Longitudinal and survival data analysis for rheumatic diseases’ prognosis

Carolina de Seixas Serra Domingos Barata
carolina.barata@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

September 2020

Abstract

Rheumatic diseases are one of the most common chronic diseases worldwide. Among them, spondyloarthritis (SpA) is a group of highly debilitating diseases, with an early onset age, significantly impacting patients’ quality of life, health care systems, and society in general. Recent treatment options consist of using biologic therapies, and establishing the most beneficial one according to the patient’s characteristics is a challenge that needs solving. Simultaneously, the emerging availability of electronic medical records (EMR) urges the development of methods that can extract insightful information while handling all the challenges of dealing with complex, real-world data.

Our goal is to better understand SpA therapy responses and identify the predictors that affect them, enabling the prognosis of therapy success or failure. Moreover, we aim at achieving a usable dataset for future studies by pre-processing the data. A sensitivity analysis of the results to the different steps of this process will be performed.

Joint models for the survival analysis of the biologic therapy failure are used, considering the information of both baseline and time-varying variables extracted from the EMR of SpA patients from the database Reuma.pt. Possible predictors of biologic drug survival were identified. It was verified that the approach used for handling data influenced the selection of variables and, consequently, the results. Furthermore, joint models proved to be a highly valuable tool for this type of analysis.

Keywords: Data Mining, Survival Analysis, Joint Models, Spondyloarthritis, Drug Survival.

1. Introduction

Rheumatic diseases are chronic diseases that consume a large amount of health and social resources, being the leading cause of disability in developed countries. Among these diseases, spondyloarthritis (SpA) is a group of several related disorders that can be highly debilitating and significantly impact patients’ quality of life, health care systems and society [1].

For there is no cure, treatment focuses on the relief of symptoms and on delaying the evolution of the disease. Biologic therapies are the most recent approach for treating these disorders and their use is recommended when all other methods have failed. However, the therapy selection follows no specific criteria and trying to establish which patients benefit the most from each drug is still a challenge that needs solving [2].

A better understanding of therapy responses for these patients would allow the prognosis of therapy success or failure, being highly valuable in terms of resources and time for both patient and medical doctor. Moreover, this could be used for aiding the medical experts in the tailoring of the treatment to the patient, towards a personalized medicine approach.

Simultaneously, the emerging availability of electronic medical records (EMR) enables the storage of great amounts of information that can be used to extract insightful knowledge. Data mining is a rapidly growing field that focuses on developing the techniques necessary for using this information in an insightful way.

The analysis of an outcome of interest is usually performed using survival analysis methods such as the Kaplan-Meier estimator [3] and the Cox model [4]. These methods however, are only able to deal with time static variables. For dealing with time-varying variables, methods such as the extended Cox model [5] have been introduced. However, they are not appropriate for dealing with biomarkers.

Joint models have been presented in the literature as a useful approach for handling these types of analysis, having been used in a wide range of medical studies, being the most common disease areas in Cancer and HIV/AIDS [6]. However, none has yet explored the field of rheumatic diseases, to our knowledge.

The main goal of this thesis is to propose a data mining approach based on joint models to infer relationships between time-to-event and longitudinal EMR data, retrieved from the Rheumatic Disease Portuguese Register, Reuma.pt [7], for studying the predictors of failure.
of the first biologic therapy for patients with spondyloarthritis and verify the applicability of joint models for the study of therapeutic response in rheumatic diseases.

Section 2 presents spondyloarthritis and its associated context, as well as the theoretical concepts of the statistical methods that will be used for its analysis. The preprocessing of the data, and implementation of the modeling strategy is described in Section 3. The results obtained from this approach and conclusions are presented in Section 4 and Section 5, respectively.

2. Background

2.1. Spondyloarthritis

Spondyloarthritis (SpA) is the name given to a family of inflammatory rheumatic diseases that share distinctive pathophysiologic, clinical, radiographic, and genetic features. This includes ankylosing spondylitis (AS), which is considered the prototype of this group, psoriatic arthritis, reactive arthritis, enteroptropic arthritis and the so-called undifferentiated SpA.

AS is characterized by chronic inflammation affecting predominantly the axial skeleton. Although its pathogenesis is poorly understood, there is a strong association between AS and the human leukocyte antigen B27 (HLA-B27) and the typical age at onset of this condition is at the second or third decade of life [8].

In Portugal, the first population-based study on rheumatic diseases in Portugal, EpiReumaPt, reported their national health survey results in 2015, revealing a general SpAs prevalence of 1.6%, and a prevalence of 2.0% and 1.2% for women and men, respectively [1].

The socioeconomic impact can be rather high for these conditions. A recent study [9] revealed that AS has a total annual economic impact of €639 million euros in Portugal. This value includes not only the disease related costs for the patient and the national health system, but also the economical impact of the lost work days.

Clinical monitoring of a disease is of extreme importance, not only to understand disease evolution but also to better assess patient response to treatment and guide therapeutic decisions.

Laboratory exams include erythrocyte sedimentation rate (ESR), levels of C-reactive protein (CRP), which are markers of inflammation, and other laboratory data that are considered to show relevant alterations.

Functional ability can be evaluated using the Bath Ankylosing Spondylitis Functional Index (BASFI) score and activity disease can be evaluated using the Bath Ankylosing Spondylitis Status in Ankylosing Spondylitis (BASDAI) score or, the more recently developed Ankylosing Spondylitis Disease Activity Score (ASDAS) [10].

The BASFI evaluates the degree of disability in AS patients, and is measure on a 0-10 scale, with higher scores indicate greater functional impairment [11].

The BASDAI is a measure for disease activity, and ranges from 0 to 10, where scores of 4 or more suggest an active disease and sub optimal disease control [12].

More recently, the ASDAS has been developed with the goal of improving the evaluation of disease activity in AS, by creating a score more sensitive to change and that covers disease activity in a broader way. The ASDAS has two different formulas: ASDAS-CRP (which uses the C-reactive protein) and ASDAS-ESR (using the erythrocyte sedimentation rate), being ASDAS-CRP the preferred one. These formulas are presented in Table 1 [13].

Table 1: Formulas for the calculation of the two forms of the ASDAS criterium: ASDAS-CRP and ASDAS-ESR.

<table>
<thead>
<tr>
<th>Score</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASDAS-CRP</td>
<td>0.12BP + 0.06DMS + 0.11PG + 0.07PP + 0.56ln(CRP + 1)</td>
</tr>
<tr>
<td>ASDAS-ESR</td>
<td>0.06DMS + 0.07DMS + 0.11PG + 0.09PP + 0.29√ESR</td>
</tr>
</tbody>
</table>

CRP, C-reactive protein (mg/L); ESR, erythrocyte sedimentation rate (mm/h); BP, Back Pain (BASDAI Question 2); DMS, Duration Morning Stiffness (BASDAI Question 6); PG, Patient Global (0-10); PP, Peripheral Pain/Swelling (BASDAI Question 3)

Treatment of SpA should be tailored according to the patient, taking into account the patient’s signs, symptoms and characteristics, being the most common goal reaching a state of inactive disease.

Treatment options can include physical therapy, non-steroidal anti-inflammatory drugs (NSAIDs), disease-modifying anti-rheumatic drugs (DMARDs) and, if patients remain in a high disease activity stage when trying the referred options, treatment with biologic agents, namely tumor necrosis factor (TNF) inhibitors and Interleukin (IL)-17 or IL-23 inhibitors.

2.2. Time-to-event analysis

Survival analysis, or time-to-event analysis, is the collection of statistical procedures for the analysis of survival data, that is, analysis in which the the outcome variable of interest is time until an event occurs.

Let’s denote $T$, a random, non-negative, continuous variable representing the subject’s survival time and let $t$ be an observed value of $T$.

A basic quantity in survival analysis is the hazard function, which can be interpreted as the instantaneous potential per unit time for the event to occur, given that the individual has survived to time $t$. It is given by

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}.$$  \hspace{1cm} (1)

The main feature that distinguishes survival data from other types of data is the possible presence of censored survival times.

Considering a specific individual, let $T_i$ be the random variable representing its true survival time and $C_i$ the potential censoring time. The censoring indicator variable, $\delta_i$ is defined as $\delta_i = I(T_i \leq C_i)$, where $I(.)$ is the indicator function.

The Cox proportional hazards model [4] allows us to not only estimate the hazard function but also to explore how the survival of a group of subjects depends on the values of one or more explanatory variables.
Let \( \mathbf{x} = (x_1, x_2, ..., x_4) \) be the vector of explanatory variables of an individual and \( \mathbf{\beta} = (\beta_1, \beta_2, ..., \beta_k) \) the vector of its correspondent unknown regression coefficients. The hazard function is given by

\[
h(t; \mathbf{x}) = h_0(t) \exp(\mathbf{x}^T \mathbf{\beta}),
\]

where \( h_0(t) \) is the baseline hazard function, that represents the hazard for a patient when its vector of explanatory variables is equal to zero (\( \mathbf{x} = 0 \)).

It is possible to extend the previously presented Cox model for handling time-dependent variables [5]. This model is referred to as extended Cox model. However, this model is not theoretical appropriate to deal with biomarkers since it assumes that the time-dependent variables are predictable processes, measured without error and with a full path completely known.

2.3. Longitudinal analysis

Longitudinal data can be defined as the data obtained from multiples measurements of subjects throughout time.

Linear mixed effects (LME) models are one common way of modelling this data. In a LME, the subject’s response is assumed to follow a linear regression model for handling time-dependent variables - this is usually done by combining a survival model with a mixed-effects model [14].

Let \( Y_i \) be the \( n_i \)-dimensional response vector for subject \( i \). In general, a linear mixed-effects model satisfies

\[
Y_i = X_i \mathbf{\beta} + Z_i \mathbf{b}_i + \mathbf{\epsilon}_i,
\]

where \( \mathbf{\beta} \) a \( p \)-dimensional vector that contains the fixed effects, \( \mathbf{b}_i \) a \( q \)-dimensional vector containing the random effects and \( \mathbf{\epsilon}_i \) a \( n_i \)-dimensional vector of random errors, \( X_i \) is and \( Z_i \) are \( (n_i \times p) \) and \( (n_i \times q) \) fixed-effects and random effects design matrices, respectively, \( D_i \) is a \( (q \times q) \) positive-definite covariance matrix, \( \Sigma_\epsilon \) is a \( (n_i \times n_i) \) positive-definite covariance matrix that depends on \( i \) through its dimension \( n_i \) (but the set of unknown parameters in it will not depend upon \( i \)). The \( \mathbf{\epsilon}_i \) is normally distributed with mean zero and covariance matrix \( D_i \) and \( \mathbf{b}_i \) is normally distributed with mean zero and covariance matrix \( \Sigma_\beta \). Both \( \mathbf{b}_i \) and \( \mathbf{\epsilon}_i \) are assumed to be independent between each other and between groups [14].

2.4. Joint models for longitudinal and time-to-event data

The basic idea of joint models is to perform combined analysis, in which a relative risk model is estimated for the time-to-event outcome, taking into account the effect of the longitudinal data measurements - this is usually done by combining a survival model with a mixed-effects model [15].

The first step is modelling the continuous longitudinal outcomes with linear mixed effects models. Let’s denote by \( \mathbf{y}_k \) the \((n_{ki} \times 1)\) longitudinal response vector for the \( k \)-th outcome \((k = 1, ..., K)\) and the \( i \)-th subject that is composed by elements \( y_{ki} \) which represent the value of the \( k \)-th longitudinal outcome taken at time point \( t_{ki} \). Let \( \mathbf{b}_{ki} \) be a vector of random effects and \( \mathbf{\beta}_k \) a vector of fixed effects. We have that the conditional expectation of \( \mathbf{y}_k \) given \( \mathbf{b}_{ki} \), \( \eta_{ki}(t) \), is modeled through the LME model as

\[
\eta_{ki}(t) = \mathbf{x}_{ki}^T \mathbf{\beta}_k + \mathbf{z}_{ki}^T \mathbf{b}_{ki},
\]

where \( \mathbf{x}_{ki} \) and \( \mathbf{z}_{ki} \) are the design vectors for the random and fixed effects, respectively.

Let \( T_i^* \) be the true event time for the \( i \)-th subject. We can now postulate the relative risk model for the survival process as

\[
h_i(t \mid \mathcal{M}_i(t), \mathbf{w}_i(t)) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T_i^* < t + \Delta t \mid T_i^* \geq t, \mathcal{M}_i(t), \mathbf{w}_i(t))}{Pr(t \leq T_i^* < t + \Delta t \mid T_i^* \geq t, \mathcal{M}_i(t), \mathbf{w}_i(t))},
\]

where \( \mathcal{M}_i(t) = \{\mathcal{M}_{1i}(t), ..., \mathcal{M}_{ki}(t)\} \) and \( \mathcal{M}_{ki}(t) = \{\eta_{ki}(s), 0 \leq s < t\} \) denotes the history of the true unobserved longitudinal process up to time point \( t \), \( h_0(.) \) denotes the baseline risk function, \( \mathbf{w}_i(t) \) is a vector of exogenous covariates with a corresponding vector of regression coefficients \( \gamma \). The \( \mathcal{M}_i(t) \) functions, parametrized by vector \( \mathbf{\alpha}_{ki} \), specify which components of each longitudinal outcome will be present in the relative risk model, allowing up to \( L_k \) functional forms for each of \( K \) longitudinal outcomes. The parameters contained in \( \mathbf{\alpha}_{ki} \) quantify the effect of the correspondent underlying longitudinal outcome to the risk for an event.

One of the basic approaches for the functional form is to model the hazard of the event as having an association only with the current value of the longitudinal outcome at the same time point. Considering a single outcome, this is given by \( f(\mathbf{\alpha}, \mathbf{w}_i(t), \mathbf{b}_i, \mathcal{M}_i(t)) = \alpha \eta_{ki}(t), \) where \( \alpha \) is the strength of association parameter, that indicates the change in the log hazard when there is a unit change in the subject’s longitudinal outcome value.

Estimation of joint models is performed by exploiting the full joint likelihood that is derived from the joint distribution of the longitudinal and survival outcomes. The main methods for this estimation follow either a frequentist or a Bayesian paradigm.

3. Proposed methodology

3.1. Spondyloarthritits patients on biologic therapies: data description and pre-processing

The data used in this thesis work was retrieved from the Rheumatic Diseases Portuguese Register, Reuma.pt [7], on the 22nd of July, 2019. This register was developed by the Portuguese Society of Rheumatology and has been active since June 2008, containing information
retrieved on a routine basis of rheumatic patients in Portugal, receiving biological therapies. The follow up of patients through this registry enables the monitoring of treatment efficacy, safety and comorbidities. Although Reuma.pt also contains patients with several rheumatic diseases, the focus of this work was on patients with spondyloarthritis. The database includes information regarding patients and patient visits, such as identification data, demographic data, previous medical history, comorbidities, laboratory results, past and current therapies, adverse events and disease activity scores, to name a few.

The goal of this work is to perform a survival analysis that takes into account both time-independent and time-dependent variables and understand how these impact the outcome of interest. Therefore, three types of variables are needed: time-independent (baseline) variables, time-dependent variables and time-to-event variables. The last ones are not directly found in the database and therefore needed to be processed from the existing data.

Our event of interest is the failure of the first biological therapy for each patient, where failure was defined as the discontinuation of the biological therapy due to inefficacy or adverse events. In this context, the time-to-event variables indicate if the biologic failed or if the patient is censored and the time until the occurrence of either failure of the biologic or censoring of the patient - we will refer to them as failure index and time to failure, respectively.

The data extracted from the database had to go through several pre-processing steps in order to reach a format compatible with the models to be fitted. Firstly, a set of variables to be considered was selected taking into account three main aspects: level of missing values, relevance to the study and variables equivalence. Variables with more than 60% of missing values were not considered, nor were variables that were considered to be irrelevant for the goal of our study, taking into account feedback from medical specialists. Furthermore, some variables available in the raw dataset were equivalent in the sense that they represented the same information.

After this set of assumptions and processing steps, we obtained our initial dataset, that consisted of the following variables:

1. **Time-to-event variables**: time to failure and failure index;
2. **Time-independent variables**: sex, marital status, year of diagnosis, age at diagnosis, year of disease beginning, age at disease beginning, year of start of the first biologic, age at start of the first biologic, years from diagnosis to start of the first biologic, disease years until start of the first biologic, human leukocyte antigen B27 (HLA-B27), employment status before disease, employment status at baseline, years of education, smoking habits, alcohol consumption habits, weight, height, body mass index (BMI), number of pathologies, biologic therapy, concomitant disease-modifying antirheumatic drug (DMARD) at baseline, concomitant corticoid at baseline, baseline CRP, baseline ESR, baseline BASDAI, baseline BASFI and baseline ASDAS;
3. **Time-dependent variables**: C-reactive protein (CRP), erythrocyte sedimentation rate (ESR), Bath ankylosing spondylitis disease activity index (BASDAI), Bath ankylosing spondylitis functional index (BASFI) and ankylosing spondylitis disease activity score (ASDAS).

The ASDAS we refer to here, and henceforward, is the one that corresponds to the ASDAS-CRP.

A thorough inspection of all variables was made, in order to identify any incomplete, incorrect or incoherent values. If possible, by crossing information and with the help and consultation of medical professionals, the values were corrected, but whenever it was impossible to draw conclusions, the observations were eliminated.

In the presented initial dataset, not all patients have every baseline variable available. This poses an issue and is a challenge that needs to be dealt with in most studies that use data from clinical settings, since most methods of variable selection and statistical models, cannot handle missing values. Therefore, to further proceed with our analysis, we need to understand the different approaches that can be used to handle this problem, according to our needs. Common approaches include performing complete-case analysis, removing individuals with incomplete data for a subset of covariates and multiple imputation techniques.

We decided not to perform any imputation techniques, since the imputation of baseline variables could introduce a high bias in the estimates of our models. On the other hand, there is always a high interest in keeping the most amount of data possible, to avoid wasting relevant information. Since there was no obvious choice regarding which approach would be the most appropriate, and with the goal of enabling the drawing of valuable insights from the data, the decision of creating four different datasets, according to different approaches, was made. Furthermore, this allows to study how the strategy for handling missing data and the resulting data differences can influence the modelling process and the obtained results. The overall process for the creation of these datasets is depicted in Fig. 1.

The first approach consisted in keeping only the patients for which all baseline variables were available, that is, keeping only the complete cases.

The second approach was to consider only the variables that had less than 40% of missing values and then keeping the complete cases of those variables. The percentage of 40% was chosen since it seemed to originate a good balance between number of eliminated variables and number of eliminated patients.

The third approach consisted in fitting an univariate Cox model for each initial baseline variable, and then keeping only the ones that were statistically significant in those models, at a 5% level. After that, once again,
the complete cases of those variables were kept.

Finally, the last approach was to keep only variables that were considered as clinically relevant by expert medical doctors knowledge and insight from literature research, where predictors of biologic drug survival in spondyloarthritides were studied [16; 17; 18]. The variables selected to be considered were: sex, disease years to first biologic, age at start of the first biologic, education years, baseline CRP, baseline BASDAI and baseline BASFI. Variables age at start of the first biologic and baseline BASDAI were later dropped due to violation of the proportional hazards assumption.

3.2. Statistical models’ implementation

For the initial dataset, both an overall survival curve and curves for survival according to the biologic therapy were fitted, using the Kaplan-Meier estimator [3]. Regarding the four processed datasets, the same approach was used for all, which will be described in detail below.

The first step in the analysis was to perform variable selection for the baseline covariates. This will remove unnecessary predictors that can add noise to the estimations.

Five different methods were used, to compare and study the variability of the obtained results. These were backward stepwise selection using AIC, forward stepwise selection using AIC, best subset selection using a primal-dual active set approach, lasso regression and the stepwise likelihood ratio variable selection strategy presented in Chapter 3.6.1 of [19]. Despite this, the variables obtained from the stepwise likelihood ratio variable selection were the ones ultimately selected for the next steps of the analysis, namely for building the survival sub-model.

This variable selection was not performed for dataset D, since the variables of interest had already been selected by the medical experts for this specific case.

Then, a Cox model for the survival sub-model was fitted, using the selected baseline covariates, constituting the survival sub-model. For each of the time-dependent variables, seven different LME models were fitted. The one with the better fit according to AIC and BIC criteria was chosen as the time-dependent sub-model. The formulas of the different LME models that were fitted and the corresponding name we will be using to refer to them are presented in Table 2.

<table>
<thead>
<tr>
<th>Model</th>
<th>( y_i(t_{ij}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear and random intercept</td>
<td>( \beta_0 + \beta_1 t_{ij} + \beta_2 + \epsilon_{ij} )</td>
</tr>
<tr>
<td>Linear and random slope</td>
<td>( \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij} + \beta_3 + \epsilon_{ij} )</td>
</tr>
<tr>
<td>Cubic and random intercept</td>
<td>( \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 t_{ij}^3 + \beta_4 + \epsilon_{ij} )</td>
</tr>
<tr>
<td>Cubic and random slope</td>
<td>( \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 t_{ij}^3 + \beta_4 + \beta_5 t_{ij} + \epsilon_{ij} )</td>
</tr>
<tr>
<td>Spline and random intercept</td>
<td>NC(t_{ij}, 2, ( \beta_0, \beta_1, \beta_2 ) T, ( \beta_3, \beta_4 ) T + \epsilon_{ij} )</td>
</tr>
<tr>
<td>Spline and random slope</td>
<td>NC(t_{ij}, 2, ( \beta_0, \beta_1, \beta_2 ) T, ( \beta_3, \beta_4 \beta_5 ) T + \epsilon_{ij} )</td>
</tr>
</tbody>
</table>

NC, natural cubic spline function; \( \beta \), fixed effects; b, random effects; \( t \), time.

Having both submodels, the same joint models were fitted, with R packages JM [20] and JMbayes [21]. The former estimates the model under a maximum likelihood approach and the latter under a Bayesian approach, more specifically, using Markov chain Monte Carlo (MCMC) algorithms.

The R package JMbayes also enables the fitting of multivariate joint models. Considering variables CPR and ESR are both measurements of inflammation, and that ASDAS uses CPR and BASDAI elements in its composition, we chose not to fit together CPR with ESR nor ASDAS with CPR and BASDAI. So two different combinations of time-dependent variables were considered for the multivariate joint models: CPR, BASDAI and BASFI, and BASFI and ASDAS.

The equivalent models, both univariate and multivariate, were also fitted with an extended Cox model, for enabling the comparison of both methods.
In case the survival submodel contains any of the variables baseline CRP, baseline ESR, baseline BASDAI, baseline BASFI or baseline ASDAS, these baseline variables will be dropped when the correspondent time-dependent variable is present in the univariate joint model or extended Cox model. This will enable us to compare the effect of the variable in its baseline form with its time-dependent form. Similarly, if more than one of the 5 baseline variables are present in the survival sub-model, a multivariate joint model or extended Cox model will also be fitted with those variables in the time-dependent form and dropping the baseline form from the survival sub-model.

The overall process for the fitting of the joint models and extended Cox models is schematically presented in Fig. 2 and Fig. 3, respectively, where it is also possible to observe the numbers given to the models that will be fitted for every dataset.

The exhaustive tests performed try to cover several types of strategies and models, and aim at identifying key covariates involved in the prognosis of the disease, namely in the response to treatment. Indeed, the rationale for this approach is to comprehensively span the described methods, since for this specific dataset of Reuma.pt there are no prior studies that focus on the unveiling of specific markers for the prognosis of the patient’s therapy response.

All the analysis was performed using R software [22], particularly the following packages: MASS [23], BeSS [24] and glmnet [25] for the forward and backward stepwise variable selection, best subset selection and lasso regression, respectively; car [26] for multicollinearity testing; survival [27] for the Kaplan-Meier curves, Cox model, extended Cox model and proportional hazards testing; survminer [28] for the plotting of survival curves; nlme [29] for fitting the linear mixed effects models; JM [20] and JMbayes [21] for fitting the joint models.

4. Results

4.1. Initial SpA dataset

The survival probability curve, $\hat{S}(t)$, of the first biologic therapy for the overall population, from the initial dataset, obtained with the Kaplan-Meier estimator is presented in Fig. 4, where vertical ticks along the curve indicate censored patients. We can observe that the slope of the curve is higher at the initial months, indicating that there are more failures closer to the beginning of the therapy.

A comparison of the survival probabilities between the different biologic therapies can be seen in Fig. 5, where it is possible to observe a clear distinction between the curves of the different biologic drugs. The $p$-value of the log-rank test is also presented in the figure and since it is lower than 0.05, it indicates that the biologic drug curves differ significantly in survival, at a 5% level.

From analysing the curves, it seems that Certolizumab is the biologic therapy with better survival, since it is the one that remains at the highest survival
probability value by the end of its curve. However, the fact that there is a very small number of observations for this drug, in comparison to others, does not enable us to make a fair comparison. Regarding the three drugs with a higher number of patients (Adalimumab, Etanercept and Infliximab), Etanercept appears to be the one with better survival prognostic. This result is in accordance with a cohort study [30], where it was found that patients taking Adalimumab or Etanercept had longer drug survival as compared to those taking Infliximab.

4.2. Comparison of results: datasets A, B, C and D

The first main step of the modelling process consisted of selecting the baseline variables of interest.

The comparison of the selected variables takes into consideration only datasets A, B and C, since it was not performed for dataset D. The percentage of times a variable was selected, considering the five variable selection approaches tested, for each dataset, is presented in Table 3, along with the average of times it was selected overall.

Even though not all variables were present in every dataset, the variability in the covariates that are selected for each set of data is still noticeable. This indicates that the initial process of handling the missing values and the initial selection of variables to be considered at this stage have a somewhat elevated influence on the results that will later be obtained – the results are sensitive to the parameter choice. This difference may be justified by the existence of different variables and even different patients, even if many are common between datasets.

Only one variable is selected by all methods, for all datasets where it is considered – years of education. The variables that were selected, on average, in at least more than 50% of the methods used for variable selection are: year of disease beginning, age at start of the first biologic, baseline BASFI and baseline ASDAS.

Table 4 shows the the sign of the coefficient for each of the covariates that were included in the survival submodel, for every dataset. The sign of the respective coefficient indicates the effect of such covariate on the outcome of interest, that, for us, is the failure of the first biologic therapy. A positive sign indicates that the variable increases the risk of failure, for higher values of such variable (for a continuous variable) or for that value, in comparison to the reference level (for a categorical variable); a negative coefficient indicates the opposite, a decrease in the risk of failure. It can be noticed that the sign of the coefficient is coherent between datasets, for all variables, even if the number of datasets where that variable is present differs.

According to the results obtained in the Cox regression, the factors that indicate a good prognosis for the biologic drug survival are being a male, starting the biologic therapy at an older age, having a larger time interval between disease start and initiation of the first biologic, and being HLA-B27 positive. On the contrary, having the disease beginning or starting the biologic therapy in more recent years, a larger number of patients taking Adalimumab or Etanercept had longer drug survival as compared to those taking Infliximab.

4.2. Comparison of results: datasets A, B, C and D

The first main step of the modelling process consisted of selecting the baseline variables of interest.

The comparison of the selected variables takes into consideration only datasets A, B and C, since it was not performed for dataset D. The percentage of times a variable was selected, considering the five variable selection approaches tested, for each dataset, is presented in Table 3, along with the average of times it was selected overall.

Even though not all variables were present in every dataset, the variability in the covariates that are selected for each set of data is still noticeable. This indicates that the initial process of handling the missing values and the initial selection of variables to be considered at this stage have a somewhat elevated influence on the results that will later be obtained – the results are sensitive to the parameter choice. This difference may be justified by the existence of different variables and even different patients, even if many are common between datasets.

Only one variable is selected by all methods, for all datasets where it is considered – years of education. The variables that were selected, on average, in at least more than 50% of the methods used for variable selection are: year of disease beginning, age at start of the first biologic, baseline BASFI and baseline ASDAS.

Table 4 shows the the sign of the coefficient for each of the covariates that were included in the survival submodel, for every dataset. The sign of the respective coefficient indicates the effect of such covariate on the outcome of interest, that, for us, is the failure of the first biologic therapy. A positive sign indicates that the variable increases the risk of failure, for higher values of such variable (for a continuous variable) or for that value, in comparison to the reference level (for a categorical variable); a negative coefficient indicates the opposite, a decrease in the risk of failure. It can be noticed that the sign of the coefficient is coherent between datasets, for all variables, even if the number of datasets where that variable is present differs.
education years, and higher values of CRP or BASFI at baseline are all predictors of a greater risk of failure of the first biologic therapy.

Patients of male sex [17], HLA-B27-positive [31] and with longer disease duration [16] have been reported in literature as good predictors of biologic drug survival, which concurs with the results obtained in the Cox models. On the other hand, older age at the start of the biologic therapy [31] has been stated as increasing the risk of failure of the biologic, contrarily to what was found in our data. We could interpret our result by hypothesizing that older patients are more compliant with their perceived efficacy of the therapy, or that their symptoms are more intense than younger people and that the relative improvement of symptoms is more noticeable, therefore increasing their satisfaction levels, and decreasing the chances of therapy switch.

Regarding the predictors that were found to increase the risk of failure in our study, starting treatment in more recent years [30] and higher values of BASFI at baseline [16] were likewise found to be predictors of biologic drug discontinuation in research publications. A higher number of education years [18] was reported in one study as decreasing the risk of failure of the therapy, differing from our results. Again, we could make an hypothesis, and say that patients with a higher academic level are more comfortable with exposing their discontent with the lack of therapy response, or even that they are more aware of new therapeutic options, and for that reason, request a switch of the therapy more often.

Higher values of CRP at baseline were found to increase [32] the hazard of biologic therapy failure in some publications but to have the opposite effect [30; 18] in others.

Given the elevated number of models fitted, and to aid the drawing of comprehensive conclusions, Table 5 and Table 6 were built, regarding the joint and extended Cox models, respectively. These show, for each dataset and every variable present in the model, the percentage of models (relative to the total number of fitted models for each dataset) in which the covariate was statistically significant, the percentage of models in which the variable has a positive regression coefficient and the percentage of models in which the variable has a negative regression coefficient. Furthermore, the average of these percentages was calculated to obtain an overall view of the most common behaviour of each variable, gathering the information from all datasets.

This consensus or ensemble approach is conducted as to aiding the identification of the most significant variables. The rationale is that, if a feature always appears as significant, independently of the specific chosen model, then there is evidence that such feature is associated with the outcome. The equivalent reasoning is applicable for identifying the covariates’ effect on the event of interest, namely, its positive or negative contribution for the risk of the therapy failure.

Focusing on the time independent variables, and starting with the covariate that represents male sex, we see that this variable is not statistically significant in any joint or extended Cox model. This is coherent with what was observed in the Cox model, where sex was not a statistical significant predictor for our outcome. Regarding the effect of the variable on the event of interest, being male was more frequently a good predictor of biologic therapy survival than a bad predictor, although this ratio was very small for the joint models.

The year of disease beginning was statistically significant for most models it was present in, even if only one dataset analysed this variable. Its associated coefficient was positive for all the extended Cox models and for an average of 92% of the joint models, which is consistent with what had been obtained in the Cox models, indicating that patients with the beginning of the disease in more recent years have a higher risk of treatment failure.

The year of start of the first biologic appears as a statistically significant predictor in most joint models and in all extended Cox models. For the majority of models, a biologic therapy initiated in more recent years increases the risk of its failure.

The year interval between disease beginning and start of first biologic is only statistically significant in a small percentage of joint models and it shows no significance in any of the Extended Cox models, mimicking the results for the Cox model. The coefficient sign for this covariate is coherent among all models, indicating that a larger year interval reduces the hazard for the biologic therapy failure.

Being HLA-B27 positive is statistically significant as a predictor for biologic therapy failure in approximately 80% of all models and is always associated with a decreased risk of failure when in comparison with HLA-B27 negative patients.

The number of years of education has statistical significance in roughly 90% of all joint and extended Cox models, and a higher number of education years increases the hazard of biologic failure, for all Cox, extended Cox and joint models.

The value of CRP at baseline has an associated positive regression coefficient in all joint models and in 80% of extended Cox models, indicating an increased risk of failure for higher CRP values, which is also verified in the Cox model. This covariate was not statistically significant in the Cox regression but is significant in approximately 40% of the joint and extended Cox models.

The baseline BASFI is statistically significant in less than half of the joint and extended Cox models, even though it is always statistically significant in the survival sub-models fitted with a Cox model. This variable appears to be a predictor for increased risk of biologic therapy failure in most joint and extended Cox models, which is concordant with the Cox models’ results.

Regarding the time-dependent variables, we notice that, for the joint models, variables CRP, ESR, BASDAI
and ASDAS appear as being statistically significant in most models. Furthermore, all of them are predictors of increased therapy failure for all the joint models that were fitted. Variable BASFI is only statistically significant in approximately half of the joint models that were fitted. Variable ESR shows statistically significant in 75% of the models, BASFI in 58% of the models and CRP in only 38% of them. In all models, all five time-varying covariates are predictors of increased risk of biologic failure.

5. Conclusions

Joint models are statistical models that are able to analyse both time static and time-varying variables and therefore infer relationships between time-to-event and longitudinal data, widely present in electronic medical records. In this work, this modelling approach was selected as to investigate the research question of the biologic drug survival and its predictors, for SpA patients in Portugal. Furthermore, the insights obtained throughout the process that culminated in the fitting of these models, are also highly valuable.

This work was developed using data of spondyloarthritits patients from Reuma.pt that was yet unexplored. The entire pre-processing work performed for enabling its use originated a dataset that can be utilized in future studies that wish to investigate research questions regarding this group of diseases.

The variable selection process appears to be sensitive to this data pre-processing step, and dependent of which variables and patients are described in the dataset.

The tested methods for variable selection originated quite different results for the same set of data. The process of selection of covariates should be analysed carefully, as fully automated methods may not be the

Table 5: Percentage of models in which a variable had a statistically significant, percentage of models in which a variable had a positive coefficient sign and percentage of models in which a variable had a negative coefficient sign, for the covariates present in the joint models fitted for datasets A, B, C and D, and average of those percentages across all datasets.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Dataset</th>
<th>% s.s.</th>
<th>% pos.</th>
<th>% neg.</th>
<th>% s.s.</th>
<th>% pos.</th>
<th>% neg.</th>
<th>% s.s.</th>
<th>% pos.</th>
<th>% neg.</th>
<th>% s.s.</th>
<th>% pos.</th>
<th>% neg.</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year of disease beginning</td>
<td>A</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>Year of start of the first biologic</td>
<td>B</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>Age at start of the first biologic</td>
<td>C</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>Disease years until start of the first biologic</td>
<td>D</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>HLA-B27</td>
<td>A</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>CRP</td>
<td>B</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>ESR</td>
<td>C</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>BASDAI</td>
<td>D</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>ASDAS</td>
<td>Average</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
</tr>
</tbody>
</table>

% s.s., percentage of models in which variable is statistically significant; % pos., percentage of models in which variable has positive coefficient; % neg., percentage of models in which variable has negative coefficient. HLA-B27, human leucocyte antigen B27; CRP, C-reactive protein; ESR, erythrocyte sedimentation rate; BASDAI, Bath ankylosing spondylitis disease activity index; BASFI, Bath ankylosing spondylitis functional index; ASDAS, ankylosing spondylitis disease activity score.

Table 6: Percentage of models in which a variable had a statistically significant, percentage of models in which a variable had a positive coefficient sign and percentage of models in which a variable had a negative coefficient sign, for the covariates present in the joint models fitted for datasets A, B, C and D, and average of those percentages across all datasets.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Dataset</th>
<th>% s.s.</th>
<th>% pos.</th>
<th>% neg.</th>
<th>% s.s.</th>
<th>% pos.</th>
<th>% neg.</th>
<th>% s.s.</th>
<th>% pos.</th>
<th>% neg.</th>
<th>% s.s.</th>
<th>% pos.</th>
<th>% neg.</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year of disease beginning</td>
<td>A</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>Year of start of the first biologic</td>
<td>B</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>Age at start of the first biologic</td>
<td>C</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>Disease years until start of the first biologic</td>
<td>D</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>HLA-B27</td>
<td>A</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>CRP</td>
<td>B</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>ESR</td>
<td>C</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>BASDAI</td>
<td>D</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>ASDAS</td>
<td>Average</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
</tr>
</tbody>
</table>

% s.s., percentage of models in which variable is statistically significant; % pos., percentage of models in which variable has positive coefficient; % neg., percentage of models in which variable has negative coefficient. HLA-B27, human leucocyte antigen B27; CRP, C-reactive protein; ESR, erythrocyte sedimentation rate; BASDAI, Bath ankylosing spondylitis disease activity index; BASFI, Bath ankylosing spondylitis functional index; ASDAS, ankylosing spondylitis disease activity score.
most appropriate ones for establishing which variables should be included in the statistical model. A wise approach consists of aiming for a balance between statistical significance and clinical significance, always taking into account the goal of the study. It was shown that joint models, and namely the functions implemented in the R software packages JM and JMBayes, can be successfully used for the simultaneous analysis of time-to-event and longitudinal data. Furthermore, with this work, it was possible to identify the potential predictors of biologic therapies failure for this Portuguese population of spondyloarthritis patients. This can aid the prognosis of these rheumatic diseases and potentially predict the most adequate treatment option according to the patient’s characteristics.

Considering the main challenges and difficulties experienced throughout the development of this thesis, some proposals for future work include the exploration of the SpA data with recurrent event analysis methods, using all available history of biological therapies for each patient and using different approaches for the handling of missing data such as multiple imputation techniques. Regarding the joint modelling approach, the choice of the LME function for describing the biomarker trajectories should be further explored, as well as the different functional forms available, that specify the association between the longitudinal biomarker and the hazard function of the event. Lastly, the already implemented functions for making predictions based on the joint models should be investigated, as these predicted outcomes can aid physicians in the tailoring of decision making based on the characteristics of the individuals, with the aim of optimizing the therapeutic strategy and achieving personalized healthcare solutions.

Acknowledgements
The author would like to thank to her thesis supervisors, Prof. Susana Vinga and Prof. Alexandre Carvalho for the incredible advice and support during the development of this thesis. A special thanks to everyone involved in the PREDICT project, specially to Dr. Helena Canhão and Dr. Ana Rodrigues for their valuable insights.

References


