RESEARCH ARTICLE

Advanced Online Survival Analysis Tool for Predictive Modelling in Clinical Data Science

Julio Montes-Torres[1,4]*, **José Luis Subirats**[1,2,4], **Nuria Ribelles**[3,4], **Daniel Urda**[1,4], **Leonardo Franco**[1,4], **Emilio Alba**[3,4], **José Manuel Jerez**[1,4]

1 Computer Science Department, Malaga University, Malaga, Spain, **2** Yachay Tech University, Urcuqui (Imbabura), Ecuador, **3** Virgen de la Victoria University Hospital, Malaga, Spain, **4** Malaga Biomedical Research Institute (IBIMA), Malaga, Spain

* julmontor@uma.es

Citation: Montes-Torres J, Subirats JL, Ribelles N, Urda D, Franco L, Alba E, et al. (2016) Advanced Online Survival Analysis Tool for Predictive Modelling in Clinical Data Science. PLoS ONE 11(8): e0161135. doi:10.1371/journal.pone.0161135

Editor: Gang Han, Texas A&M University, UNITED STATES

Received: February 25, 2016

Accepted: July 29, 2016

Published: August 17, 2016

Data Availability Statement: Data used to test the methods are available from the Comprehensive R Archive Network, and can be downloaded along with the rest of the package Survival from this URL: https://cran.r-project.org/web/packages/survival/index.html.

Funding: This work was supported by grants TIN2010-16556 from MICINN-SPAIN (Spanish Government) and P08-TIC-4026 (Andalusia Regional Government, Spain). All of them include FEDER funds (European Union). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

One of the prevailing applications of machine learning is the use of predictive modelling in clinical survival analysis. In this work, we present our view of the current situation of computer tools for survival analysis, stressing the need of transferring the latest results in the field of machine learning to biomedical researchers. We propose a web based software for survival analysis called OSA (Online Survival Analysis), which has been developed as an open access and user friendly option to obtain discrete time, predictive survival models at individual level using machine learning techniques, and to perform standard survival analysis. OSA employs an Artificial Neural Network (ANN) based method to produce the predictive survival models. Additionally, the software can easily generate survival and hazard curves with multiple options to personalise the plots, obtain contingency tables from the uploaded data to perform different tests, and fit a Cox regression model from a number of predictor variables. In the Materials and Methods section, we depict the general architecture of the application and introduce the mathematical background of each of the implemented methods. The study concludes with examples of use showing the results obtained with public datasets.

Introduction

Over the years, a number of works [1] [2] have shown the advantages of using the latest advances in machine learning to assist clinicians in determining the clinical outcome of patients after surgery. At present, there is a wide range of software solutions aimed to provide researchers with computer tools to perform classical statistical analysis and to implement those new machine learning techniques in their works. Many of them are commercial applications such as SPSS, Stata or SAS, which enjoy considerable popularity among researchers in the field of survival analysis [3] [4]. However, this kind of software has two significant disadvantages for many users: firstly, they are proprietary applications with major restrictions on its use; secondly, new methods published in scientific literature are not incorporated and updated in a short period of time. Clinicians may find a free alternative in open source statistical computing

Competing Interests: The authors have declared that no competing interests exist.

languages such as R (R Development Core Team, unpublished data), which is rapidly growing into the gold standard in clinical research and Bioinformatics. Still, R is an interpreted programming language with a hard to learn command line syntax, and full of options of no use for biomedical researchers who are only interested in survival analysis.

This situation has motivated the birth of many applications intended to ease the work of researchers in this field, such as CanSurv [5] and PODSE [6]. CanSurv is a Windows program developed to generate graphs representing standard survival models for population based data. On the other hand, PODSE is a MATLAB based tool for parameter optimisation in discrete time survival analysis. Some other applications are web based tools, like PROGgene [7], which includes more than 130 cancer datasets and is intended to help researches to find prognostic mRNA biomarkers performing usual survival analysis techniques. KMPlot [8] is another online software for biomarker assessment that also uses its own gene expression data. Finally, there are web applications which let the users analyse their own datasets, like OASIS [9], focused in classical survival models as well.

Nevertheless, none of the mentioned programs is transferring the results of machine learning research into a practical clinical application. Furthermore, they generally lack some features that we consider essential for scientific software that is not aimed at computer specialists: i) A wide range of tools updated with the latest results published in scientific literature, ii) advanced computational methods, like ANN-based algorithms, which are freed from the proportional and linearity constrains of classical models, having the potential to obtain a more accurate clinical outcome for individual patients, and iii) a clean and straightforward user interface which, at the same time, provides a great deal of options.

In this paper we propose the design and development of OSA (Online Survival Analysis), a system which fully meet all the requirements previously mentioned, providing an easy tool to obtain personalised survival curves from ANN-based models, and to carry out traditional survival analysis. The rest of the paper is organised as follow: Materials and Methods describes the architecture of the application, the web site structure and the mathematical foundations of the methods; Results and Discussion shows some results obtained with public datasets; Conclusions summarises our work and mentions future improvements. OSA is free and can be found at: http://www.icb.uma.es/Survival.

Materials and Methods

Architecture

OSA consists of the following elements:

1. An IIS web server running an ASP.NET MVC 4 application, which takes advantage of the latest ASP.NET, JavaScript and HTML5 technologies to provide an easy to use and feature rich user interface.
2. An SQL Server database which safely stores all the uploaded information using the 256 bit AES encryption to protect sensitive data such as the user's password.
3. A computational cluster with 27 nodes that executes the tasks by means of the R environment.

Fig 1 shows a simplified diagram of the application structure. The user can create projects which contain uploaded files with datasets and tasks to analyse the data. An IIS web server accesses the application database with the ADO.NET Entity Framework. It also connects to a web service that processes all the task execution requests. The web service selects a free node, constituted by a quad core x64 PC with 4 Gigabytes of main memory running Linux. Next, the

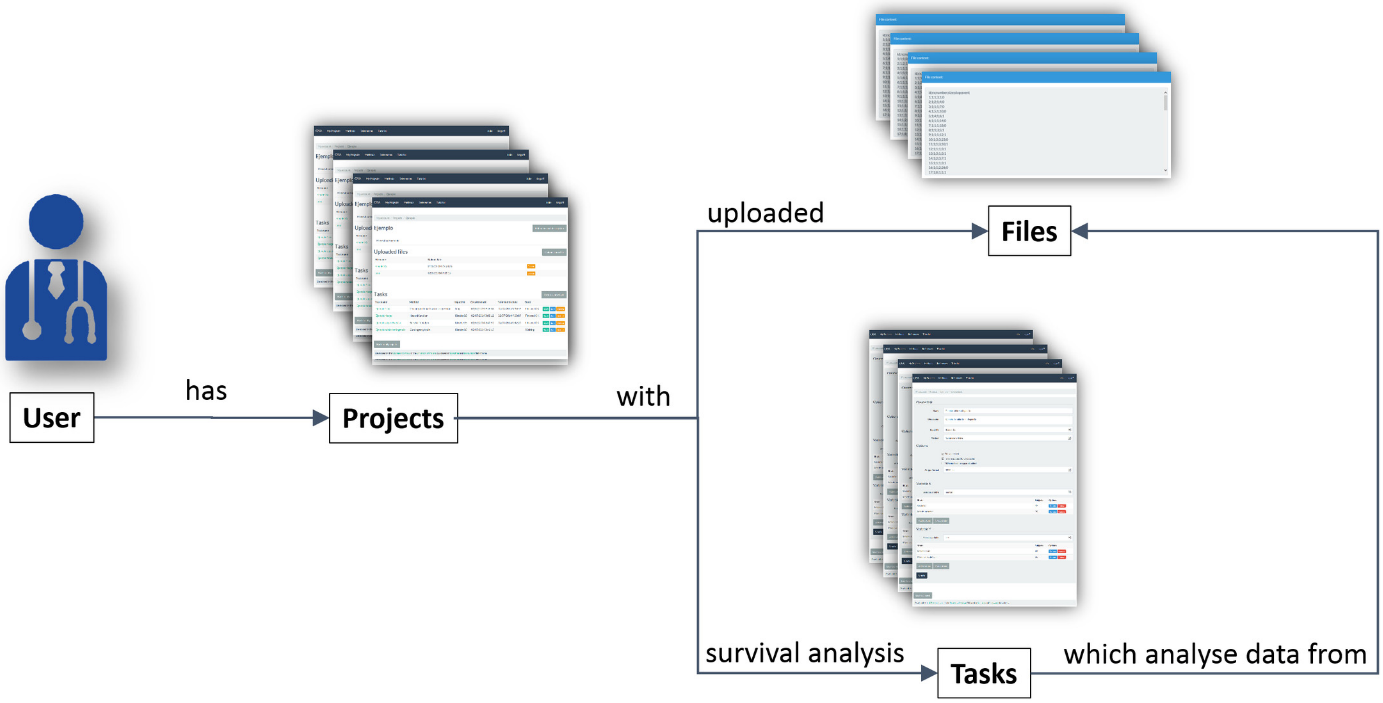


Fig 1. Simplified application structure diagram. This figure shows the main entities involved in the application workflow and their relationships.

doi:10.1371/journal.pone.0161135.g001

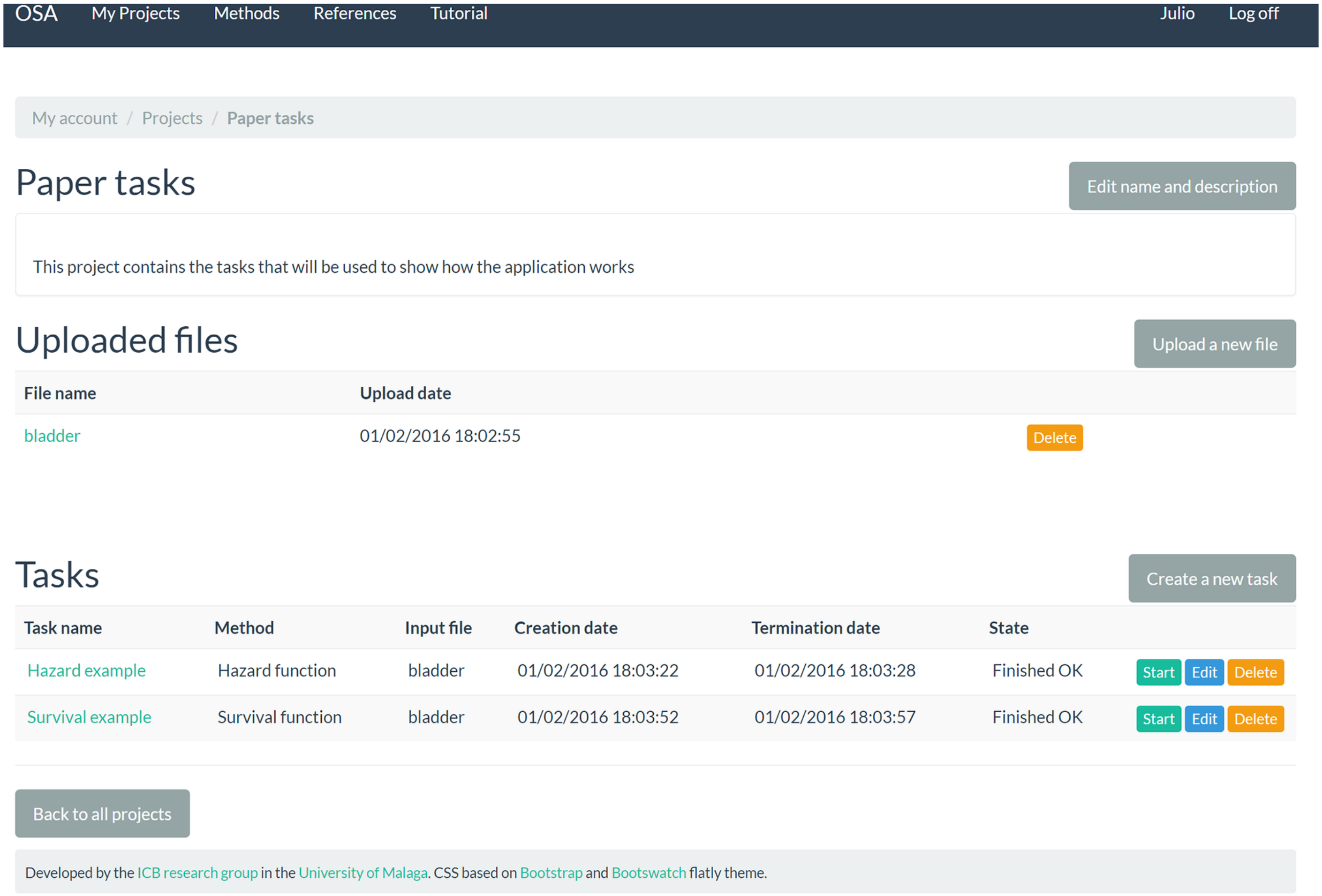
task is executed in that node using the R environment and the result is collected by the web service, which sends it to the main application.

Website structure

Every web page in the application features a top menu bar as in Fig 2, with the next five links:

1. My Projects. When logged on, it leads to the user's project list; otherwise the log on formulary will be shown to the user. Each project plays the role of a container where files with survival data and task are stored.
2. Methods. This link provides access to the method description web page, in which we include all the information regarding the mathematical background of the available statistical procedures.
3. References. Following this link, every visitor can refer to the supporting references of this work. These are presented in an appropriately formatted list.
4. Tutorial. It points to a concise explanation of the usage of this web application with practical examples.
5. Log on. This is the link to the log on formulary, where the user has to enter the name and password in order to access to the personal area. The formulary also includes a link to the register page for those who need to create a new account.

After signing up, an empty project list and a button to create new projects will appear. Once the project has been created, it is possible to include CSV files with survival data to perform statistical analysis. The file upload process requires the user to provide some basic information



Paper tasks

This project contains the tasks that will be used to show how the application works

Uploaded files

File name	Upload date
bladder	01/02/2016 18:02:55

Tasks

Task name	Method	Input file	Creation date	Termination date	State
Hazard example	Hazard function	bladder	01/02/2016 18:03:22	01/02/2016 18:03:28	Finished OK
Survival example	Survival function	bladder	01/02/2016 18:03:52	01/02/2016 18:03:57	Finished OK

Developed by the ICB research group in the University of Malaga. CSS based on Bootstrap and Bootswatch flatly theme.

Fig 2. Screenshot of a project page. Here, the user can upload datasets and create tasks.

doi:10.1371/journal.pone.0161135.g002

about the data, like the names of the columns with the follow-up time and the censorship status. When there is at least one CSV file in the project, the user can select one of the four traditional statistical methods or the ANN-based method provided by the application by creating a new task. Finally, after clicking on the start button to execute the task, the result can be obtained in the form of a table or a graph which can be downloaded in various formats (including PNG and PDF).

Methods

The methods provided by OSA to perform statistical analysis on the uploaded data can be classified in three groups: The basic, classical survival analysis methods (which includes the survival function and the hazard curve), the Cox proportional hazards regression model and the ANN-based predictive model.

Kaplan-Meier survival curve. The first basic method is the survival curve, which plots a Kaplan-Meier [10] estimate of the survival function. The method is based on the R packages `survival` (Therneau T, Lumley T, unpublished data) and `epitools` (Aragon T, unpublished data). Let t_i be an observed event time, where $i = 0, 1, 2, \ldots D$. Let m_i be the number of subjects with an observed event time t_i, and n_i the number of subjects at risk before t_i. Then, the Kaplan-Meier estimate, $\hat{S}(t)$, is given by

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{t_i \le t} \left(1 - \frac{m_i}{n_i}\right) & \text{otherwise} \end{cases} \tag{1}$$

The survival curve method includes a number of options, such as showing censored data, plotting the confidence interval at 95 percent, generating the plot using black and white graphics, showing the median follow-up time for one curve or including a table with the number of patients at risk. The user can also compare different survival curves in the same plot to establish whether there is a statistically significant relationship between them. The available tests to perform the comparison are Log-Rank [11], Peto-Peto [12] and Tarone-Ware [13]. Assuming the existence of two curves, X and Y, the comparison estimate would be

$$\frac{\left[\sum_{i=1}^{p} w_i (m_{Xi} - e_{Xi})\right]^2}{\sum_{i=1}^{p} w_i Var(m_{Xi} - e_{Xi})} \approx \chi^2, \tag{2}$$

where p is the total number of failure times, m_{Xi} is the number of observed events at a time t_i for group X, e_{Xi} is the number of expected events at t_i and w_i is the weight employed by the test at t_i. The number of expected events, e_{Xi}, is computed as follows

$$e_{Xi} = \frac{n_{Xi} m_i}{n_i}, \tag{3}$$

while the variance in Eq 2 is given by

$$Var(m_{Xi} - e_{Xi}) = \sum_{i=1}^{p} \frac{n_{Xi} n_{Yi} (m_{Xi} + m_{Yi})(n_{Xi} + n_{Yi} - m_{Xi} - m_{Yi})}{(n_{Xi} + n_{Yi})^2 (n_{Xi} + n_{Yi} - 1)}, \tag{4}$$

where n_{Xi} and n_{Yi} are the number of subjects at risk before t_i for group X and Y, respectively, and m_{Yi} is the number of subjects in group Y with an observed event time t_i. It can be proven that all the calculations are equivalent exchanging X for Y. The tests can be generalised to compare more than two groups, in which case the comparison estimate is approximately χ^2 distributed with $k - 1$ degrees of freedom, where k is the number of groups. For each curve, the result shows the median survival time, the number of events that have taken place and the total number of patients. It also prints the computed P value for the test, followed by the odds ratio when comparing exactly two curves. The default position of every label and legend can be selected within the plot area.

Hazard function estimation. The second basic method is the hazard function, which is based on the R packages `muhaz` (Hess K, Gentleman R, unpublished data) and `survival` (Therneau T, Lumley T, unpublished data). It computes a kernel smoothed hazard function from right censored data using a fixed bandwidth kernel smoothed estimator [14] [15]:

$$\hat{h}(t) = \frac{1}{b} \sum_{i=1}^{p} K\left(\frac{t - t_i}{b}\right) \Delta\tilde{H}(t_i), \tag{5}$$

where b is the bandwidth distance of t, $K()$ is the Epanechnikov kernel function

$$K(x) = \frac{3}{4}(1 - x^2), -1 \le x \le 1, \tag{6}$$

which is appropriately replaced by the corresponding asymmetric kernels of Gasser and Müller [16] for $t < b$ and for $t_D - b \le t \le t_D$, and $\tilde{H}(t_i)$ is the Nelson-Aalen estimator [17] [18]. The user can set the bandwidth of the kernel as well as other parameters, like showing the histogram of the function or the confidence interval at 95 percent. There are options for generating black and white results and for including a table of patients like in the previously discussed method. Another common characteristic of the first two methods is the possibility of doing visual comparison between various curves. Thus, two or more hazard functions can be shown in the same plot, including for each one the maximum hazard ratio and the time it has been reached. The user can also change the original position of every information printed over the plot area.

Cox proportional hazards model. The fourth available method, based on the R packages `survival` and `MASS` [19], is the Cox's proportional-hazards regression [20]. It uses the variables the user selects from the uploaded data to fit a proportional-hazards regression model of the form

$$h(t, X) = h_0(t)e^{\sum_{i=1}^{q} \beta_i X_i}, X = (X_1, X_2, \ldots, X_q), \tag{7}$$

where $h_0(t)$ is the non-parametric baseline hazard, X is the vector of time-independent predictor variables X_i and β_i are the initially unknown regression coefficients. The available methods for tie handling are Breslow [21], Efron [22] and Exact, while the mode of stepwise variable search can be selected from a number of options, namely Backward, Forward and Both. For each variable or strata included in the model, the method computes its regression coefficient, the confidence interval of the coefficient, the number e to the power of the coefficient and the P value of the performed Wald test. The result is shown in the form of a table, followed by the chi-square goodness of fit test and the proportional hazards assumption test results [23].

The user can also study the correlation between two variables by contingency table analysis. This method uses the R package `gmodels` (Warnes GR, Bolker B, Lumley T, Johnson RC, unpublished data) for some calculations. The application allows selecting and modifying the variables from which a contingency table with marginal values will be computed. There are three tests that can be carried out using the information in the table. These are the Pearson's chi-square test of independence, the Fisher's exact test for small samples and the McNemar's test [24]. In addition, the user can select whether the proportional values in the table will be represented by percentages, as in SPSS, or the way SAS depict them, on a scale from 0 to 1.

ANN-based method for survival analysis. Finally, the ANN-based method for predictive modeling, based on the Mani approach [25], fits a single-hidden-layer neural network with one input for each predictor variable, and as many outputs as groups are selected by the user to split the maximum follow-up time. The output represents the hazard rate for each interval of the follow-up time, which is computed by the formula

$$\hat{h}(t) = \begin{cases} 0 & \text{if } 1 \leq t \leq T_{surv} \\ 1 & \text{if } C = 1 \text{ and } T_{surv} < t \leq T \\ \dfrac{m_t}{n_t} & \text{if } C = 0 \text{ and } T_{surv} < t \leq T \end{cases} \tag{8}$$

where T represents the maximum value of the follow-up time, T_{surv} is the subject survival time, and C indicates whether the subject is censored ($C = 0$) or not ($C = 1$). For uncensored observations, the hazard is zero until the interval of the time of death and 1 thereafter. For censored observations, the hazard is zero until the interval of censoring time and then is calculated by the number of subjects with an observed death at instant t, or m_t, divided by the number of subjects at risk at t, n_t.

To fit the neural network, the application firstly preprocess the input data to determine the size of each the n intervals of the follow-up time. In order to increase the accuracy of the network, the same number of events are grouped in each interval, which can make them of different lengths. Afterwards, the *n* hazard rate outputs are calculated for every patient in the study with Eq 8.

The resulting model is obtained using a 10-fold cross validation process. In this manner, the preprocessed input is split in every iteration into a test set, a validation set and 8 training sets, each one with the same number of subjects. This validation procedure is repeated generating models with the R package `nnet` (Ripley B, unpublished data), with a different number of hidden-layer neurons. The network with the highest accuracy is presented as the final result. The accuracy of the model is calculated by

$$ACC = \frac{\sum_{i=1}^{N} \frac{G - |T_{surv} - T'_{surv}|}{G}}{N} \tag{9}$$

where N is the number of observations, G is the number of groups or intervals of the follow-up time, T_{surv} is the actual survival time of the observation and T'_{surv} is the estimated survival time for the observation.

While this network directly estimates the hazard rate of an individual subject for each follow-up time interval, it can be easy turned into a Kaplan-Meier estimator using the formula

$$\hat{S}(t) = \prod_{i|t_i<t} (1 - \hat{h}(t)) \tag{10}$$

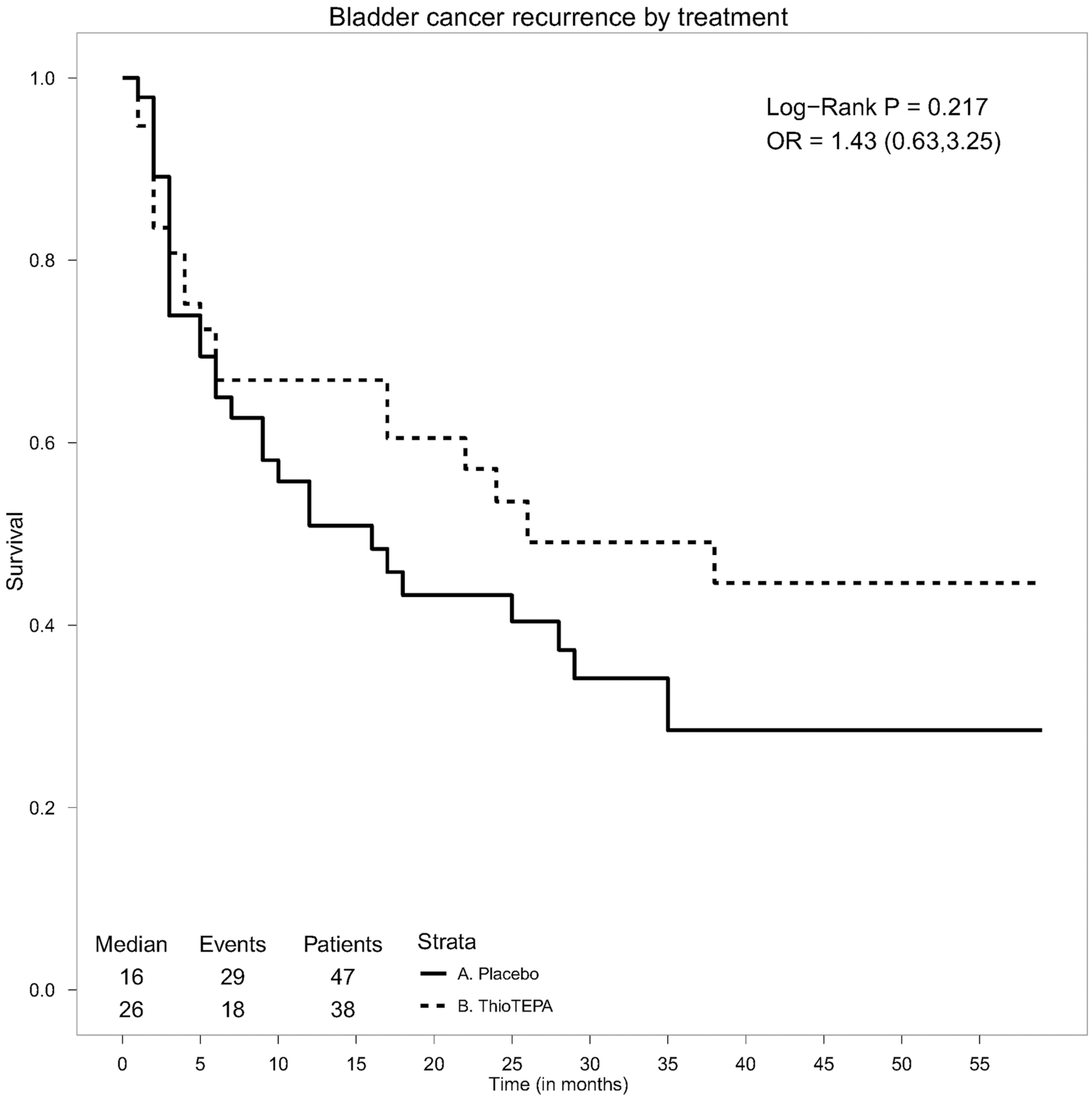
Results and Discussion

In order to illustrate how the application works, we will use a subset of the bladder cancer dataset present in the R package `survival`, consisting of the 85 observations of the first recurrence of bladder cancer for each patient [26] [27]. The variables included in the dataset are the received treatment (placebo or ThioTEPA), the number of tumours and the size of the largest tumour.

Fig 3 shows the survival curves stratified for the two-sample treatment. In the task creation form, we set the x-axis interval to 5 months, and ticked the checkboxes to show the table with the patients at risk and to use black and white graphics. We also selected the option to compare survival distributions by the Log-Rank test and used the received treatment as the study variable. Fig 4 shows the hazard plots, obtained by using the same parameter settings as done for the survival curve. The specific options for hazard curves are the bandwidths for the kernel smoothed function. The integer bandwidth values used in this example were 3 for both the Placebo and ThioTEPA.

The contingency table was created to compare the relationship between the number of tumours and the size of the largest tumour, with those variables having 8 and 7 categories, respectively. For each variable, we grouped all categories but the first and renamed them taking advantage of the web formulary options. Thus, the number of tumours becomes a two-stratum variable, with categories “1 tumour” and “More than 1”, and the size of the largest tumour gets reduced to the categories “Largest tumour of 1 cm” and “Largest tumour greater than 1 cm”. Fig 5 shows the contingency table followed by the computed Pearson’s chi-square test and the Fisher exact test.

We used a different dataset for the Cox proportional-hazards regression model. In this case, we chose the lung cancer dataset from package `survival` [28]. It provides variables such as the age of the patient and the sex, which were selected for this study. We created four strata for the age (ranges 39 to 45, 45 to 55, 56 to 65 and 66 to 82). Fig 6 shows the result after using the method Efron for tie handling and the backward stepwise variable search.

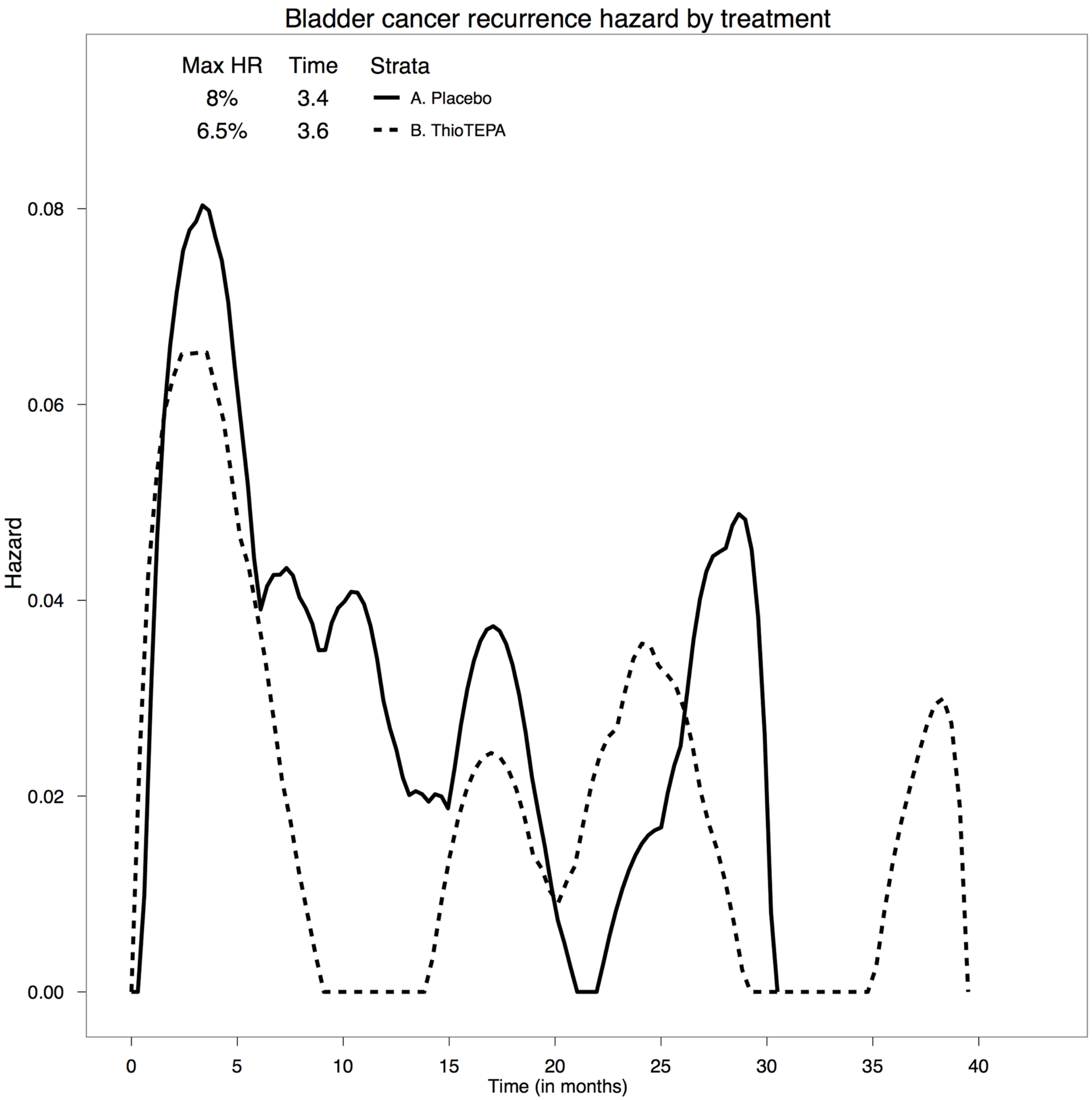


A	47	33	25	20	16	15	8	6	3	2	1	1
B	38	27	23	21	18	15	11	11	9	6	3	1

Number of patients at risk

Fig 3. Comparison of the survival curves of two groups of patients. The plot is followed by the number of patients at risk for each group.

doi:10.1371/journal.pone.0161135.g003

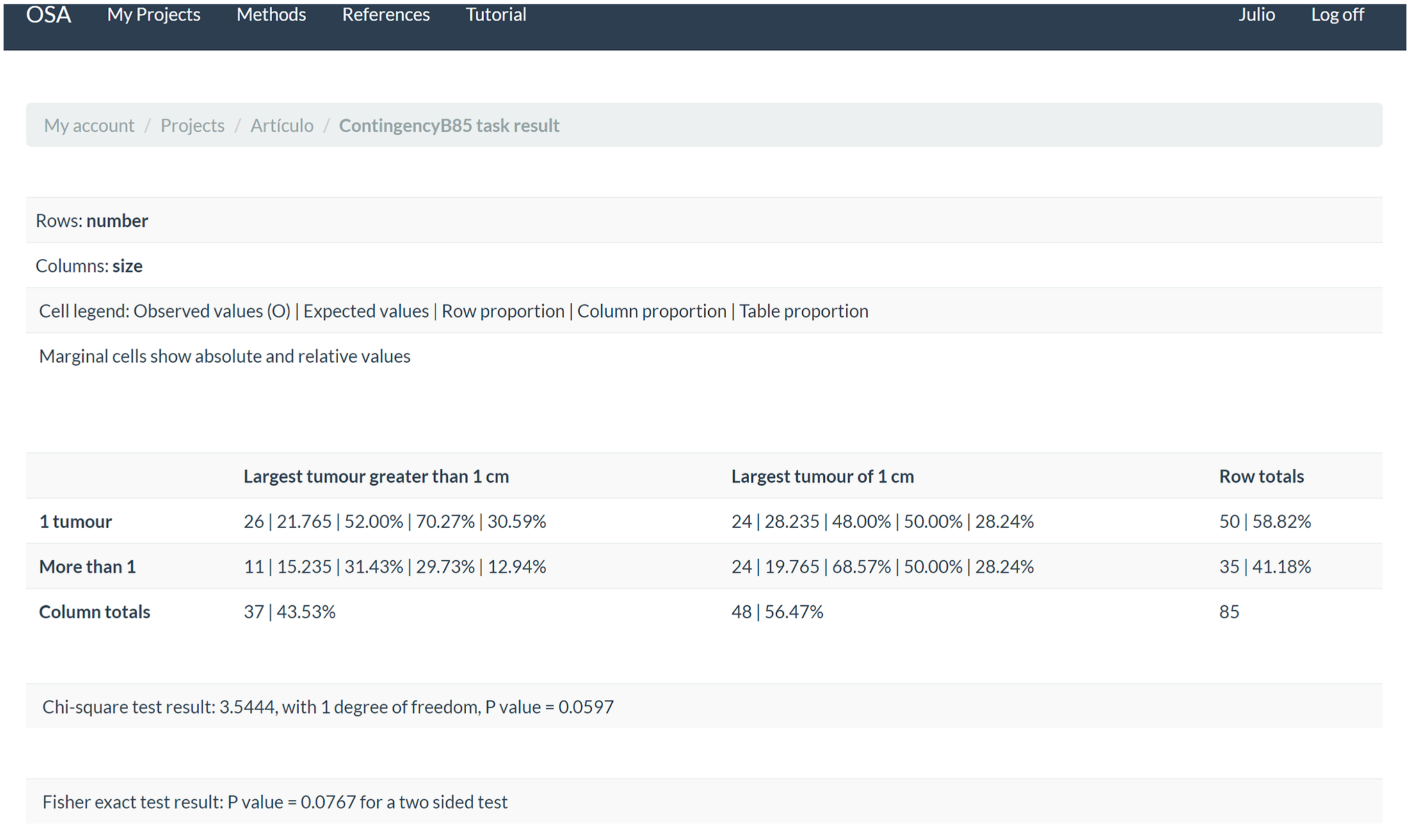


A	47	33	25	20	16	15	8	6	3
B	38	27	23	21	18	15	11	11	9

Number of patients at risk

Fig 4. Comparison of two hazard functions. The graph is followed by the number of patients at risk.

doi:10.1371/journal.pone.0161135.g004


OSA
My Projects
Methods
References
Tutorial
Julio
Log off
My account / Projects / Artículo / ContingencyB85 task result

Rows: **number**

Columns: **size**

Cell legend: Observed values (O) | Expected values | Row proportion | Column proportion | Table proportion

Marginal cells show absolute and relative values

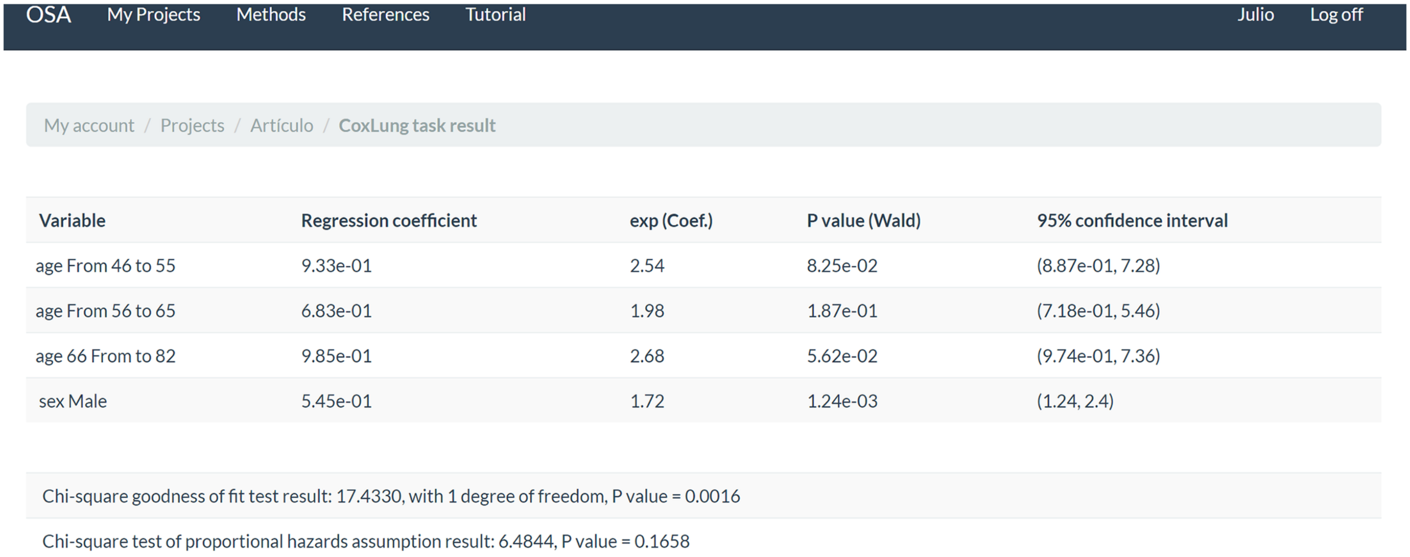
	Largest tumour greater than 1 cm	Largest tumour of 1 cm	Row totals
1 tumour	26 | 21.765 | 52.00% | 70.27% | 30.59%	24 | 28.235 | 48.00% | 50.00% | 28.24%	50 | 58.82%
More than 1	11 | 15.235 | 31.43% | 29.73% | 12.94%	24 | 19.765 | 68.57% | 50.00% | 28.24%	35 | 41.18%
Column totals	37 | 43.53%	48 | 56.47%	85

Chi-square test result: 3.5444, with 1 degree of freedom, P value = 0.0597

Fisher exact test result: P value = 0.0767 for a two sided test

Fig 5. The contingency table generated by the application as it is shown on the web.

doi:10.1371/journal.pone.0161135.g005



Variable	Regression coefficient	exp (Coef.)	P value (Wald)	95% confidence interval
age From 46 to 55	9.33e-01	2.54	8.25e-02	(8.87e-01, 7.28)
age From 56 to 65	6.83e-01	1.98	1.87e-01	(7.18e-01, 5.46)
age 66 From to 82	9.85e-01	2.68	5.62e-02	(9.74e-01, 7.36)
sex Male	5.45e-01	1.72	1.24e-03	(1.24, 2.4)

Chi-square goodness of fit test result: 17.4330, with 1 degree of freedom, P value = 0.0016

Chi-square test of proportional hazards assumption result: 6.4844, P value = 0.1658

Fig 6. The Cox regression result as it is generated by the application.

doi:10.1371/journal.pone.0161135.g006

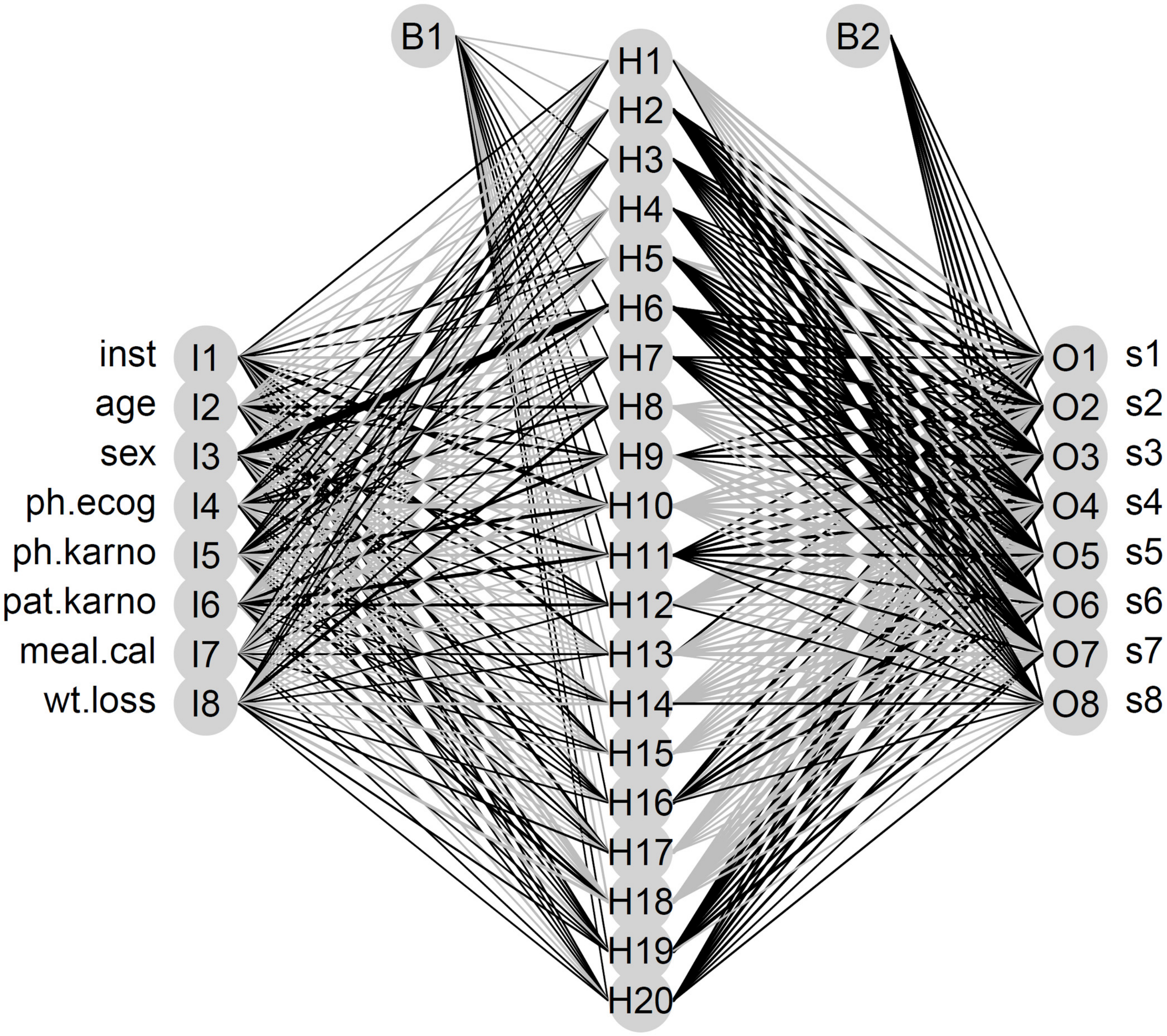


Fig 7. Artificial Neural Network. ANN fitted by the application, with a hidden layer of 20 neurons. The 8 inputs are the values of the lung cancer dataset variables, while the 8 outputs are the values of survival in each corresponding stratum of the follow-up time.

doi:10.1371/journal.pone.0161135.g007

Finally, we used the lung cancer dataset with the ANN-based predictor model. We stratified the age the same way we did with the previous model and discretised the follow-up time in 8 groups. Fig 7 depicts the resulting neural network which estimates the survival rate for each one of the calculated classes of the follow-up time. Fig 8 compares the actual survival curve of a specific censored patient from the dataset with the predicted one by the model. The Kaplan-Meier estimate obtained from the observed times (dotted curve) remains constant with the value of 1 until the censoring takes place. That is the reason why the first 6 stratum of the discretised follow-up time (shorter than the latter ones, as the majority of events take place at the beginning of the study) show visually significant differences between the actual (dotted) and the predicted curve (the solid one). However, after censoring, both curves develop almost identically.

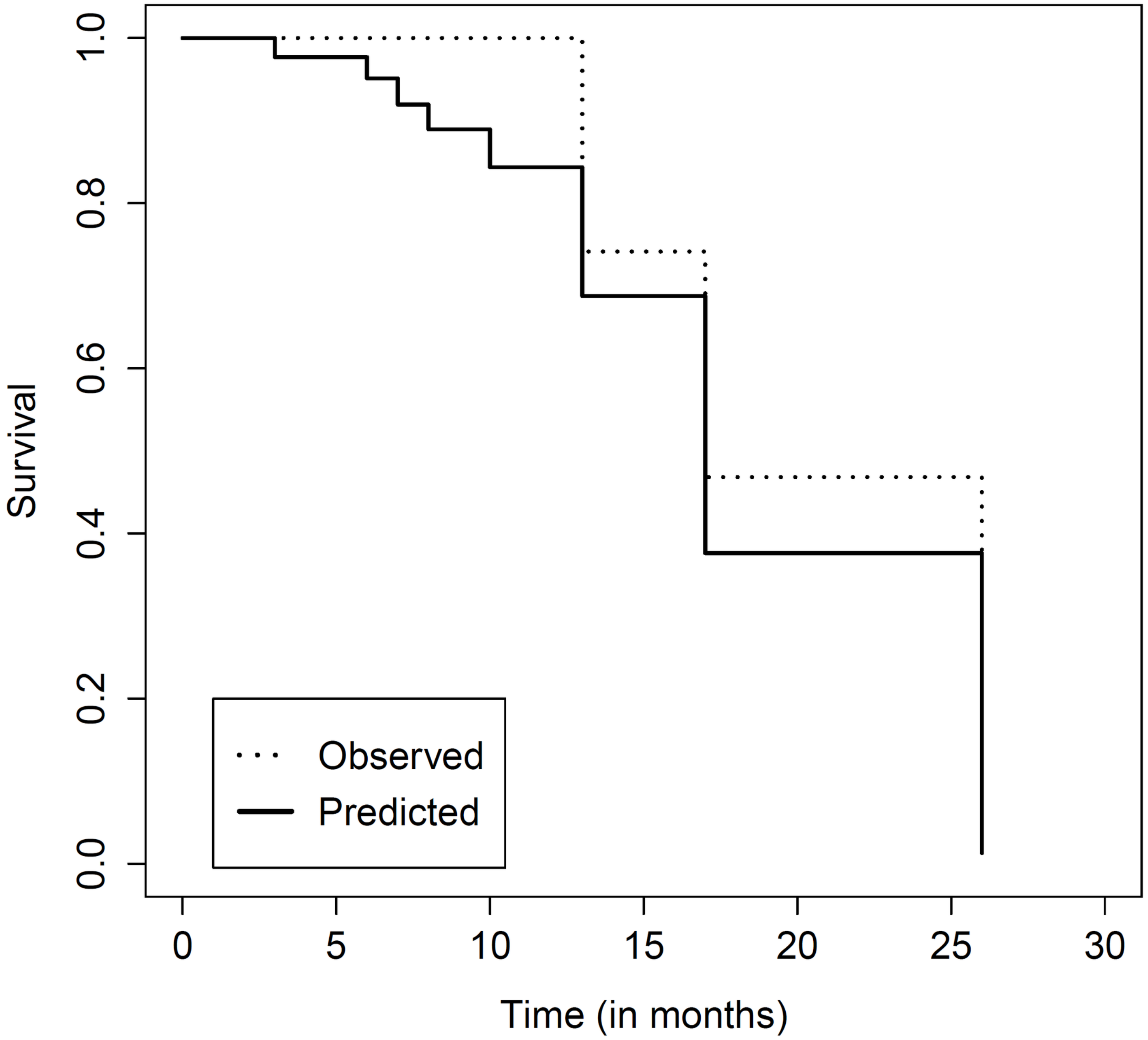


Fig 8. Predictive modelling. Comparison between the actual discrete time survival curve of a patient of lung cancer (dotted) and the predicted one generated by the fitted ANN-based model (solid).

doi:10.1371/journal.pone.0161135.g008

Conclusions

OSA is an easy to use and learn graphical front end for doing survival analysis. It makes use of the power and the wide variety of packages provided by R language to compute and plot the results. The ANN based model provides the user with a straightforward tool for discrete time survival analysis based on the Mani method. Although the other methods are available in commercial software like Stata or SAS, OSA stands out allowing users to conduct their studies without the need of learning complex command line syntax or complicated interfaces. Furthermore, it makes the graphical results available for download in PNG, PDF or EPS. As any web application, it can be used in every device with internet access, which permits researches to carry on with their work almost anywhere. This way, this open access application, which takes advantage of the computational power of the 27-node cluster, may be employed in helping in the field of survival analysis globally at no cost to its users. Future work includes Cox proportional-hazards for time dependent variables and new machine learning-based methods for survival analysis.

Acknowledgments

The authors acknowledge support through grants TIN2014-58516-c2-1-R and TIN2010-16556 from MICINN-SPAIN, and P08-TIC-4026 (Junta de Andalucía), all of which include FEDER funds.

Author Contributions

Conceived and designed the experiments: JMT JMJ.

Performed the experiments: JMT JLS.

Analyzed the data: JMT JLS DU.

Contributed reagents/materials/analysis tools: JMT NR EA.

Wrote the paper: JMT LF JMJ DU.

References

1. Arsene CTC, Lisboa E P J Biganzoli. Model selection with PLANN-CR-ARD. Lecture Notes in Computer Science. 2011;6692:pp. 210–219.
2. Ansari D, Nilsson J, Andersson R, Regnér S, Tingstedt B, Andersson B. Artificial neural networks predict survival from pancreatic cancer after radical surgery. American Journal of Surgery. 2013; 205(1): pp. 1–7. doi: 10.1016/j.amjsurg.2012.05.032 PMID: 23245432
3. Crowther MJ, Lambert PC. Stgenreg: A stata package for general parametric survival analysis. Journal of Statistical Software. 2013; 53(12):pp. 1–17. doi: 10.18637/jss.v053.i12
4. Zhang X, Akcin H. A SAS macro for direct adjusted survival curves based on Aalen's additive model. Computer Methods and Programs in Biomedicine. 2012; 108(1):pp. 310–317. doi: 10.1016/j.cmpb.2012.01.003 PMID: 22365671
5. Cansurv. Version 1.1. Statistical Methodology and Applications Branch, Data Modeling Branch, National Cancer Institute; 2012.
6. Józwiak K, Moerbeek M. PODSE: A computer program for optimal design of trials with discrete-time survival endpoints. Computer Methods and Programs in Biomedicine. 2013; 111(1):pp. 115–127. doi: 10.1016/j.cmpb.2013.02.005
7. Goswami C, Nakshatri H. PROGgene: gene expression based survival analysis web application for multiple cancers. Journal of Clinical Bioinformatics. 2013; 3(1):pp. 22. doi: 10.1186/2043-9113-3-22 PMID: 24165311
8. Györffy B, Lanczky A, Eklund AC, Denkert C, Budczies J, Li Q, et al. An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. Breast Cancer Research and Treatment. 2010; 123(3):pp. 725–731. doi: 10.1007/s10549-009-0674-9
9. Yang JS, Nam HJ, Seo M, Han SK, Choi Y, Nam HG, et al. OASIS: Online application for the survival analysis of lifespan assays performed in aging research. PLoS ONE. 2011; 6(8). doi: 10.1371/journal.pone.0023525
10. Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association. 1958; 53(282):pp. 457–481. doi: 10.1080/01621459.1958.10501452
11. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer chemotherapy reports Part 1. 1966; 50(3):pp. 163–170. PMID: 5910392
12. Peto R, Peto J. Asymptotically Efficient Rank Invariant Test Procedures. Journal of the Royal Statistical Society Series A (General). 1972; 135(2):pp. 185–207. doi: 10.2307/2344317

13. Tarone RE, Ware J. On Distribution-Free Tests for Equality of Survival Distributions. Biometrika. 1977; 64(1):pp. 156–160. doi: 10.1093/biomet/64.1.156

14. Muller HG, Wang JL. Hazard Rate Estimation under Random Censoring with Varying Kernels and Bandwidths. Biometrics. 1994; 50(1):pp. 61–76. doi: 10.2307/2533197 PMID: 8086616

15. Hess KR, Serachitopol DM, Brown BW. Hazard function estimators: A simulation study. Statistics in Medicine. 1999; 18(22):pp. 3075–3088. doi: 10.1002/(SICI)1097-0258(19991130)18:22%3C3075::AID-SIM244%3E3.0.CO;2-6 PMID: 10544307

16. Gasser T, Müller HG. Kernel estimation of regression functions. vol. 757 of Lecture Notes in Mathematics. Springer Berlin Heidelberg; 1979.

17. Aalen O. Nonparametric Inference for a Family of Counting Processes. The Annals of Statistics. 1978; 6(4):pp. 701–726. doi: 10.1214/aos/1176344247

18. Nelson W. Theory and Applications of Hazard Plotting for Censored Failure Data. Technometrics. 1972; 14(4):pp. 945–966. doi: 10.1080/00401706.1972.10488991

19. Venables WN, Ripley BD. Modern Applied Statistics with S. 4th ed. New York: Springer; 2002.

20. Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society Series B (Methodological). 1972; 34(2):pp. 187–220.

21. Breslow N. Covariance Analysis of Censored Survival Data. Biometrics. 1974; 30(1):pp. 89–99. doi: 10.2307/2529620 PMID: 4813387

22. Efron B. The Efficiency of Cox's Likelihood Function for Censored Data. Journal of the American Statistical Association. 1977; 72(359):pp. 557–565. doi: 10.1080/01621459.1977.10480613

23. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika. 1994; 81(3):pp. 515–526. doi: 10.1093/biomet/81.3.515

24. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947; 12(2):pp. 153–157. doi: 10.1007/BF02295996 PMID: 20254758

25. Mani DR, Drew J, Betz A, Datta P. Statistics and Data Mining Techniques for Lifetime Value Modeling. Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 1999;p. pp. 94–103.

26. Wei LJ, Lin DY, Weissfeld L. Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. Journal of the American Statistical Association. 1989; 84(408):pp. 1065–1073. doi: 10.1080/01621459.1989.10478873

27. Andrews DF, Hertzberg AM. DATA: A Collection of Problems from Many Fields for the Student and Research Worker. New York: Springer-Verlag; 1985.

28. Loprinzi CL, Laurie JA, Wieand HS, Krook JE, Novotny PJ, Kugler JW, et al. Prospective evaluation of prognostic variables from patient-completed questionnaires. Journal of Clinical Oncology. 1994; 12(3): pp. 601–607. PMID: 8120560