrix and ROC curve is more informative.

Comparison between the three survival models showed that the Cox and Weibull models had similar concordance indices. In contrast the Random Survival Forest had a noticeably lower index. The result being close to 0.5 indicates the model had it hard to order the time until event for when a person would develop the disease. This indicated that the previous two mentioned models could more easily order which person is more likely to develop RA. It is possible that the reason for the similar result between the Cox and Weibull model is because their hazard rate was almost the same. The Weibull model had an almost constant scale indicating their base hazard was almost constant. This could cause the Weibull hazard to be roughly proportional to the Cox hazard. The similarities can be seen in equation (11) and (23). The scaling difference would not affect the ordering of the observations significantly.

The Briers score showed that the Cox model had the lowest score. For all except the last year, Random Survival Forest had a lower Brier score than the Weibull model. The result indicates that the Cox model had the most accurate probability estimates. The Weibull model distinguished itself by starting from a Briers score

above zero. From how the Brier score was defined in equation (47), it should be impossible for the score to be above zero at the start. There must have been an error when applying the Weibull model to the implementation of Brier score from the Survmetrics package. The Brier score for the Weibull model is hence less reliable than for Cox and Random Survival Forest. In a follow up work, the Brier score for the Weibull model should be recalculated. Since the Cox and Random Forest used Nelson-Aalen estimator which remembers the survival times it was trained on, it is possible that their training errors are affected by overfitting.

The confusion matrices showed that the Random Forest model had a higher specificity than the Cox and Weibull models. However, the Random Forest model was less sensitive. The AUC statistic indicated that the Random Forest model could more accurately predict which person is more likely to develop the disease. When comparing with the result of the scoring method, the sensitivity of the Cox and Weibull models were better. However, their specificity was also lower. This caused the ROC curve and AUC to be similar between these two survival models and the scoring method. The Random Forest model had roughly the same sensitivity as the scoring model with 250 points threshold. It also had the same specificity as the 400 points model. The ROC curve and AUC for the survival model were higher than the scoring model. The Random Forest model could be an improvement over the scoring method, but due to the model being validated on the training set, it is uncertain how much the improvement can be attributed to overfitting. It is also possible that the ability of Random Forest to detect nonlinearities improved the result. The result of the Confusion matrices and ROC curves depended on the timepoint of interest used. In this study the end of the study period was chosen. The performance of the survival models and the original scoring method could be different if a different time point had been used.

Since the assumption of PCA is violated, the representation by the dimensionality reduction possibly retains less information than

desired. This could indirectly have reduced the models performance. Alternative dimensionality reduction methods to PCA which are designed for binary variables such as logistic PCA (Landgraf & Lee, 2020) could possibly have improved the result. Also, With fewer and more carefully chosen covariates, the number of reduced dimensions, if needed at all, would be lower. The representation would then possibly be more informative. Consequently, it could possibly enhance the model performances.

Due to the methodology used to create the models, it is important to interpret the result carefully. Because of the limited sample size, the result is affected by the variance in the data. Creation of a test set to measure overfitting was not possible. If it had been possible, it could have yielded a different result. With this in mind, further studies with more individuals should be done.

5.2 Conclusion

If the goal is to predict when and if a patient will develop Rheumatoid Arthritis, Cox regression performed the best. The Weibull model performed similarly, but worse for early timepoints. Random Survival Forest has difficulty predicting when patients will be diagnosed. If the goal is to improve the methodology of the previous RiskRA study, Random Forest is possibly better at predicting who is more likely to develop the disease in three years.

6 References

Aalen, O. (1978). Nonparametric Inference For a Family Of Counting Processes. *The Annals of Statistics*, *6*(4), 701-726. https://www.jstor.org/stable/2958850

Aalen, O., Borgan, O., Borgan, Ø., Gjessing, S., & Gjessing, H. (2008). *Survival and Event History Analysis: A Process Point of View*. Springer.

Agresti, A. (2013). *Categorical Data Analysis*. Wiley.

Breiman, L. (2001). Random Forest. *Machine Learning*, *45*, 5-32. https://doi.org/10.1023/A:1010933404324

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press.

Cîrciumaru, A., Kisten, Y., Hansson, M., Mathsson-Alm, L., Joshua, V., Wähämaa, H., Haarhaus, M. L., Lindqvist, J., Padyukov, L., Catrina, S.-B., Fei, G., Vivar, N., Rezaei, H., af Klint, E., Antovic, A., Réthi, B., Catrina, A. I., & Hensvold, A. (2024, November). Identification of early risk factors for anti-citrullinated-protein-antibody positive rheumatoid arthritis—a prospective cohort study. *Rheumatology*, *63*(11), 3164–3171. https://doi.org/10.1093/rheumatology/keae146

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187-202. 10.1111/j.2517-6161.1972.tb00899.x

Gerds, T. A., & Schumacher, M. (2006). Consistent Estimation of the Expected Brier Score in General Survival Models with Right-Censored Event Times. *Biometrical Journal*, *48*(6), 1029-1040. 10.1002/bimj.200610301

Harrell Jr, F. E., Califf, R. M., & Pryor, D. B. (1982). Evaluating the Yield of Medical Tests. *JAMA*, *247*(18), 2543-2546. 10.1001/jama.1982.03320430047030

Hastie, T., Tibshirani, R., Friedman, J., & Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer New York.

Ishwaran, H., & Kogalur, U. B. (2026). *Fast Unified Random Forest for Survival, Regression and Classification (RF-SRC)* [R package version 3.6.2]. https://cran.r-project.org/package=randomForestSRC

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). RANDOM SURVIVAL FORESTS. *The Annals of Applied Statistics*, *2*(3), 841-860. 10.1214/08-AOAS169

Kalbfleisch, J. D., & Prentice, R. L. (2011). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons Canada, Limited.

Kaplan, E. L., & Meier, P. (1958). NONPARAMETRIC ESTIMATION FROM INCOMPLETE OBSERVATIONS. *Journal of the American Statistical Association*, *53*(282), 457-481. https://doi.org/10.2307/2281868

Knevel, R., Knitza, J., Hensvold, A., Cîrciumaru, A., Bruce, T., Evans, S., & Maarseveen, T. (2022). Rheumatic?—A Digital Diagnostic Decision Support Tool for Individuals Suspecting Rheumatic

Diseases: A Multicenter Pilot Validation Study. *frontiers*, *9*(Article 774945). 10.3389/fmed.2022.774945

Landgraf, A. J., & Lee, Y. (2020). Dimensionality reduction for binary data through the projection of natural parameters. *Journal of Multivariate Analysis*, *180*. 10.1016/j.jmva.2020.104668

Lee, J. A., & Verleysen, M. (2007). *Nonlinear Dimensionality Reduction* (J. A. Lee & M. Verleysen, Eds.). Springer New York. https://link.springer.com/book/10.1007/978-0-387-39351-3

Leiden University Medical Center & Karolinska Institutet. (n.d.). *Rheumatic?* Rheumatic? Retrieved 19 08, 2025, from https://rheumatic.elsa.science/en/

R Foundation for Statistical Computing. (2026). *R: A Language and Environment for Statistical Computing*. The R Project for Statistical Computing. https://www.R-project.org/

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). *ROCR:* [visualizing classifier performance in R." Bioinformatics, 21(20), 7881.]. https://ipa-tys.github.io/ROCR/

Smolen, J. S., Aletaha, D., Barton, A., Burmester, G. R., Emery, P., Firestein, G. S., Kavanaugh, A., McInnes, I. B., Solomon, D. H., Strand, V., & Yamamoto, K. (2018). Rheumatoid arthritis. *Nature Reviews Disease Primers*, *4*. 10.1038/nrdp.2018.1

Therneau, T. (2026). *_A Package for Survival Analysis in R_,* [R package version 3.8-6,]. CRAN. Retrieved February 16, 2026, from https://CRAN.R-project.org/package=survival

Tsagris, M., Athineou, G., Adam, C., Yu, Z., Sajib, A., Amson, E., Waldstein, M. J., & Papastamoulis, P. (2026). *_Directional:* [A

Collection of Functions for Directional Data Analysis_. R package version 7.4]. https://CRAN.R-project.org/package=Directional

Zhou, H., Cheng, X., Wang, S., Zou, Y., & Wang, H. (2025). _*SurvMetrics:* [Predictive Evaluation Metrics in Survival Analysis_. R package version 0.5.1]. https://CRAN.R-project.org/package=SurvMetrics