Journal Pre-proof

Large scale prediction of sick leave duration with nonlinear survival analysis algorithms

Javier Béjar, Raquel Pŕez, Armand Vilalta, Sergio Álvarez-Napagao, Dario Garcia-Gasulla

PII: S0957-4174(22)00225-1
DOI: https://doi.org/10.1016/j.eswa.2022.116760
Reference: ESWA 116760

To appear in: *Expert Systems With Applications*

Received date : 1 April 2021
Revised date : 18 January 2022
Accepted date : 24 February 2022

Please cite this article as: J. Béjar, R. Pŕez, A. Vilalta et al., Large scale prediction of sick leave duration with nonlinear survival analysis algorithms. *Expert Systems With Applications* (2022), doi: https://doi.org/10.1016/j.eswa.2022.116760.

Large Scale Prediction of Sick Leave Duration with Nonlinear Survival Analysis Algorithms

Javier Béjar[a,b,c] (bejar@cs.upc.edu), Raquel Pérez[a,b] (raquel.perez@bsc.es), Armand Vilalta[a] (armand.vilalta@bsc.es), Sergio Álvarez-Napagao[a,b] (sergio.alvarez@bsc.es), Dario Garcia-Gasulla[a] (dario.garcia@bsc.es)

[a] Barcelona Supercomputing Center (BSC), Spain

[b] Computer Science Department. Universitat Politècnica de Catalunya. Barcelona, Spain

[c] IDEAI. Universitat Politècnica de Catalunya. Barcelona, Spain

Corresponding Author:

Javier Béjar

Computer Science Department. Universitat Politècnica de Catalunya.

Barcelona, Spain

Tel: (34) 93-4137879

Email: bejar@cs.upc.edu

Large Scale Prediction of Sick Leave Duration with Nonlinear Survival Analysis Algorithms

Javier Béjar[a,b,c,*], Raquel Pérez[a,b], Armand Vilalta[a], Sergio Álvarez-Napagao[a,b], Dario Garcia-Gasulla[a]

[a] *Barcelona Supercomputing Center (BSC), Spain*
[b] *Computer Science Department. Universitat Politècnica de Catalunya. Barcelona, Spain*
[c] *IDEAI. Universitat Politècnica de Catalunya. Barcelona, Spain*

Abstract

The management of sick leaves is a critical task that public and private health systems carry out. This enables the good care of sick workers and guarantees a safe return to their jobs. Most health systems enforce regulations that establish the duration of sick leave according to general rules for groups of diagnoses. However, these regulations do not account for the particularities of workers. On the one hand, an early return to work is sometimes possible, but this does not happen unless the worker pro-actively requests it. On the other hand, the worker's health condition could demand for one or more leave extensions, but the system requires mandatory and sometimes unnecessary follow-ups, adding nuisance to patients and overhead to health systems. In both cases, the lack and excess of action by the health system represents extra costs for society. This paper proposes the analysis of a voluminous historical dataset of sick leaves (including medical and personal data) to predict the duration of future sick leaves for patients. The data mining process is performed for a large number of diagnoses to assess the possibility of using data driven models for broad decision-making. The nature and characteristics of the data makes it difficult to obtain models using classical methods, which is why the analysis focuses on non-

*Corresponding Author: Javier Béjar; E-mail: bejar@cs.upc.edu.
Email addresses: bejar@cs.upc.edu (Javier Béjar), raquel.perez@bsc.es (Raquel Pérez), armand.vilalta@bsc.es (Armand Vilalta), sergio.alvarez@bsc.es (Sergio Álvarez-Napagao), dario.garcia@bsc.es (Dario Garcia-Gasulla)

linear machine learning-based survival analysis methods. In sight of the models performance, we move forward to its practical implementation, proposing a tool to manage the decision of what patients should be contacted at a given date using the predictions of the trained models. This tool will manage the whole cycle, continuously training on new data, performing daily predictions, and presenting the results to the health-care decision-maker for their assessment.

Keywords: Sick leave prediction, Survival analysis, Machine learning, Decision support systems, AI in medicine

1. Introduction

Paid sick leave is a worker's right in many countries. According to the World Health Organization (WHO) Scheil-Adlung & Sandner (2010), it helps to reduce the cost of National Public Health systems by preventing workers from not seeking medical assistance, which could aggravate their diagnoses or spread disease. Furthermore, it also serves as a social and economic stabilizer in times of crisis.

Global and national health organizations have defined a set of guidelines about sick leaves expected duration Scheil-Adlung & Sandner (2010). These guidelines help physicians to determine when to intervene, while rationalizing healthcare costs. However, it is now clear that the effects and consequences of an illness are not the same for all individuals. Factors like age, gender, or work position may significantly influence the development of diseases. Hence, the duration of sick leaves is not a one size fits all task. For many years, data about the duration and development of sick leaves have been collected. It is now possible to analyse this information at a large scale, so that we may obtain decision models for more rationalized processes with significant well-being and economic returns.

There is a vast literature on health policy and epidemiology about the course and duration of particulars or groups of diagnoses. Some of these studies focus on the effects of interventions during the leave period that can benefit the

patient, by defining a more accurate return date and preventing relapses. For instance, in Edwards et al. (2019) the relation among different sick leave policies for seasonal influenza and their economic impact is studied using dynamic systems. It is shown how sooner reincorporation to work or longer leaves may be beneficial. Previous works have also studied sick leave duration as a timely event. This means that the model includes not only if the event is going to occur but also when it is going to happen. In this setting, survival analysis is the method of choice for modelling this sort of problem. For instance, in Vemer et al. (2013) survival analysis is performed to study the time needed for workers suffering from depression to return to their jobs. In Spierdijk et al. (2009) these same methods are used to study the causes of sick leave duration of Dutch self-employed workers, as well as the effect of different interventions.

The aim of this paper is to model the duration of sick leaves using survival analysis based on a large dataset of sick leaves from workers in the Public Health System (Spanish Social Security). Preliminary experiments with the data allow us to conclude that models trained using classical survival analysis methods (Kaplan-Meier, Cox Proportional Hazard Kleinbaum & Klein (2010)) are not completely appropriate due to the possible existence of non-linear relationships among the data. Moreover, the data presents complex characteristics with a mixture of continuous, binary and categorical attributes, which are difficult to deal with for these models. This leads us to use machine learning survival analysis methods based on random forest and neural networks. Results indicate that these are indeed better predictive models for the task.

Our contribution is intended to be used in practice. Currently, the data available for our study is being used by a team of nurses to prioritize every day which workers should be contacted for an intervention (e.g., regular check-ups, refer to specialist). Since their capacity is limited, and the number of candidates for an intervention ranges in the tens of thousands, a prioritization of the intervention is paramount. Hence, the final goal of this work is to develop a set of models that will be used as a decision support system to recommend when an intervention is needed, given the probable duration of the leave.

Our assumptions for developing this system are two fold. First, that the historical observations of leave duration allow the prediction of future leave behavior and that the available socio-economic and labour attributes are a surrogate for more complex medical information. Second, that these characteristics are enough for obtaining explanations for the leaves duration for each diagnose and that these can be used by the end users (nurses) as actionable criteria for prioritizing interventions.

The contributions of the paper are as follows:

- The study of a large dataset of leaves corresponding to many different diagnoses using non-linear survival analysis methods
- The proposal of a flexible system that implements a decision support system based on survival analysis models for improving how sick leaves are managed at large scale

The outline of the paper is as follows: Section 2 explains the characteristics of the dataset used in the experiments and the different preprocessing steps applied to obtain the data before feeding the models. Section 3 gives a brief explanation of survival analysis and the characteristics of the machine learning methods used for the modelling. Section 4 explains the methodology employed for the experimentation, training, and validation for the models. Section 5 outlines the results and insights obtained from the experiments. Section 6 describes the current procedure for determining the interventions, the architecture of the system that implements the decision support system developed for aiding the decision-making process and how it will be used. Section 7 outlines some future directions and additional extensions to this work. Section 8 summarizes the main results of the paper.

2. The Dataset

The Spanish Social Security is in charge of granting sick leaves in the public health system. In the follow-up of these leaves, a combination of public and

private actors may intervene, as some leaves are derived to private and public health companies to reduce the load and cost of the public healthcare system. Within this ecosystem, Asepeyo is a partner of the Spanish Social Security, acting as a health provider and insurance manager of the economic benefits associated to the sick leave. It has 142 owned healthcare centres and 4 hospitals, with over 3,000 employees and handles health data of roughly 5 million patients-workers.

This project arises from Asepeyo's need to properly prioritize patients on sick leave. To tackle this issue, Asepeyo gathered a dataset composed by roughly 1.8 million sick leaves, taking place during a period of almost six years (January 2014, September 2019). To avoid any ethical issues with the use of the data has been anonymized. Those sick leaves correspond to more than 12,000 different diagnoses, but only the 129 most frequent ones are considered in this work, to maximize intervention relevance and practical impact. The targeted diagnoses can be categorized into eight general groups: fractures, cervical, lumbar, shoulder, knee, osteoarticular, tendinose and psychiatric. These diagnoses account for over half a million sick leaves in the dataset.

Each sick leave in the dataset contains a large amount of metadata. This includes information about the leave (diagnosis, start date, final date, last follow-up date...), personal data (date of birth, gender, marital status...) and labour data (job classification, job status...). The first preprocessing step targeted the missing values in the data. Attributes having too many were dropped, and for the rest, missing values in categorical variables were replaced by a dummy value. Furthermore, outliers in the length of the leaves were dropped, discarding the samples with values over the 99 percentile. Some features were added based on existing attributes (*e.g.*, day of the week, week of the year), as a preliminary statistical study of the dataset showed its relevance. For instance, the length of the leaves presents weekly periodicity and a large number of short leaves have as their more frequent duration 5 or 7 days, indicating probably Monday-Friday or Monday-Monday leaves (see Figure 1). Other features were also added, such as counts of recurrent leaves for a worker (with the same or different diagnosis), the

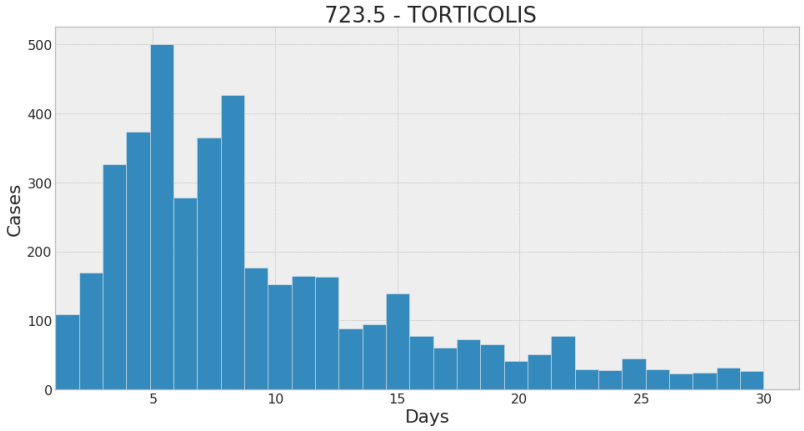

Figure 1: Distribution of leave duration for Torticollis cases. An increase of leaves termination is associated with weekly periods and short leaves ends increase after five or seven days.

duration of the last leave (with the same or different diagnosis), and the sum of the duration of the previous leaves. After preprocessing, a total of 27 variables were chosen for building the models. Some of them were nominal variables and were coded using one-hot encoding.

In survival models, the goal is to predict the time until the occurrence of an event. In our case, the end of a sick leave. However, some data samples may not have reached the event yet, and not using these for training could result in an underestimation of the duration. These examples are called *right censored* in survival analysis terminology. This means that the data needs an additional indicator attribute that represents if the event has occurred already or not. There also exists the concept of *left censored* data, those examples included before the beginning of the process. In our data, that does not happen because there are no leaves that start before the starting date of the dataset.

The final dataset includes a variety of diagnoses grouped in the eight mentioned categories. Figure 2 plots the distribution of the number of leaves (in logarithmic scale) and the maximum duration of the leaves for each category. One issue to notice is the differences in their distribution, indicating that a

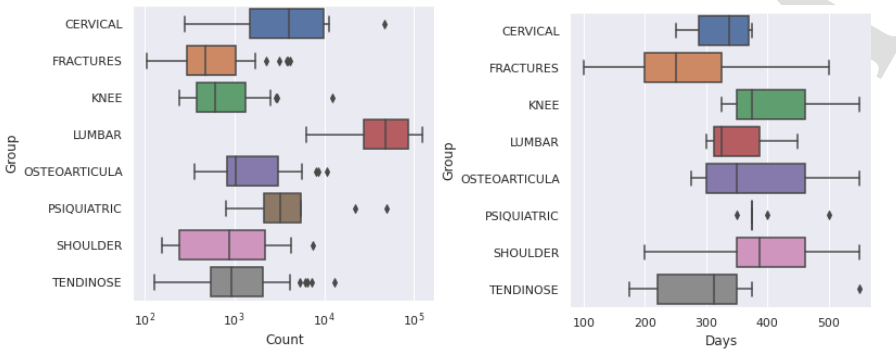

Figure 2: Distribution of number of leaves (left, in log scale) and their duration (right) for the eight groups of diagnosis considered.

global model or even a category model is not feasible given these wide differences among the diagnoses characteristics. Instead, we will train a specific model for each of the 129 diagnosis.

3. Machine Learning methods for Survival Analysis

Survival analysis Kleinbaum & Klein (2010) is a field of statistics for the analysis of the expected time until one or more events happen. In the context of our problem, the event is the end of the sick leave, as the worker is considered apt for work by a clinician. Survival analysis performs the modelling using different functions like the survival function, the hazard function or the risk score. In our case, we use the survival function $S(t)$

$$S(t) = P(T > t) \tag{1}$$

where t is a specific time and T is a random variable denoting the time of the event. This function represents the probability that the event of interest has not occurred after some time t has passed. This value is intuitive to interpret by the final decision maker. For a specific sick leave, this probability indicates the chance that the worker is still sick at a given time.

There are different approaches to model the survival function. It can be estimated parametrically, assuming a specific model for the distribution of time (*e.g.*, exponential, Weibull, log-normal...), it can be estimated non-parametrically (*e.g.*, Kaplan-Meier model) or semi-parametrically (*e.g.*, Cox Proportional Hazard, Multi task logistic regression, Survival trees). However, the usual models do not take into account non-linear interactions among the attributes. Considering the many relations between variables that can be observed in our data, we focus on non-linear methods. These methods substitute linear and logistic regression by non-linear machine learning methods. In this context, we are going to build models using three different approaches: Conditional Survival Forest, Neural Multitask Logistic Regression and Non-Linear Cox Proportional Hazard.

3.1. Conditional Survival Forest

The assumption of survival trees is that a unique survival function is not adequate for modelling the data. Instead, it is better to split the model into several functions according to their characteristics. Modelling is done using ensembles based on random forest. There are several implementations available for this model. We have chosen the Conditional Survival Forest (CSF) Wright et al. (2017) among them because of its competitive performance.

CSF combines a set of trees built using the original decision trees algorithm. Each decision evaluates all variables looking for the best split, in this case, computing a statistical independence test among the regressor variable and the response variable. From the variables that have a significance higher than a threshold for the test, the best one is picked to make the decision.

This model has the advantage of being able to compute the relevance of the attributes of the model, helping to the interpretability of the results.

3.2. Neural Multitask Logistic Regression

The Multi-Task Logistic Regression Yu et al. (2011) model can be seen as a series of logistic regression models built on different time intervals to estimate the probability that the event of interest happened within each interval.

As logistic regression is a linear transformation, the Neural Multitask Logistic Regression (NMTLR) model was proposed in Fotso (2018) substituting it for a non-linear neural network.

3.3. Non Linear Cox Proportional Hazard

The Cox Proportional Hazard model Cox (1972) assumes that a baseline survival function can be computed. In this model, the different individuals have a function that is proportional to the baseline, where the actual proportion can be computed as a linear combination of its characteristics. The Non-Linear Cox Proportional Hazard model (NLCPH) Katzman et al. (2018) substitutes the linear model by a neural network to extend the capabilities of the Cox model.

4. Experiments

To conduct the experiments, samples from each individual diagnosis were randomly split into training and validation, with a proportion 80%/20%. One special circumstance in survival analysis is that the data has to take into account examples where the event has not yet occurred. In our dataset, examples with ongoing leaves are randomly distributed both in the training and the validation sets. All the experiments were performed using the `pysurvival` python library Fotso et al. (2019) which implements all the methods explained in the previous section.

During training, a grid search was used for the exploration of the hyper-parameters. For CSF these include the size of the ensemble, the maximum number of features for each tree, the minimum number of examples in the leaves, and the criteria used for the split of a node. For NMTLR and NLCPH they included the design and training details of the neural network. We considered one or two layers with different number of neurons, different activation functions for the neurons, the parameters for the optimizer (Adam in our case), the initial learning rate, the weight for the L2 regularization and the maximum number of epochs. Additionally, for the NMTLR we also explored the number of bins

used to divide the timeline. For each combination of parameters we performed a 10-fold cross validation to assess the accuracy of the models.

All the experiments were run on Barcelona Supercomputing Centre's Marenostrum 4 supercomputer. To measure the performance in a similar situation as in the target deployment setting, a limited number of computing nodes were used, and the training of the neural networks was performed without accelerators (*i.e.*, no GPU). We assumed that, in production, only commodity cloud computing resources are going to be available for the retraining of the models. In summary, models for 129 diagnoses were trained, exploring around a hundred hyper-parameter combinations for each one of the three algorithms. The goal was to have an estimation of the average performance that could be achieved for each diagnosis and each algorithm.

The code for performing these experiments is available in `https://github.com/bejar/OmenCode`.

5. Results

This section presents the metrics used, the results obtained by the trained models, as well as a preliminary interpretation of their performance.

5.1. Performance metrics

There are two main performance measures for survival analysis, each capturing a slightly different metric. These are the *Concordance Index* and the *Integrated Brier Score*. The *Concordance Index* (CI) is a generalization of the area under the ROC curve (AUC) that takes into account that there are censored data. It assesses the ability of the model to rank survival times based on the survival function. The maximum value is 1 (perfect prediction) and the minimum is 0.5 (random prediction). On the other hand, the *Brier Score* evaluates the accuracy of a predicted survival function at a given time t. It represents the average squared distance between the observed survival status and the predicted survival probability, and it is always a number between 0 and 1,

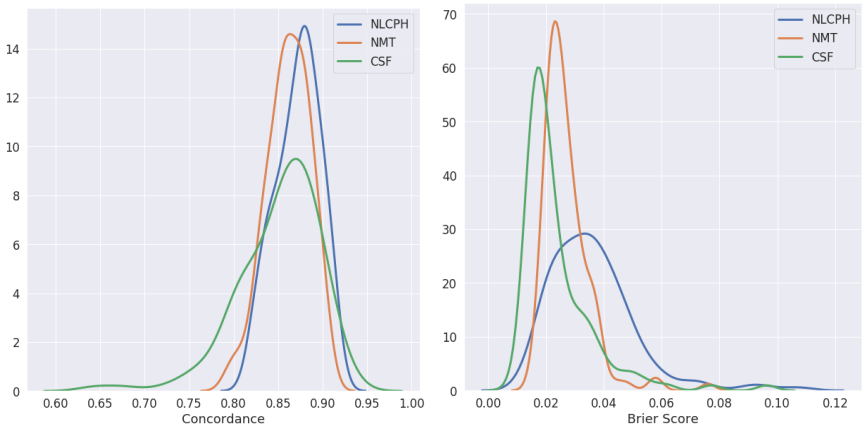

Figure 3: Distribution of Concordance index and Integrated Brier Score of the best model for the 129 diagnosis separated by algorithm. For Concordance, closer to the right is better. For Brier, closer to the left is better.

with 0 the best possible value. The *Integrated Brier Score* (IBS) computes the mean of the score and gives the overall performance of the model. According to literature Kleinbaum & Klein (2010), a value of 0.7 for the CI indicates a good model, with 0.8 indicating a strong model. The Brier score should not be over 0.25 for all the time interval for a reasonable model.

5.2. General results

The results obtained for all diagnoses are plotted as a distribution in Figure 3. By looking at the plot on the left, results for the CI metric are mostly above 0.7, with the majority being above 0.8. This indicates that most of the obtained models are reliable predictors. The algorithm NLCPH is the one with the best models according to Concordance (its distribution has the higher mean, close to 0.88), followed by NLMTLR. CSF has a few models that perform poorly (between 0.6 and 0.8), but has a reasonable performance in general.

For the plot on the right of Figure 3, based on the IBS, we observe the opposite behaviour. The CSF is the closest one to optimal performance (left), while the NLCPH has a larger error (although it is always below 0.12). Remarkably,

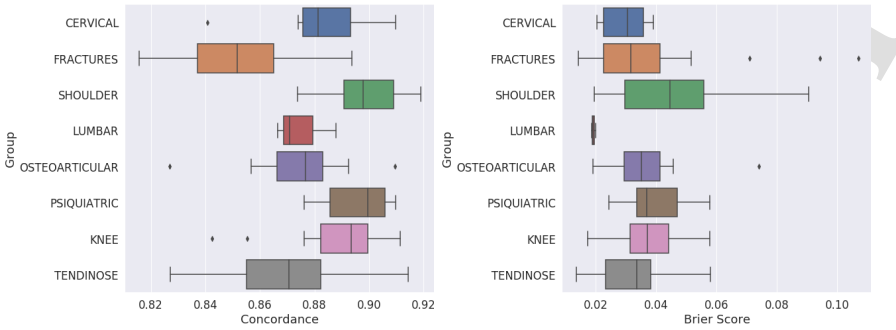

Figure 4: Distribution of Concordance index and Integrated Brier score for NLCPH models by groups of diagnoses

the vast majority of models achieve performances which can be considered as strongly reliable.

To understand the disparity between the CI and IBS metrics in terms of model scoring, let us discuss their interpretation. According to CI, NLCPH is the best at ranking the sick leaves by their expected duration, providing the most reliable ordering. On the other hand, according to IBS, CSF predicts better the overall duration of the leaves, with the least deviation from the predicted date. NLMTLR can be seen as a compromise between both models.

To analyse the performance of models across diagnosis, in Figure 4 we plot the results of both scores group-wise for NLCPH. Results indicate a high variability among and inside groups, with large differences between both scores. Some groups of diagnosis are more difficult to order, but easier to estimate the date (*e.g.*, cervical or lumbar), while others are easy to order but harder to estimate (*e.g.*, shoulder or psychiatric). These differences may be caused by the duration of the leaves, since shorter leaves may be harder to order, while for long leaves may be harder to estimate the exact ending date. Another variable that may be of relevance is the number of cases for each diagnosis. In the following section, we will analyse the influence of these two factors in the data, to test their impact on performance variance and to better understand our models.

5.3. Effects of dataset size and leave duration

To assess the impact of sample size and leave duration on the models, in Figure 5 we plot these two variables together with each of the two scores used. Results indicate that diagnoses with more samples and longer leaves result in models with higher CI values (*i.e.*, better ranking of leaves), with a significant positive effect according to a robust linear regression using Huber loss (see right plots of Figure 5). On the one hand, this means that, as expected, the larger the size of the dataset, the better results can be obtained, which provides good expectations as of how will these models evolve through time as more training data is made available. On the other, the effect of the maximum length of a leave could be explained by a better distribution of the duration of the cases, as well as a larger margin for ordering samples.

The effect on the IBS is weaker (see left plots of Figure 5). A slight positive effect can be observed with the maximum leave length (performing worse as the leaves get longer, bottom left plot of Figure 5). Similarly, the magnitude of the number of samples has a slight negative effect (performing better as more data is available, top left plot of Figure 5). This later case seems to be driven by those diagnoses with the least cases (see the left side of the top plot of Figure 5), which reinforces the idea that in a broader sense, the relation is very weak.

5.4. Attribute relevance and interpretability

One of the goals of the decision support system is to obtain some insights regarding the best criteria in the decision of prioritizing some leaves over others. For studying this issue, we collected the relevance importance ranking of the attributes computed by the best CSF models for each diagnosis. This relevance is computed in the same fashion as for Random Forest as this is the base for the CSF model.

We explored the patterns in the relevance rankings of the attributes used by the models. For that, we estimated the distribution of the position of the attributes in these rankings. Figure 6 presents a matrix of the number of times an

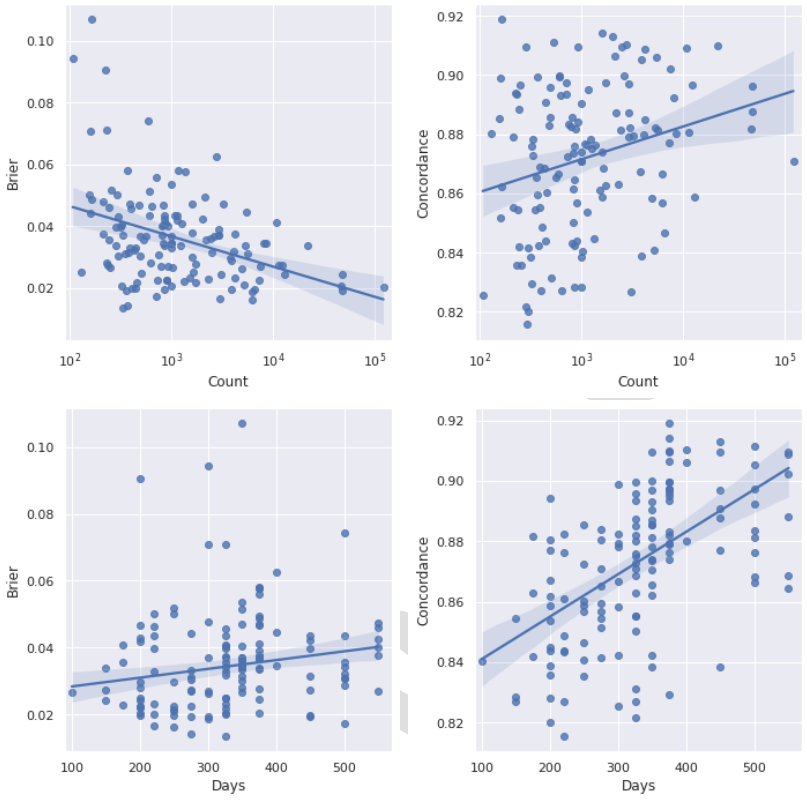

Figure 5: Distribution of Concordance index and Integrated Brier score with respect of the magnitude (log10) of the number of cases (Count) and the maximum length (Days) of a diagnosis for NLCPH models. For Brier, lower is better. For Concordance, higher is better.

attribute appears in a position in the relevance rankings. This count is normalized by the maximum for each position in the ranking (column wise). Results indicate that some attributes are more frequent in a specific position than others, presenting on average more relevance for the models (top attributes). It can also be seen that some attributes are irrelevant, not appearing in most of the models (bottom attributes).

Except for one of the attributes, there is a large variation on the ranking

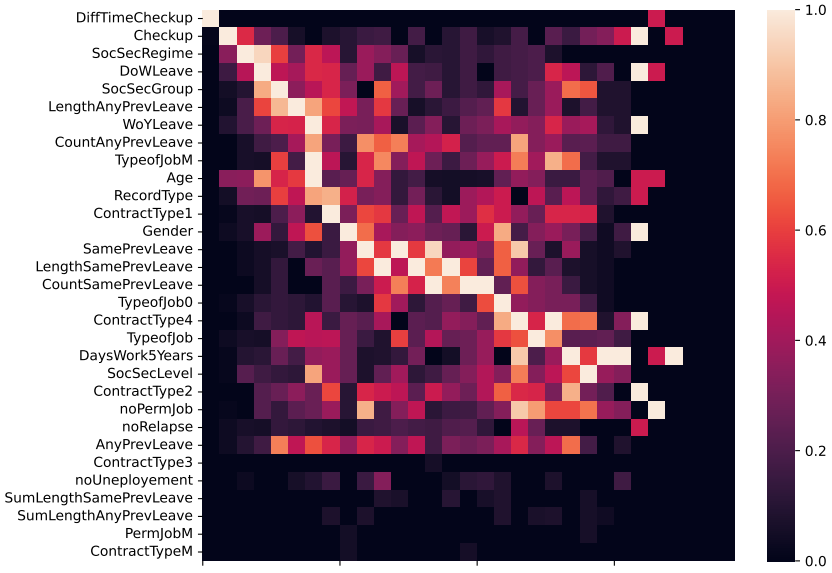

Figure 6: Distribution of the positions in the attribute relevance ranking computed by the CSF model.

for each attribute. This means that each diagnosis can usually be explained by a different combination of attributes. In other words, some attributes can be highly relevant for some diagnoses and almost irrelevant for others. This relevance can be used for explaining the global predictions of the models and presented by the decision support system as additional information about what attributes are better criterion for prioritizing leaves, so it can be considered by the user. For instance, for the diagnosis *shoulder dislocation* the age and the gender are the second and third most relevant attributes after the time from the last check up, or for the diagnosis *shoulder tendinitis* to have any previous leaves is the second most relevant attribute. Having that information will help to interpret why some leaves are prioritized over others.

To obtain a more detailed explanation for a particular leave prediction is more complex given that SCF is based on Random Forest that is a black box model. Model agnostic methods can be used in this case, like, for instance,

the LIME method Ribeiro et al. (2016). This method uses a linear model as surrogate that obtains a linear model that locally explains the prediction. As in the current application of the decision support system, it is more important to obtain general criteria for ordering the leaves, this is not a crucial feature at this time.

5.5. Algorithm selection

Overall, the three models considered provide reliable predictions for most of the diagnoses. For the first deployment in production we have chosen to use the models computed by NLCPH and CSF. The NLCPH models are closer to the ranking among the sick leaves, which is the most relevant functionality according to the end users of the decision support system. We will also integrate the CSF models, since this allows us to assess the influence of the different variables in the predictions. Such information can provide useful insights to the decision maker to refine their actions, and help to understand the reasons behind the difference among patients with the same diagnosis.

6. Tool Deployment

Every sick leave has a specific duration determined by a doctor. The doctor decides this duration based on the severity and the type of illness. Nonetheless, the time expected for a leave might deviate from the time it actually takes to heal. Calling each patient at the right time brings about considerable well-being and economic effects, as it might mean less stress for the patients while leaves are finished at their appropriate time according to medical standards.

Currently at Asepeyo, a team of experienced nurses skim through the complete list of patients that are on leave every morning. According to a set of criteria, they sort the list in order to decide which patients to call first. These criteria combine the team experience with rules that are based on specific factors such as the type of diagnosis, the total time passed since the last medical action, the time left before the next action, whether or not a call has already

been made, the leave history, and so on. Specific details about the complete set of rules are out of the scope of this paper but, in the general case, the nurses look for patients with a nearby expected date of recovery, where the meaning of *nearby* in this context is heavily influenced by the type of diagnosis.

The distribution of the workload among nurses is not defined by the current system and it is dynamically adjusted depending on the number of cases and their characteristics. There is no split of diagnoses among the nurses, considering they have a similar level of expertise in all cases, although simpler cases can be assigned to auxiliary nurses. At the beginning of the day, each nurse first solves the patients pending from the day before. Once all pending cases have been completed, the filtering of new cases starts. Each case is evaluated sequentially to decide if the patient is going to be contacted. Once contacted, if the patient is recovered, they go to the doctor to end the leave. Otherwise, the leave continues as scheduled by the doctor.

6.1. The decision support system

The models evaluated in the previous section will provide a decision support system for these nurses, to help them refine and improve their decision-making process. Models will be deployed for inference (*e.g.*, prediction) and integrated with the Asepeyo systems so that they can become an additional tool for the nurses to be used in their day-to-day work along with the interface they are currently using.

Models will provide a priority list of leaves split by diagnosis. Diagnoses that are simpler to evaluate can be routed to auxiliary nurses that also have access to an estimate of the probability of the leave. The rest of the diagnoses can be scheduled for calls by experienced nurses that can consider the information the system will provide. Specifically, the order of cases, their probability and the more relevant attributes considered by the model for the diagnosis. Other considerations that are already taken into account can also be included in the decision, such as prioritization according to the gravity of the diagnosis.

Meanwhile, the models will be retrained periodically (e.g. once a week,

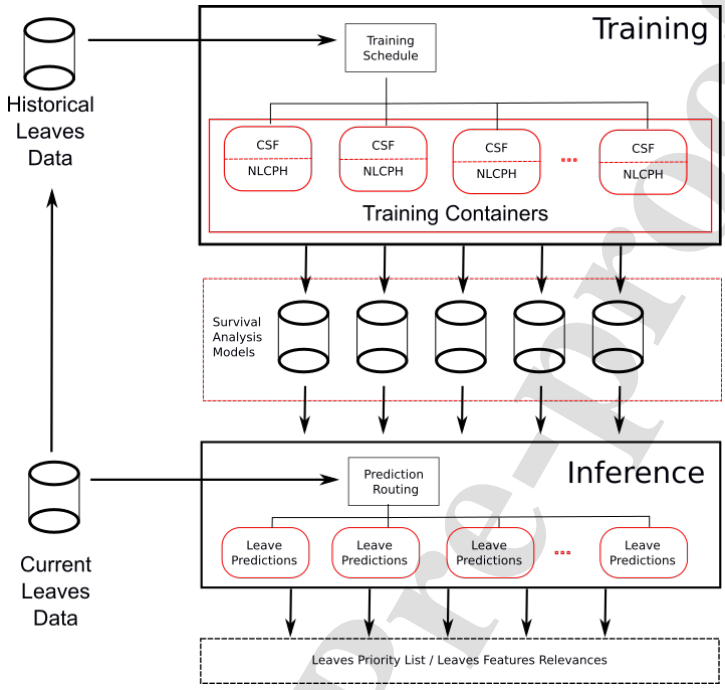

Figure 7: Organization of the Decision Support System with the training and inference subsystems for generating priority lists of patients

during the weekend...) through the addition of the training data that is produced and stored every day. Therefore, updated and improved predictions are constantly made available.

The architecture of the decision support system is based on two subsystems as depicted on Figure 7:

- The **training subsystem** will be in charge of generating and updating the survival analysis models. It will schedule the training of the models from a historical sick leaves database considering the amount of new examples included since the last models for each diagnosis was computed. The schedule will tune the parameters, if needed, when the quality of the new models varies significantly from the current model. To make the decision, the historical values for quality measures (IBS, CI) of the models will be

stored, and a statistical significance test will be used for performing the decision.

For the parameter tuning, a Sequential Model-based Algorithm Configuration (SMAC) (Hutter et al. (2011)) optimization algorithm will be used. This algorithm estimates the surface corresponding to the quality measures of the models in the hyper-parameter space allowing a better exploration. Specifically, regression random forest is used for this estimation. This prediction surface is stored in order to bootstrap future hyper-parameter searches when the quality of the new model, trained with the current hyper-parameters, differs from the latest model.

After training the models for both prediction algorithms, the relevance ranking from the CSF will be stored for the inference subsystem.

- The **inference subsystem** will be in charge of computing the survival function for the pending leaves. The leaves for each diagnostic will be sorted according to their probability at the current date. A probability threshold will be used to cut the list of leaves to the ones with a higher chance to be close to their end. For each diagnosis, the ranking of the attributes' relevance of the CSF models will be included for the nurse to interpret.

The prediction of the survival function for the pending sick leaves is not computationally demanding, and thousands of leaves can be predicted in a matter of seconds using commodity hardware. This prediction can even be parallelized if needed, given that each diagnosis is independent of the rest. This means that the ordering for the patients can be ready at the beginning of each day.

The main burden of computational cost corresponds to the training subsystem. For the two chosen algorithms, the fitting of the models has different constraints. NLCPH uses neural networks so its training can not be parallelized when using only the CPU. However, it can be trained using GPUs if needed, drastically reducing its cost. CSF does not gain performance from using GPUs, but it can run in parallel, gaining from using several cores during training.

The current size of the training data varies from diagnosis to diagnosis. The one with the larger number of examples is *back pain* with more than 100,000 cases. The next three diagnoses have around 50,000 cases each and, after the 20^{th} diagnosis, this number is lower than 10,000 cases.

Measuring the wall clock time during the experiments for the hyper-parameter tuning, the training time for NLCPH for the diagnosis with the larger number of examples using only CPU takes around 2 minutes per configuration including a 10-fold cross validation using a single thread on a Intel Xeon 8160 2.1 GHz processor. This time could be reduced, if necessary, at least by an order of magnitude using a modern GPU. For CSF, the training for that diagnosis takes around 30 seconds not using multi-core training, and also including 10-fold cross-validation. These quantities can vary depending on the complexity of the models but in all cases it will approximate these values.

The computational cost of the solution makes it feasible to schedule and distribute the training of the survival analysis models for the diagnoses at large scale and in a flexible way. In most cases, the models will be updated when the number of instances for a diagnosis increases significantly using the same parameters as the last model. When a hyper-parameter tuning has to be performed, the SMAC optimization algorithm can reduce the number of models that have to be tested and this exploration can be interleaved with the normal update of the models. The independence of the training of the diagnoses also allows for increasing or decreasing the computational resources used. Most of the models can be trained in-house using commodity computers and cloud services can be added if necessary.

In order to evaluate the effects of integrating the models predictions into the current interface used by the nurses, two approaches will be followed. During the first months, the nurses will have access to the predictions but will not be encouraged to act upon it. Periodically, these predictions will be compared to the actual duration of the leaves in order to detect any possible deviations from the original dataset used for the design and implementation of the models. This is relevant due to the sensitivity of the data with respect to socio-economic

variations, e.g. economy crises. In a second phase, there will be a period of A/B testing in which we compare the results of the calls made by a control group, formed by nurses that apply the same method they have always followed, and another group, formed by nurses that factor the models predictions into their sorting methodology.

7. Future work

There are different extensions that can be performed in this work related to the extent of the analysis of the leaves data that the system provides and the type of target users for the system. One obvious extension is to include a larger number of diagnoses. This is only limited by the amount of historical data that is available to the system. In the same way the retraining of the models is performed as the precision changes, new diagnoses could be included as the performance of the models reach an acceptable level, extending the capacity of the system for prioritizing new diagnoses incrementally.

For the current system, we are only using limited information about each leave that does not include medical information for the patients apart from the diagnosis. We have assumed that the available attributes are an adequate surrogate for that information when predicting the leave duration and that these attributes are enough for the interpretation of the predictions for the nurses. Although, adding information related to the diagnosis, like different tests and clinical analysis or comorbidites information, would allow generating predictions that can be used by the clinicians in order to schedule further visits or assessing the evolution of the diagnosis according to the predictions of the survival model. To obtain an explanation of the prediction in terms of that information can also be helpful in order to take clinical decisions. This is an extension that we will pursue as this information is available for the patients included in the dataset.

8. Conclusions

In this work we have designed, trained and evaluated reliable AI models for predicting sick leaves duration using historical data. To this end, we successfully apply different methods for survival analysis and obtain models using three different algorithms that score high on two different metrics. The quality of the results makes these models apt for deployment as a tool for decision-making. They can estimate the duration of a particular leave, rank patients by the likelihood of recovery, and extract the importance of each feature.

These models have been encapsulated within a software tool that can be deployed easily on site. This tool allows for the periodical retraining of models, and provides inferences in real time. The cost of the training is limited and it can be performed on commodity computing infrastructure. Training time could be improved if necessary using GPUs acceleration or scaling the computational resources in case of high demand. The cost of the inference is also cheap, so it can be applied at a large scale for presenting daily information useful for the decision-making process.

Our study shows that data volume is critical for model performance. Hence, the models will work better for the more common diagnoses. In our dataset, there are over 12,500 different diagnoses, and approximately 300 of those have at least 1,000 cases. This 2.4% of diagnoses contains 72% of all cases. As more data is collected daily, we can expect the performance on the less frequent diagnoses to improve.

Another relevant insight of this work is the difficulty of finding good general estimators. The information associated with each case is also limited, and the explanations that can be extracted are largely related to socio-economic characteristics and simple computations about the leave history of the worker. To improve the insightfulness of the models we consider the addition of medical history data and test records. This would allow the model to consider risk factors, like the ones caused by comorbidities.

This system is not with limitations that arise from the data that is used for

building the models and from uncertainty sources that affect the predictions. Specifically, the amount of leaves for diagnose has an impact on the quality of the prediction (see Section 5.3). This can be improved as more historical data is available, but for more rare diagnoses an acceptable amount of data could be difficult to reach. Also, there are some calendar effects that can be observed on short duration diagnoses (see Section 2), that introduce uncertainties on the predictions. This makes more difficult to prioritize leaves for these diagnoses. Finally, the criteria used for explaining the leaves is tailored to the nurses needs. A more complete medical information for the leaves would allow to extend the use to other interested users like for instance practising physicians.

Acknowledgements

This work has been funded by a collaboration between Asepeyo and BSC. We want to thank the following people from Asepeyo for their help and support that made this project possible. Xavier Calatrava Petisme, Alex Nogué Martinez, Enric Lleal Serra, Francisco Manuel Ventura Nofuentes, Oscar González Cherta, Francisco Sánchez Algarra, Eulàlia Borén and Vanessa Vegazo. We also want to thank to Nadia Tonello (BSC Data Manager) and Ulises Cortés (UPC/BSC) for their insightful comments.

References

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*, 187–202.

Edwards, C. H., Tomba, G. S., Kristiansen, I. S., White, R., & De Blasio, B. F. (2019). Evaluating costs and health consequences of sick leave strategies against pandemic and seasonal influenza in norway using a dynamic model. *BMJ open*, *9*, e027832.

Fotso, S. (2018). Deep neural networks for survival analysis based on a multi-task framework. *arXiv preprint arXiv:1801.05512*, .

Fotso, S. et al. (2019). PySurvival: Open source package for survival analysis modeling. https://www.pysurvival.io/.

Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2011). Sequential model-based optimization for general algorithm configuration. In C. A. C. Coello (Ed.), *Learning and Intelligent Optimization* (pp. 507–523). Berlin, Heidelberg: Springer Berlin Heidelberg.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, *18*, 24.

Kleinbaum, D. G., & Klein, M. (2010). *Survival analysis* volume 3. Springer.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016* (pp. 1135–1144).

Scheil-Adlung, X., & Sandner, L. (2010). The case for paid sick leave. *World health report*, .

Spierdijk, L., van Lomwel, G., & Peppelman, W. (2009). The determinants of sick leave durations of Dutch self-employed. *Journal of health economics*, *28*, 1185–1196.

Vemer, P., Bouwmans, C. A., Zijlstra-Vlasveld, M. C., van der Feltz-Cornelis, C. M., & Hakkaart-van Roijen, L. (2013). Let's get back to work: survival analysis on the return-to-work after depression. *Neuropsychiatric disease and treatment*, *9*, 1637.

Wright, M. N., Dankowski, T., & Ziegler, A. (2017). Unbiased split variable selection for random survival forests using maximally selected rank statistics. *Statistics in medicine*, *36*, 1272–1284.

Yu, C.-N., Greiner, R., Lin, H.-C., & Baracos, V. (2011). Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In *Advances in Neural Information Processing Systems* (pp. 1845–1853).

ORCID

- Javier Béjar, ORCID: 0000-0001-5281-3888
- Raquel Pérez, ORCID: 0000-0002-0041-8146
- Armand Vilalta, ORCID: 0000-0002-4871-1480
- Sergio Álvarez-Napagao, ORCID: 0000-0001-9946-9703
- Darío García-Gasulla, ORCID: 0000-0001-6732-5641

Research Highlights

- Modelling of Patient sick leaves at large scale from historical data
- Studying non-linear survival analysis methods for decision support systems
- Analysing attribute relevance extracted from decision trees survival models
- Building a Decision Support System for sick leaves management

Javier Béjar: Conceptualization, Methodology, Formal analysis, Software, Investigation, Writing - Original Draft, Writing - Review & Editing. **Raquel Pérez:** Software, Investigation, Data Curation, Validation, Writing - Review & Editing. **Armand Vilalta:** Software, Investigation, Data Curation, Validation, Writing - Review & Editing. **Sergio Álvarez-Napagao:** Software, Investigation, Validation, Writing - Review & Editing. **Dario Garcia-Gasulla:** Supervision, Funding acquisition, Writing - Review & Editing.

Declaration of interests

X The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: