

Instituto Tecnológico y de Estudios Superiores de Monterrey

Monterrey Campus

School of Engineering and Sciences



**Machine Learning and Cox Based Benchmarking Tool: Exploration of  
Survival Models Associated with Chronic Degenerative Diseases**

A thesis presented by

**Jorge Andrés Orozco Sánchez**

Submitted to the  
School of Engineering and Sciences  
in partial fulfillment of the requirements for the degree of

Master of Science

in

Computer Science

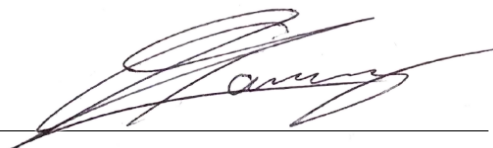
Monterrey, Nuevo León, June, 2020



# Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Monterrey

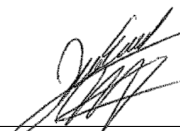
The committee members, hereby, certify that have read the thesis presented by Jorge Andrés Orozco Sánchez and that it is fully adequate in scope and quality as a partial requirement for the degree of Master of Science in Computer Sciences.



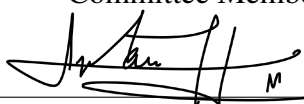
José Gerardo Tamez Peña  
Advisor's Institution  
Principal Advisor



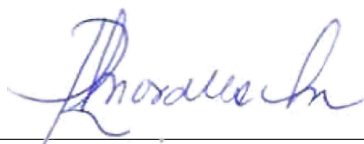
Víctor Manuel Treviño Alvarado  
Tecnológico de Monterrey  
Committee Member



Juan Emmanuel Martínez Ledesma  
Tecnológico de Monterrey  
Committee Member



Antonio Martínez Torteya  
Universidad de Monterrey  
Committee Member



Rubén Morales Menéndez  
Associate Dean of Graduate Studies  
School of Engineering and Sciences

Monterrey, Nuevo León, June, 2020



# Declaration of Authorship

I, Jorge Andrés Orozco Sánchez, declare that this thesis titled, Machine Learning and Cox Based Benchmarking Tool: Exploration of Survival Models Associated with Chronic Degenerative Diseases and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

---

Jorge Andrés Orozco Sánchez  
Monterrey, Nuevo León, June, 2020

©2020 by Jorge Andrés Orozco Sánchez  
All Rights Reserved



# Dedication

To all the people who were part of my education: teachers, friends, classmates, **my family and especially to you**. Thank you for all your unconditional trust, support, patience and encouragement. You were my main impetus to continue my education. A special feeling of gratitude to my dear parents, Jorge and Monica, whose efforts to provide us with the best education and life will always be rewarded. To my sisters Nicole, Michelle and my Aunt Emily who were always a part of the process and were with my parents in every moment that I could not be present.

Without any of you, this would never have been accomplished.

Thank you!





# Acknowledgements

Thanks to all the people who were part of this process. To José Tamez who without his knowledge and guidance this investigation could not have been carried out. To Victor Treviño, Emmanuel Martínez and Antonio Torteya who with their experience knew how to guide, correct and, above all, collaborate with this research. A special thanks to my two Master's classmates Erick and Arturo that without worry and dedication to the master's degree I would have missed many occasions. Thank you for those hours of study.

Thanks to all the people behind the scholarship and to the institution: Tecnológico de Monterrey. This work was partially supported by Secretaría de Educación Superior, Ciencia, Tecnología e Innovación part of Gobierno de la República del Ecuador and by Strategic Research Group of Bioinformatics for Clinical Diagnosis from Tecnológico de Monterrey. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health ([www.fnih.org](http://www.fnih.org)). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Thanks to OAI, The OAI is a public-private partnership comprised of five contracts (N01-AR-2-2258; N01-AR-2-2259; N01-AR-2-2260; N01-AR-2-2261; N01-AR-2-2262) funded by the National Institutes of Health, a branch of the Department of Health and Human Services, and conducted by the OAI Study Investigators. Private funding partners include Merck Research Laboratories; Novartis Pharmaceuticals Corporation, GlaxoSmithKline; and Pfizer, Inc. Private sector funding for the OAI is managed by the Foundation for the National Institutes of Health. This manuscript was prepared using an OAI public use data set and does not necessarily reflect the opinions or views of the OAI investigators, the NIH, or the private funding partners.



# **Machine Learning and Cox Based Benchmarking Tool: Exploration of Survival Models Associated with Chronic Degenerative Diseases**

**by**

**Jorge Andrés Orozco Sánchez**

## **Abstract**

The steady evolution of technology is changing the way physicians face health issues. In fact, the computer capacity to answer increasingly difficult questions continue to grow at a staggering rate, which has opened doors to groundbreaking research. Despite this, the complex nature of chronic-degenerative diseases and outstanding concern originated due to its significant incidence generates even more questions to answer. On the other hand, Machine Learning (ML) algorithms have already found beneficial information on those diseases. With this in mind, the present work reports the exploration of the CoxBenchmarking function applied to chronic-degenerative disease datasets associated with survival. CoxBenchmarking implementation is a computer-based benchmarking algorithm that compares the Survival Models that were constructed by several machine learning strategies. It was developed as an extension of FRESA.CAD package and uses its Random Holdout Cross-Validation. CoxBenchmarking provides an algorithm that generates eleven distinct survival models through feature selection of ML-based techniques: 6 wrappers and 5 filters. Besides, the function summarizes the results with tables and graphs by providing a well-ordered data structure and a plot function. The exploration includes the survival analysis applied to information of NBA players simulation, Wisconsin Prognostic Breast Cancer, San Jose Prognostic Breast Cancer, Osteoarthritis Initiative, and Alzheimer's Disease Neuroimaging Initiative. All the results were compared with previous works and tested with the same subjects, which allowed the fair comparison of all the ML techniques. In consequence, the exploration also helps in the efforts of creating new knowledge for each clinical case. After the study, clinical results were published in two conferences and a journal paper is being developed. Regarding the ML methods, the results do not inform a statistically significant difference between them. Consequently, the use of each of the methods depends on the case to be applied.



# List of Figures

2.1	Censoring types in a study. . . . .	19
2.2	KM Curves sex-stratified between 442 Alzheimer’s disease patients . . . . .	23
2.6	ROC Curve plotted by FRESA.CAD R package with some data classification analysis . . . . .	36
2.7	Confusion matrix plotted by FRESA.CAD R package in the ROC curve of random classifier with some data . . . . .	42
2.3	Stages of BSWiMS procedure based on Figure 2(a) in [104] . . . . .	43
2.4	Stages of Coxnet procedure . . . . .	44
2.5	Confusion matrix plotted by FRESA.CAD R package in the ROC curve of random classifier with some data . . . . .	45
3.1	ADNI/TADPOLE (a) Patient selection process. (b) Feature types used in this study. . . . .	53
3.2	Heat map with 301 features selected by all the Machine Learning Methods. On the top section, patients dendrogram and 4 bars with the subjects’ information about conversion, time to event, sex and APOE. On the left section, dendrogram of features and the information about the type of feature. Subject identification x-axis, features on y-axis. . . . .	55
3.3	Summary of the selection of Participants for the OAI experiment and data conditioning process. (a) Participants selection (b) Feature types and data conditioning process. *Absolute difference of X-ray measurements . . . . .	57
3.4	Approximation of positions on x-axis that Duryea provides in the dataset. Green section is the representation of the Medial Compartment and the red section is the representation for Lateral Compartment. Lines between the positions show the relation between them. . . . .	59
3.5	Example of CoxBenchmarking plot result. In the figure we can see (a) Accuracy barplot for accuracy of some model (b) Concordance Index Follow-up Times barplot (c) Number of features selected by the default Cox Benchmarking methods (d) Jaccard Index selected by the default Cox Benchmarking methods . . . . .	72
4.1	KM and ROC Curves for the Simulation experiment of 8 features (4 real - 4 random features) (a) BSWiMS (b) LASSO (c) RIDGE (d) ELASTICNET (e) GSPDAS (BeSS) (f) SPDAS (g) SPDAS with BIC . . . . .	77

4.2	KM and ROC Curves for the Simulation experiment of 100 features (4 real - 96 random features) (a) BSWiMS (b) LASSO (c) RIDGE (d) ELASTICNET (e) GSPDAS (BeSS) (f) SPDAS (g) SPDAS with BIC . . . . .	81
4.3	KM and ROC Curves for wrappers methods with the 4 real features and 996 random created with (a) BSWiMS (b) LASSO (c) RIDGE (d) ELASTICNET (e) GSPDAS (BeSS) (f) SPDAS (g) SPDAS with BIC . . . . .	82
4.4	KM and ROC Curves for wrappers methods in the Simulation experiment with 21 features. (a) BSWiMS (b) LASSO (c) RIDGE (d) ELASTICNET (e) GSPDAS (BeSS) (f) SPDAS (g) SPDAS with BIC . . . . .	85
4.5	KM and ROC Curves for filters methods in the Simulation experiment with 21 features (a) BSWiMS with Cox (b) LASSO with Cox (c) GSPDAS with Cox (d) Univariate Cox Analysis . . . . .	86
4.6	KM and ROC Curves for wrappers methods in the Simulation experiment with 101 features (a) BSWiMS (b) LASSO (c) RIDGE (d) ELASTICNET (e) GSPDAS (BeSS) (f) SPDAS (g) SPDAS with BIC . . . . .	88
4.7	KM and ROC Curves for filter methods in the Simulation experiment with 101 features (a) BSWiMS with Cox (b) LASSO with Cox (c) GSPDAS with Cox (d) Univariate Cox Analysis . . . . .	89
4.8	KM and ROC Curves for wrappers methods with the 11 real features and 990 random created with (a) BSWiMS (b) LASSO (c) RIDGE (d) ELASTICNET (e) GSPDAS (BeSS) (f) SPDAS (g) SPDAS with BIC . . . . .	90
4.9	Kaplan Meier (KM) and ROC curves for wrappers/embedded section. CoxNet showed the best accuracy on the classification and the best c-index on Risk and Follow-up times. (a) Model 1 BSWiMS KM (b) CoxNet KM (c) BeSS KM (d) BSWiMS ROC (e) CoxNet ROC (f) BeSS ROC . . . . .	92
4.10	Kaplan Meier (KM) and ROC curves for filters section. Cox Model build with BSWiMS features showed the best accuracy on the classification and the best c-index on Risk and Follow-up times. (a) Model 4 Cox with BSWiMS KM (b) Cox with BSWiMS ROC (c) Model 5 Cox with CoxNet KM (d) Cox with CoxNet ROC (e) Model 6 Cox with Univariate Cox KM (f) Cox with Univariate Cox ROC . . . . .	94
4.11	A heat map representation of the features associated with MCI to AD conversion. The figure shows the ten features selected by all the 4 methods in at least in the half of the iterations (horizontal axis) and subjects on the vertical axis. (F1) Mean volume (CP) of entorhinal, (F2) mean cortical thickness SD of Bankssts, (F3) APOE4, (F4) mean volume (WMP) of amygdala, (F5) mean cortical thickness AVG of Bankssts, (F6) mean volume (CP) of inferior temporal, (F7) absolute difference cortical thickness AVG of middle temporal, (F8) absolute difference of cortical thickness AVG of pars opercularis, (F9) absolute difference cortical thickness AVG of inferior parietal, (F10) mean cortical thickness SD of Rostral middle frontal . . . . .	95

4.12	KM and ROC Curves for (a) Experiment I model GPDAS as a filter c-index Risks = 0.75 (0.72, 0.78), ACC = 0.67 (0.62, 0.71) (b) Experiment II model LASSO c-index Risks = 0.65 (0.62, 0.67), ACC = 0.76 (0.72, 0.8) (c) Experiment III model BSWiMS as a filter c-index Risks = 0.74 (0.72, 0.77), ACC = 0.67 (0.62, 0.71) (d) Experiment IV model SPDAS.BIC c-index Risks = 0.67 (0.64, 0.70), ACC = 0.74 (0.7, 0.78) (e) Experiment V model BSWiMS as a filter c-index Risks = 0.68 (0.65, 0.71), ACC = 0.72 (0.68, 0.76) . . . . .	96
4.13	KM and ROC Curves for all the models of Experiment IV (CSF Measures + Cog-assessments + Radiomics). (a) BSWiMS Model [ACC = 0.76 (0.71, 0.8) c-index FT = 0.69 (0.66, 0.72) c-index Risks = 0.65 (0.63, 0.68)] (b) LASSO Model [ACC = 0.75 (0.71, 0.79) c-index FT = 0.77 (0.75, 0.79) c-index Risks = 0.66 (0.63, 0.69)] (c) RIDGE Model [ACC = 0.71 (0.66, 0.75) c-index FT = 0.91 (0.89, 0.92) c-index Risks = 0.63 (0.61, 0.66)]. (d) GPDAS Model [ACC = 0.74 (0.69, 0.78) c-index FT = 0.59 (0.55, 0.62) c-index Risks = 0.65 (0.62, 0.68)]. (e) SPDAS Model [ACC = 0.76 (0.72, 0.8) c-index FT = 0.63 (0.6, 0.67) c-index Risks = 0.66 (0.63, 0.68)]. (f) SPDAS.BIC Model [ACC = 0.74 (0.69, 0.78) c-index FT = 0.61 (0.58, 0.64) c-index Risks = 0.66 (0.63, 0.68)]	97
4.14	A heat map representation of the features associated with MCI to AD conversion. The figure shows the 8 features selected by all the 6 methods in at least in the half of the iterations (horizontal axis) and subjects on the vertical axis. (F1) $A\beta_{1-42}$ , (F2) CDSRB, (F3) RAVLT immediate, (F4) ADAS13, (F5) FAQ, (F6) Mean volume CP Inferior Temporal, (F7) Mean volume CP Entorhinal, (F8) Mean cortical thickness SD Bankssts. . . . .	99
4.15	A heat map representation of the features associated with TKR outcome on OAI patients. The figure shows the 15 features selected by all the 8 methods (Seven wrappers and Unicox) in at least in the half of the iterations (horizontal axis) and subjects on the vertical axis. (F1) Mean KL, (F2) Absolute difference KL, (F3) KOOS Sports, (F4) absolute difference osteophytes grades of femur lateral compartment, (F5) absolute difference lateral tibial plateau margin, (F6) Mean osteophytes grades of femur lateral compartment, (F7) absolute difference medial minimum JSW, (F8) Mean sclerosis grades tibia medial compartment, (F9) mean x coordinate of minimum JSW, (F10) KOOS Quality of life, (F11) absolute difference of FTA, (F12) raw feature osteophytes grades of femur medial compartment, (F13) absolute difference sclerosis grades of femur medial compartment, (F14) Raw difference between position x=150 and x=850, (F15) Raw osteophytes grades femur lateral compartment. . . . .	104
4.16	KM and ROC curves of the wrapper models built with Prognostic Wisconsin Breast Cancer Database. (a) BSWiMS (b) LASSO (c) RIDGE (d) ELASTIC-NET (e) GSPDAS (BeSS) (f) SPDAS (g) SPDAS with BIC . . . . .	107
4.17	(a) KM and (b) ROC curves for the only model which got results on the San Jose Survival analysis, BSWiMS. . . . .	109
5.1	Bar plot for Concordance Index Follow-up Times. (a) 4 real and 4 random features (b) 4 real and 96 random features (c) 11 real and 10 random features (d) 11 real and 90 random features. . . . .	117

5.2	Bar plot for the number of features and Jaccard Index. Left size shows the barplot for Jaccard Index of each model and Right size shows the mean number of features (a) 4 real and 4 random features (b) 4 real and 96 random features (c) 11 real and 10 random features (d) 11 real and 90 random features.	118
5.3	Coxnet ROC with 296 patients who suffered the conversion or have a censored event in more than 4 years. . . . .	121
5.4	BSWiMS ROC with 347 patients who suffered the conversion or have a censored event in more than 3 years. . . . .	123
5.5	Heatmap of the features used in the Experiment VI. . . . .	124
5.6	Plot of CoxBenchmarking analysis of OAI data. Barplots for (a) Balanced Error (b) Accuracy (c) ROC.AUC (d) Sensitivity (e) Specificity (f) C-Index Risks (g) C-Index Follow-Up Times (h) Jaccard Index (i) Mean of features . .	127
5.7	Plot of CoxBenchmarking analysis of BRCA Wisconsin data. Barplots for (a) Balanced Error (b) Accuracy (c) ROC.AUC (d) Specificity (e) C-Index Risks (f) C-Index Follow-Up Times (g) Sensitivity (h) Jaccard Index (i) Mean of features . . . . .	129
5.8	KM Curves stratified by the median of (a) HH1 Level z-Score $p=0.063$ (b) PAM50 Score $p=0.22$ , (c) Oncotype $p=0.50$ . . . . .	131



# List of Tables

2.1	Kinds of survival models . . . . .	20
2.2	Group 1 (Males) alternative ordered layout. $m_f$ is the number of events at time $t$ . $q_f$ number of censored subjects at time $t$ . $n_f$ set of subjects who are at risk of failure . . . . .	22
2.3	Group 2 (Females) alternative ordered layout. $m_f$ is the number of events at time $t$ . $q_f$ number of censored subjects at time $t$ . $n_f$ set of subjects who are at risk of failure . . . . .	22
2.4	Tables 2.2 and 2.3 combined. Number one 1 in the underscore section of the columns' names denotes group 1 (males) and the number two 2 (females). $e$ is the expected value $t$ . . . . .	24
2.5	Table with expected values and a column with observed minus expected values	25
2.6	HR value and $\beta$ coefficient effect in the Hazard ratio function . . . . .	26
2.7	ADNI/TADPOLE data of male patients that will be used on one of this experiments of this thesis. Time to event is in days, status=1 represents that the patient suffered the conversion of MCI to AD . . . . .	33
2.8	Review . . . . .	41
3.1	Features to be related with the NBA careers of 1000 NBA players simulated information. * the probability of success in binomial distribution . . . . .	49
3.2	Features to be related with the NBA careers of 1000 NBA players simulated information. . . . .	50
3.3	Characteristics of tadpole challenge subjects used in this study. 187 patients presented the MCI to AD conversion event and 255 maintained the MCI diagnosis during the observation period. The normal control patients (n=233) were used as reference controls. APOE status (1: Noncarriers 2: Heterozygotes 3:Homozygotes) . . . . .	54
3.4	The relation between positions in the x-axis. The difference between the first column which belongs to the Medial compartment and the second column belonging to the Lateral compartment is calculated (Position 150 - Position 850)	59
3.5	Summary of the features used in this Experiment. Detailed information of the features can be found in the previous paragraphs. . . . .	60
3.6	Summary of the features of Prognostic Wisconsin BRCA Database. A total of 198 patients and 34 features are used. . . . .	61
3.7	Survival groups of patients suffering of BRCA in San Jose experiment . . . .	62

3.8	San Jose BRCA information summary of features. First column describe the feature type, the second details the group of the features. Third column shows the number of features on that group and the last column shows the description of the group . . . . .	63
3.9	Default algorithms used in CoxBenchmarking method. . . . .	67
3.10	Output of Cox Benchmarking Model. Left side lists the name of the objects in the model and Right side describes it. . . . .	70
4.1	Classification and survival stats for wrapper methods in the Simulation experiment of 8 features (4 real - 4 random). I = BSWiMS, II = LASSO, III = RIDGE, IV = ELASTICNET, V = GSPDAS (BESS), VI = SPDAS (BESS.SEQUENTIAL), VII = SPDAS.BIC (BESS.SEQUENTIAL.BIC). Worst scores for each stat are bolded. . . . .	78
4.2	Classification and survival stats for wrapper methods in the Simulation experiment of 100 features (4 real - 96 random). I = BSWiMS, II = LASSO, III = RIDGE, IV = ELASTICNET, V = GSPDAS (BESS), VI = SPDAS (BESS.SEQUENTIAL), VII = SPDAS.BIC (BESS.SEQUENTIAL.BIC). Worst scores for each stat are bolded. . . . .	79
4.3	Classification and survival stats for filter methods with 4 real features and 96 random variables. I = Cox with BSWiMS, II = Cox with LASSO, III = Cox with BESS IV = Univariate Cox. Best scores for each stat are bolded. . . . .	80
4.4	Classification and survival stats for wrapper methods in the Simulation experiment of 21 features (11 real - 10 random). I = BSWiMS, II = LASSO, III = RIDGE, IV = ELASTICNET, V = GSPDAS (BESS), VI = SPDAS (BESS.SEQUENTIAL), VII = SPDAS.BIC (BESS.SEQUENTIAL.BIC). Best scores for each stat are bolded. . . . .	83
4.5	Classification and survival stats for filter methods with 11 real features and 10 random variables. I = Cox with BSWiMS, II = Cox with LASSO, III = Cox with BESS IV = Univariate Cox. Best scores for each stat are bolded. . . . .	84
4.6	Classification and survival stats for wrapper methods in the Simulation experiment of 21 features (11 real - 100 random). I = BSWiMS, II = LASSO, III = RIDGE, IV = ELASTICNET, V = GSPDAS (BESS), VI = SPDAS (BESS.SEQUENTIAL), VII = SPDAS.BIC (BESS.SEQUENTIAL.BIC). Best scores for each stat are bolded. . . . .	87
4.7	Classification and survival stats for filter methods with 11 real features and 100 random variables. I = Cox with BSWiMS, II = Cox with LASSO, III = Cox with BESS IV = Univariate Cox. Best scores for each stat are bolded. . . . .	89
4.8	Models predictions statistics. contains c-index of follow-up times predictions, the p-value on log rank test between low-high risk curves, area under the curve, accuracy, sensitivity and specificity with their 95% confidence intervals. W=Wrappers, F=Filters, FS = Feature Size, JI = Jaccard Index. Best scores for each stat are bolded. . . . .	91

4.9	Characteristics and ranking of ten features selected in almost the half of the iterations. The ranking was ordered based on the number of times selected and then ordered depending on the p-value of univariate cox analysis. [FT = feature type; M=mean, P= polymorphism, A=absolute difference], [MT = measure type; V=volume (mm3), G = gene, M = cortical thickness (mm)], [M1 = BSWiMS, M2 = COXNET/LASSO, M3 = BeSS, M4 = Univariate Cox] P. Value significance: $\alpha < 0.1$ , * < 0.05, ** < 0.01, *** < 0.001, **** < $10^{-04}$ . . . . .	93
4.10	Characteristics and ranking of eight features selected in almost the half of the iterations. The ranking was ordered based on the number of times selected, then ordered depending on the p-value of univariate cox analysis and finally, the concordance index of the univariate model. [FT = feature type; M=mean, C=Cog. Assessment, P=CSF Measure], [mt = measure type; v=volume (mm3), p = protein, ct = cortical thickness (mm)], [M1 = BSWiMS, M2 = LASSO, M3 = RIDGE, M4 = GPDAS, M5=SPDAS, M6=SPDAS.BIC, M7=Univariate Cox] P. Value significance: < 0.1, * < 0.05, ** < 0.01, *** < 0.001, **** < $10^{-04}$ . . . . .	100
4.11	Main classification (Accuracy ACC) and survival stats (c-index FT) for all the models on Experiments I, II, III, and VI. The complete stats for all the experiments are shown in the Table 4 at the appendix section. Bold number on each column indicates the best stat on that specific experiment. Tiebreakers were performed by the AUC value and 95%CI, *AUC =0.81(0.77,0.85), **AUC =0.84(0.81,0.88). M = Models (1=BSWiMS, 2=LASSO, 3=RIDGE, 4=GP-DAS, 5=SPDAS, 6=SPDAS.BIC, 6=Univariate Cox analysis). MT = Model Type (W = Wrapper, F=Filter). I = CSF Measures. II = Cog-assessments. III = Radiomics. VI = CSF Measures + Cog-assessments + Radiomics. . . . .	101
4.12	Main classification (Accuracy ACC) and survival stats (c-index FT) for all the models with Wrapper and Filter Methods. The complete stats for all the models are shown in the Table 4.13 for wrapper methods and Table for filter Methods. Best scores for each stat are bolded. . . . .	102
4.13	Classification and survival stats for all the models with Wrapper and Filter Methods in the OAI analysis. Best scores for each stat are bolded. . . . .	103
4.14	Classification and survival stats for filter methods that analyzed the Prognostic Wisconsin Breast Cancer Database. I = BSWiMS, II = LASSO, III = RIDGE, IV = ELASTICNET, V = GSPDAS (BESS), VI = SPDAS (BESS.SEQUENTIAL), VII = SPDAS.BIC (BESS.SEQUENTIAL.BIC). Best scores for each stat are bolded. . . . .	106
4.15	Classification and survival stats for filter methods that analyzed the Prognostic Wisconsin Breast Cancer Database. I = Cox with BSWiMS, II = Cox with LASSO, III = Cox with BESS IV = Univariate Cox. Best scores for each stat are bolded. . . . .	106
5.1	Random variables selected in more than the half of the iterations for all the experiments . . . . .	114

5.2 Mean variables selected by each method in the models. False discovery rate is shown inside the parenthesis. FDR is calculated with the ratio of random variables selected in more than the half of the models. . . . . 114

5.3 Cox Model summary using the eleven real features with the simulated data. First column shows the features related with the outcome, second column the effect size of each feature and the third the Hazard Ratio of each feature  $\exp(\beta)$  116

5.4 Characteristics and ranking of eight features selected in almost the half of the iterations. The ranking was ordered based on the number of times selected, then ordered depending on the p-value of univariate cox analysis and finally, the concordance index of the univariate model. [FT = feature type; M=mean, C=Cog. Assessment, P=CSF Measure], [mt = measure type; v=volume (mm3), p = protein, ct = cortical thickness (mm)], [M1 = BSWiMS, M2 = LASSO, M3 = RIDGE, M4 = GPDAS, M5=SPDAS, M6=SPDAS.BIC, M7=Univariate Cox] P. Value significance: < 0.1, \* < 0.05, \*\* < 0.01, \*\*\* < 0.001, \*\*\*\* <  $10^{-04}$  . . . . . 125

# Contents

<b>Abstract</b>	<b>ix</b>
<b>List of Figures</b>	<b>xiv</b>
<b>List of Tables</b>	<b>xviii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Problem definition and motivation . . . . .	4
1.2 Objectives . . . . .	7
1.2.1 Main objective . . . . .	7
1.2.2 Particular objectives . . . . .	8
1.3 Hypothesis . . . . .	8
1.4 Solution Overview . . . . .	9
1.5 Main Contributions . . . . .	10
1.6 Outline of the Thesis . . . . .	10
<b>2 Background Work</b>	<b>13</b>
2.1 Chronic degenerative diseases . . . . .	14
2.1.1 Breast cancer . . . . .	14
2.1.2 Osteoarthritis . . . . .	15
2.1.3 Dementia . . . . .	15
2.2 Survival Analysis . . . . .	16
2.2.1 Hazard Function . . . . .	17
2.2.2 Formulation . . . . .	17
2.2.3 Censoring . . . . .	18
2.2.4 Approaches to Survival Models . . . . .	20
2.3 Kaplan Meier . . . . .	21
2.3.1 Survminer . . . . .	24
2.3.2 Log Rank Test . . . . .	24
2.4 Cox Model . . . . .	26
2.5 Statistical Learning and Machine Learning Methods . . . . .	27
2.5.1 FRESA.CAD . . . . .	27
2.5.2 GLMNET . . . . .	28
2.5.3 BeSS . . . . .	30
2.6 Model evaluation metrics . . . . .	31

2.6.1	Jaccard Index . . . . .	31
2.6.2	Concordance index . . . . .	31
2.6.3	Log Rank . . . . .	32
2.6.4	Classification results . . . . .	32
2.6.5	Accuracy . . . . .	34
2.6.6	Sensitivity and Specificity . . . . .	34
2.6.7	Receiver operating characteristic . . . . .	35
2.7	Validation . . . . .	36
2.7.1	Cross-Validation . . . . .	36
2.7.2	Leave-one-out Cross-Validation . . . . .	37
2.7.3	<i>k</i> -fold Cross-Validation . . . . .	38
2.7.4	Holdout and Repeated Holdout CV . . . . .	38
2.8	Summary . . . . .	39
2.8.1	Chronic-degenerative disease . . . . .	39
2.8.2	Survival analysis . . . . .	40
2.8.3	Machine Learning techniques . . . . .	40
2.8.4	Metrics . . . . .	41
<b>3</b>	<b>Methodology</b>	<b>47</b>
3.1	Experiments: Data acquisition and preparation . . . . .	48
3.1.1	Simulation data . . . . .	48
3.1.2	TADPOLE/ADNI . . . . .	51
3.1.3	Osteoarthritis Initiative: OAI . . . . .	55
3.1.4	Prognostic Wisconsin Breast Cancer Database . . . . .	61
3.1.5	Prognostic San Jose Hospital Breast Cancer Database . . . . .	62
3.2	Cox Benchmarking implementation . . . . .	64
3.2.1	Default ML/SL algorithms . . . . .	65
3.2.2	Cox Benchmarking Algorithm . . . . .	68
3.2.3	Random Holdout Cross Validation implementation . . . . .	68
3.2.4	Cox Benchmarking Model . . . . .	71
3.3	Summary . . . . .	71
3.3.1	Experiments: Data acquisition, preparation and analysis . . . . .	71
3.3.2	Cox benchmarking . . . . .	73
<b>4</b>	<b>Experiment Results</b>	<b>75</b>
4.1	Simulation data . . . . .	75
4.1.1	4 variables . . . . .	76
4.1.2	11 variables . . . . .	83
4.2	TADPOLE-ADNI . . . . .	91
4.2.1	Survival Models Associated with MCI to AD Conversion with qMRI features . . . . .	91
4.2.2	Prediction of MCI to AD Risk of Conversion Survival Models: qMRI vs CSF Measures and Cognitive Assessments . . . . .	95
4.3	Osteoarthritis Initiative: OAI . . . . .	102
4.4	Prognostic Wisconsin Breast Cancer Database . . . . .	105

4.5	Prognostic San Jose Hospital Breast Cancer Database . . . . .	108
<b>5</b>	<b>Discussions and Conclusions</b>	<b>111</b>
5.1	Discussion . . . . .	112
5.2	Simulation data . . . . .	113
5.3	TADPOLE-ADNI . . . . .	119
5.3.1	Survival Models Associated with MCI to AD Conversion with qMRI features . . . . .	119
5.3.2	Prediction of MCI to AD Risk of Conversion Survival Models: qMRI vs CSF Measures and Cognitive Assessments . . . . .	121
5.4	Osteoarthritis Initiative: OAI . . . . .	126
5.5	Prognostic Wisconsin Breast Cancer Database . . . . .	129
5.6	Prognostic San Jose Breast Cancer Database . . . . .	130
5.7	Conclusions . . . . .	132
5.8	Future work . . . . .	132
	<b>Bibliography</b>	<b>145</b>





# Chapter 1

## Introduction

The strong evolution of technology is transforming the way things are done today. Accordingly, it is more common that day to day new technology tools interfere in our daily life. For example, the healthcare industry is continually evolving the way how they face the problems. Particularly, they are extending the paths to store and analyze all the data they are producing. These changes will affect the direction healthcare services are delivered and applied [123]. In this context, the vast capacity of modern processors to interpret information and the facility to store large datasets influence the ability to explore clinical information. In practical terms for the medical area, this helps to investigate diseases and obtain learning on specific diagnoses, treatments, and prognosis. Subsequently, this investigation helps in the knowledge of these diseases and improves the treatments of them. Despite the continuous improvement of the knowledge, there are diseases that, although being constantly studied, have not been able to find a definitive solution. Hence, using the technology as a tool or direction must be the path to find solutions and apply new methods to address this challenging problem [49].

The importance of these tools is bigger when the diseases to be analyzed are common, have high incidence and prevalence, and affect not only people who suffer from the disease but also generate a significant public health expenditure. This thesis will consider three of the most important chronic degenerative diseases, Alzheimer's dementia (AD), breast cancer (BRCA), and osteoarthritis (OA). Trying to apply automatic learning tools in the information of patients suffering from these diseases and thereby find a better understanding of the survival analysis in each disease. Dementia is one of the most critical syndromes worldwide and has one of the highest prevalence rates among the elderly [6]. According to Alzheimer's Disease International, there are currently more than 50 million cases of dementia and possibly 10 million new cases each year. Among them, Alzheimer dementia (AD) represents 60%-70% of cases [7]. Despite its very impressive statistics, the most worrisome is the lack of effective therapy to control AD which led to the fact that between 2000 and 2015 the number of deaths caused by the disease has increased by 123%. An amount larger than the corresponding percentages of diseases such as Prostate cancer and Heart diseases [6]. Accordingly, a clear understanding of the AD process and stages is essential in developing effective therapies. Breast cancer (BRCA) is the most commonly occurring cancer in women and the second most common cancer overall. In Mexico, the incidence and mortality of breast cancer have risen in the past years. Changes in health-care policies in Mexico (since 2003) have changed the way to treat this disease and now they focus on early detection and treatment since it has cure but it

depends on the early diagnosis [2]. If this disease does not have an early diagnosis, it takes the patient to death and that is why BRCA has a high mortality rate. GLOBOCAN's prediction for Mexico's breast cancer mortality rate by 2030, is that 24 386 women will be diagnosed and 9778 (40%) will die [21]. On the other hand, Osteoarthritis (OA) is the most common form of arthritis; it causes considerable disability in elderly populations. Osteoarthritis does not have a consistent technique that can be used for its early diagnosis and is more common than expected. In Mexico, the prevalence of osteoarthritis was 10.5% [82] and despite a high prevalence, there is no treatment or medication that can cure it. Furthermore, there is no way to reverse or halt the disease evolution what it causes always an event on its prognosis.

The diseases create a big problem for insurance companies, government, and patients due to the money they must expend on the treatments for these incurable diseases [49]. This is the main reason why there is a lot of interest in studying them and finding solutions to this problem. The attention started since there was enough capacity to analyze a good quantity of data, on 1995, Olvi Mangasarian et al. [67] presented one of the first advances on this topic using just computer power to process and analyze data. They proposed some linear programming-based machine learning techniques, that are used to increase the accuracy of breast cancer diagnosis and prognosis. They use the information found on a liquid sample that was extracted by the process of breast fine-needle aspiration (FNA) [67]. There are other methods and techniques that use different kinds of information for this disease. Joseph Cruz, David Wishart in 2006 [25] mentioned that there are some machine learning applications in cancer prediction and prognosis. They summarized some research that was published before the article; in the majority all the research use information like proteomic and genomic data about the patient. Furthermore, they mentioned some machine learning techniques that are commonly used on the breast cancer prognosis research such as Support vector machine (SVM), Genetic Algorithms, Artificial Neural Networks, k-Nearest Neighbor, Naïve Bayes, and Decision Trees.

A decade later the applications of machine learning changed a bit. A growing trend is noted in the use of supervised learning techniques, such as SVMs and Bayesian networks (BNs). Kourou et al. [62] reviewed more than 7510 articles about breast cancer prediction and prognosis between 2010 and 2014. Then they found three clusters inside the research: prediction, recurrence, and survival; for which they selected the most relevant publications for each group. Taking into concern that information, for cancer prediction is used the following type of data: mammographic, demographic, SNPs (single nucleotide polymorphisms) clinical and pathologic, for cancer survival the information used is from the type: clinical, genomics, molecular and for the last group, recurrence, imaging tissue genomic, blood genomic, genetic and pathologic [62].

In the context of AD, there are very good methods for clinical dementia diagnosis, based on patient reports, cognitive observation, and symptomatology [96]. Some risk factors of developing the disease have even been determined, where the presence of APOE4 is a well-known genetic factor [23]. In some cases, there are patients who do not have enough conditions to be diagnosed with AD; but fall between the cognitive changes of aging and early dementia, their condition is known as Mild Cognitive Impairment (MCI) [59]. Therefore; MCI, in future AD patients, is an intermediate stage between normal aging and clinical

dementia. Henceforth, MCI diagnosis represents a critical condition due to the increased risk of early AD findings [36, 40]. However, detecting AD in the early stage is complex; considering that only 33.6% of the MCI subjects convert to clinical AD [74]. Hence, standards have not been defined on the best neuropsychological results that should be used to measure early AD [23]. On the other hand, imaging has the ability to visualize early AD [74], and several imaging-biomarkers have been found in clinical images. These Imaging-biomarkers have been associated with the conversion from MCI to AD [96, 37, 102, 52, 47]. The vast majority of existing imaging studies have used information from magnetic resonance imaging (MRI) and positron-emitting tomography (PET); and both modalities remain as recommendations to monitor the progression of the disease, in addition, to detect the current stage of neuronal degeneration [56]. Although recent studies have shown that PET has a great capacity to diagnose the disease [79, 80, 8, 99] and that MRI details related to AD neuronal degeneration in patients with MCI is not detectable by experts, MRI is preferable to PET because PET facilities are scarce [99] compared to MRI [96].

In the case of osteoarthritis, there is less information about survival. Machine learning tries to identify it, classify or predict how much risk exists in acquiring the disease. Beth G. Ashinsky et al. in 2017 exposed a machine learning approach for predicting early symptomatic osteoarthritis with the classification of magnetic resonance images (MRI) [10]. One year later Tiulpin et al. in 2018 [115] presented an automatic knee osteoarthritis diagnosis from radiographs that used deep learning to analyze the data. In the same year, Ting Hu [49] proposed an evolutionary learning and network approach to identify the key metabolites for this disease. Using genetic, epigenetic, and biochemical markers taken from the plasma of blood samples they could find information about relevant metabolites. Moreover, there is other information that could help to understand how to work with this disease. Iliou and Anagnostopoulos [50] presented many machine learning techniques to detect and extract features from information about osteoporosis, which is a similar illness to OA. They used clinical information about patients who suffer from osteoporosis to create a score that could predict the disease.

Considering all the most important disciplines for the development of this thesis research, it should be stated the tools that will be used to develop the survival analysis of the diseases. In the field of statistics, the term of survival analysis is used as the technique that let analyze the expected duration of time until an event happens [103, 95]. Within this area, several terms will be of great importance. Censored events, what happens when we only know partial information about an event; hazard function, that returns the probability of an event happening between time  $t$  and  $dt$ ; Kaplan Meier that is a nonparametric statistic used to estimate the survival function from the collection of the life data of a particular object [103, 95, 53] and Cox model, One of the most know models for analysis of failure time regression data [11], commonly used mathematical modeling technique for estimating survival curves when considering some descriptive variables simultaneously [58]. On the other hand, this kind of statistical methods could be applied to different manners and machine learning is one of them. In this work, we will use different machine learning strategies to find hazards models than mainly will use three software packages that include survival analysis algorithms. First, in 2011, N. Simon et al. presented different regularization paths for cox's proportional hazards model. Penalized Cox Regression (CoxNet) component of the GLMNET R package.

The main algorithm in the package fits the Cox Model regularized by an elastic net penalty which lets the user change its parameters and turn the algorithm into another approach. Second, in 2016, Bootstrapped Stage-wise Model Selection (BSWiMS) was proposed by Jose Tamez who later described better the algorithm in [105]. This method is a function that returns a set of models that best predict the result. BSWiMS is part of the FRESA.CAD R package and it is a supervised model-selection method aimed to select a unique statistical model that predicts a user-specified outcome, in this case, a survival outcome. Finally, in 2017, C. Wen et al. described an R package for the best subset selection (BeSS) for different problems, one of them was the Cox regression model. This method uses an efficient active set algorithm to choose the best possible Cox model through three different algorithms that will be detailed later in this thesis.

Because all this awareness already generated, it is important to continue with the investigation on the survival association for this kind of disease. This thesis will focus on the study and analysis of machine learning-based survival analysis applied to different clinical challenges. By using simulated and clinical data, specifically, Alzheimer's dementia, breast cancer, and osteoarthritis information we are going to test different ML-based survival strategies through the development of a Benchmarking Method. The study of different kinds of patient's information such as imaging data (x-rays, mammograms, magnetic resonance imaging, positron emission tomography) and clinical data, will allow us to show which features are more related to a certain event on the prognosis of the disease. This leads to used those features to build different Cox Models that can predict the hazard of each patient depending on his condition. Culminating in the construction of inferences that allows doctors and patients to have more knowledge about the evolution of the disease. It is expected that with this result exists a contribution in the knowledge of the chronic-degenerative diseases which lets to take better advantage of the information already generated for the diagnosis; and mainly, generate a good benchmarking framework for the ML-powered survival analysis tasks.

This first chapter presents an introduction to the problem that this thesis tries to solve. As well as making clear the limits of research and the final objectives of the work. In the first section, 1.1 we will find the description of the problem and the main motivation for finding a solution. Then the hypothesis and objectives will be described in section 1.3 and 1.2 respectively. Subsequently, a conceptualized description of the solution will be presented in section 1.4; The main contributions of the research will be described in section 1.5. Finally, a description of the structure of this document is presented in section 1.6.

## 1.1 Problem definition and motivation

Chronic degenerative diseases such as breast cancer, osteoarthritis, and Alzheimer's dementia are relevant to public expenditure on the medical field. Besides, they are diseases with high mortality and incidence rates, which means that their treatment and research should be a priority. Especially in countries with health problems and organizations that have to do with this issue [21, 82]. This is the case of Mexico, a country that suffers from health problems (as detailed in the Introduction). Those facts introduced the need of this country to find new alternatives to treat chronic degenerative diseases and how the government efforts have reached

the point of affecting its political decisions and health-care policies. These changes produced some new ways to tackle the problem, which helped in the acquisition of new information about the disease behavior [2]. In contrast, one of the principal concerns of patients, when they acquire a disease, is how the disease will evolve and how their health status will be in the future. Which in some cases is totally unknown. The lack of awareness generates discomfort in the patient, which increases his desire to know what comes in the future. They want to know what is the recovery time, the consequences of suffering from the disease, or simply knowing if they will survive. All those questions can be resumed in the medical term, prognosis. Prognosis is known as the behavior or evolution of the disease. It has a direct relationship with the diagnosis and, consequently, with the treatment. Besides, it allows an idea of the future by describing the likely course that the disease will have in each particular patient [66]. Precisely here, is where the importance of the prognosis and the survival exploration of the disease lies. Each patient can follow a different path that could result in particular treatments.

The success of personalized medicine depends on having an accurate diagnosis that permits the doctor to distinguish which therapies or treatments will benefit the patients in a better way or to know which therapy works have a better chance of a good response [44]. Nowadays, personalized medicine lets us obtain vast information about the patient. That information is combined with all the screening information and creates a big dataset for each subject. In that context, and considering what was described in the last section, there is a major interest in studying the behavior of chronic-degenerative diseases. Especially because of all the knowledge that is attached to the treatment and diagnosis process. This new knowledge could lead us to find new solutions and help the efforts of personalized medicine. There is research on each of the diseases, but not all of them focus on the study of disease survival. Most of the efforts have been made in the area of diagnosis. It is for this reason that it was possible to have a great development in diagnostic tools for an early or more precise knowledge about the health condition.

Some studies already found that the information used to find a better way to predict the behavior of the disease (survival analysis) is generated from data that doctors have access to [10, 49, 115]. Clinical information combined with imaging data created the datasets that are used to diagnose the disease in all cases. This information could also be used to found derived information that is more informative. In some cases, this information is obtained through processes that, in general, exceed the budget of people who suffer from the disease or generate excessive spending in health sector organizations. This is even more complicated when you need this information in places where diseases such as breast cancer have a greater impact; this for being low-income and middle-income countries. In these places, the mortality and incidence rates are higher than in other countries. Mexico as an example of a middle-income country has a statistic in which it is mentioned that, of all the people who suffer from BRCA, 58% belong to this economic group [21]. What causes the methods used in the techniques already investigated, are more difficult to achieve or not focused on the population with higher incidence rates. An alternative to the information is to use all the information that is already commonly generated at the time of making a medical history to diagnose a disease. Among these data, diagnostic images are found; which are the first to give information to doctors about the condition of their patient. These images are currently much more studied due to the

high possibility of working with larger datasets and also because of the evolution of pattern recognition algorithms. Known as Radiomics, this science allows the use of a digital image as data that will be mined to find relevant information. The great acceptance of these analyses and the use of developments in this field have generated very good results in the last decade [41]. One of the examples is Wibner et al. which showed what Haralick texture analysis has the potential to enable differentiation of cancerous from noncancerous prostate tissue through the analysis of the information given by images generated in magnetic resonances [119]. This type of research is made with magnetic resonance (MRI), but regrettably, they are not accessible either. That is why although MRI images are used because of its precision, it does not imply that the same development and application made on those images, cannot be done on radiographs and mammograms. In consequence, it is better to apply the knowledge and development of this science to images of more accessible costs such as those previously mentioned.

In other cases, there are countries and institutions with enough capital and interest to invest in medical investigations. This has led to the existence of initiatives that collect information about patients, the data that allow knowing about their status at the time of being diagnosed with the disease, and the follow-up information of the subsequent visits at the observation of treatment stage. These data in some cases have been released for free study. One of them is the osteoarthritis initiative (OAI). OAI is a public-private partnership jointly sponsored by government institutions led by the National Institutes of Health (NIH) and the pharmaceutical industry. The main objective of the initiative is the identification of the most related biomarkers of development and progression of symptomatic knee OA [83]. Another organization is the Alzheimer's Disease Neuroimaging Initiative (ADNI). ADNI began in 2003 as a public-private partnership too and is led by Principal Investigator Michael W. Weiner, MD. The main objective of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD) [84]. This type of database allows much easier research and exploration of different techniques to find information derived from the current data. Considering that several techniques have already been tested in these diseases, this study tries to concentrate on the field (survival analysis) in which, to our knowledge, has not been explored extensively. Depending on the disease, several databases were selected that have the necessary information to produce new results.

There are already research focused on trying to solve these problems and three of them are similar to the research that will be carried out on this thesis. However, there are still spaces in which important knowledge can be added. On the side of breast cancer, it is found that there is some work with the same type of information that will be used in this study. Rodriguez-Rojas et al. [93] used feature selection techniques to find features in the segmented portion of the image obtained from the mammography screening process; to find characteristics in high-risk cases in 2013. In 2018, Tamez-Peña et al. [105] extracts features from mammography images to find possible correlations with clinical molecular signatures in breast cancer and later with the use of multivariate analysis under stringent cross-validation to train models predicting recurrence scores. And although this study was done to predict the risk of recurrence with mammography; By not doing the study with the real data of survival of the patients (hard

test), it leaves an area of opportunity that will be taken advantage of in this approach. On the other hand, in the osteoarthritis field, there is also research but this time to a lesser extent. Ashinsky et al. [10] investigated the early prediction of osteoarthritis through the classification of characteristics in magnetic resonance imaging, using machine learning. This article makes use of the data available in OAI; however, the study uses only the characteristics of magnetic resonance images and not the x-rays information. On the other hand, the relationship between the characteristics and diagnosis is carried out; instead of the survival study of it. In Alzheimer's Dementia field, there is also some research. Ke Liu et al. [63] presented an article with risk factors on the MCI conversion by combining Independent component analysis and the multivariate Cox proportional hazards regression model with the information of the ADNI database. In this case, this study showed a similar strategy with the Cox model, but in this work, the model will be building and selected by machine learning techniques.

Considering the clinical importance and the concern generated by each of these diseases, the amount of information available about the patients who suffer from them and the development of machine learning techniques for the survival study, The problem that faces is there is no consistent technique that relates the patient's information to the risk of happening event in each disease, and although there are machine learning techniques that allow the construction of survival models with Cox regression, they have not been used to explore this type of information. That is why on this thesis, we propose an exploration of different machine learning techniques to perform the analysis of survival in three diseases, in which there is enough information to be able to relate characteristics to the risk of an event happening when suffering from the diseases. This exploration, together with a fair comparison between each technique, will allow us to find a model that shows which patient features are more related to the event. The benchmarking approach to the solution is very informative and will allow generating conclusions that together with more medical studies could have great clinical importance and therefore, help the decisions that can be made when selecting treatments for different diseases.

## 1.2 Objectives

### 1.2.1 Main objective

The main objective of this thesis is the evaluation of several machine learning techniques used in the context of survival analysis. Through a fair Benchmarking evaluation between Cox survival models enhanced by these strategies. These models have to be analyzed with diverse data sets, especially those that are part of clinical data of chronic degenerative diseases. The Benchmarking method will provide significant information for the clinical evaluation of the prognosis of chronic degenerative diseases. The information that will be used to compare the strategies come from diagnostics images, Radiomics, Genomics, clinical information and some other patient data regarding disease screening which are already part of the normal clinical practice. The Benchmarking process will consider several statistics calculated from the models generated to analyze the behavior of each of the strategies in the context of survival analysis. The comparison of these statistics, selected characteristics, execution times, among others, will allow us to find the machine learning method of the selected approaches in our

study, which is better suited to different clinical situations or simulated in survival-based models. The information and statistics reported are expected to add relevant information in the context of disease survival and provide the doctor and patient with relevant information to make decisions about how the treatment will be. On the other hand, the results are also expected to justify the generation of a fair method for the evaluation of survival models. In this context, the objectives that must be met to achieve the main objective are:

### 1.2.2 Particular objectives

- To implement a computer-based comparison tool that allows finding a survival model and report statistics for each of the algorithms. The tool must summarize the information resulting from each model and return tables and graphs that facilitate the researcher's work.
- To condition and prepare the data of different patients belonging to different the different studies or initiatives selected for this study. This prepared data will form a database ready for the application of diverse machine learning techniques.
- To analyze and compare the ability of each of the studied algorithms to select characteristics and build the survival model in different clinical cases. Subsequently, to know on which situations the selected algorithms work better. The comparisons have to use the results of the Benchmarking process can be compared to previous works and the ground truth (data simulation).
- To compare the selected techniques in computational terms, taking into account the times for the construction of the model and the complexity of each of the algorithms.

## 1.3 Hypothesis

Taking into account the issues already discussed on the problem and knowledge about the learning generated in the previous investigation, it is important to propose new directions of understanding and examining the data of chronic degenerative diseases through the use of technology development to improve strategies already used. That is why, the hypothesis of this thesis is defined as: Several machine learning algorithms allow the construction of reliable survival Cox models that allowed study chronic degenerative diseases; through the selection of clinical features such as mammograms, X-rays, MRI and PET, forms and clinical data about the patient. The reported results and the comparison of just the results will allow the researcher to find relevant information that subsequently inspires the creation of new knowledge in the context of the expected outcome for each disease. This derived information will help the medical area to have additional knowledge about the patient's disease, which will help in future clinical decisions.

With this hypothesis the research questions to answer are:

- What is the concordance index produced by the characteristics extracted in the diseases so that the results of this index can influence medical decisions reliably?



- Among BSWiMS, GSPDAS, SPDAS, LASSO, RIDGE and ELASTICNET which are algorithms used in this thesis, what alternative will result in the most effective model to predict risk in an early event? Which model will achieve a better separation between high and low-risk patients? What metric is the one that allows determining which model is on the other?
- Of the models mentioned in the previous question, what kind of problem does the algorithm work best? Of the selected machine learning methods, which one selects the most reported and known characteristics in the literature?
- Is this type of analysis really accurate enough to improve common procedures?

## 1.4 Solution Overview

As described, there is a very important problem in the treatment of chronic degenerative diseases. Despite all the efforts to know, treat them and especially to prevent them, the number of people who suffer from it and the prevalence rate per year continue rising. Some new techniques are already described and others are still in development for each disease. Considering this, it is natural, that with the development of technology, they will find themselves more and more and it will be easier to look for more alternatives that help control the mentioned diseases. That is why the technology already developed has allowed collecting a huge amount of information about patients since such information can be considered to draw conclusions that help in medical practice.

Unfortunately, each advantage comes with a disadvantage, and despite being able to know more information about the diseases, access to it, either for study or only for treatment, is complicated by different issues. However, some institutions and initiatives have been dedicated to investing in the health area to collect information that can be studied and, in turn, can help these patients with the treatment of their respective diseases. These initiatives make the use of such information and let us take advantage of the opening of different sources and in various diseases a lot easier. This thesis seeks to combine the development of computational technology, coupled with statistical analysis, for the study of information on patients with chronic degenerative diseases. In other words, the statistical analysis of survival powered by machine learning techniques tries to study the relationship of medical measures, extracted from various imaging sources of patients suffering from chronic degenerative diseases, with the possibility of suffering an event in the context of same disease. That is, for example in the case of Alzheimer's disease, it will be sought to find such a relationship between the patient's measurements and the time of change of Mild cognitive impairment towards Alzheimer's dementia, what also will help obtaining information on the rate of conversion for each characteristic. This study bases its main contribution to the combination of two robust tools for data analysis. On the one hand, the statistical analysis of survival, which can provide clear and concise information that could help clinically for medical decision making. And on the other hand, the implementation and use of machine learning techniques that allow the analysis of large amounts of data, improving the production capacity of statistical models that provide adequate information. Specifically, this thesis will focus on three diseases and will use 4 different machine learning techniques for analysis. The comparison and evaluation of each

one of the techniques will allow discussing the results that in the future could have clinical relevance.

## 1.5 Main Contributions

In recent years, Machine learning has provided researchers numerous tools to explore complex models of associations between survival outcomes and clinical features or biomarkers. These tools are powerfully and complex at the same time, hence a systematic way to explore them is required to understand their potential and application on clinical and survival studies. Embedded statistical learning like L1 penalization (CoxNet), wrappers model selection (BSWiMS), and Best Subset Selection (BeSS) are among the machine learning frameworks available to researchers. Survival analysis based on multivariate Cox regression has a great potential to enhance diagnosis and understanding of the diseases, but current studies have been limited to small cohorts and a small set of imaging biomarkers.

Subsequently, considering these fields as research opportunities and, therefore, giving major importance to possible new knowledge, the main contribution of this thesis is the "CoxBenchmarking" implementation. CoxBenchmarking is an open source and free computational tool for the comparison of Survival Models constructed through 11 machine learning algorithms based on the Proportional Hazards model. This contribution also carried out a test of its performance with simulated information and considered the clinical exploration of chronic degenerative diseases. Specifically, it used datasets of Breast Cancer, Alzheimer's dementia, and Osteoarthritis. Each disease defined a particular survival event. Besides, this tool also provides a fair and graphic comparison of the ML methods, through the plot function that was also implemented for this thesis. Regarding the clinical contribution of this thesis, the research has already led to the publication of two relevant scientific papers in the medical field. It also stands the basis for great new research in the same context. In the next few years, the clinical importance of the results will be measured. Regarding the computing context, the implementation of a comparative evaluation technique contributed through the interpretation and unification of the results of eleven algorithms in a single model. CoxBenchmarking gives researchers the power of using several Machine Learning tools and also the ability to interpret the results through the statistics reported. Which even allows them to make the comparison and select the method that best suits the solution they are looking for.

## 1.6 Outline of the Thesis

The outline of this thesis that describes the evaluation of different machine learning approaches to build clinical-based survival models, describing the rate of suffering an event on some chronical degenerative diseases, is detailed below:

In this chapter 1, the motivation of the problem to be solved with the investigation of this thesis, is introduced and described. The objectives are described and limited and the hypothesis is detailed. In the next chapter 2, we describe the background information about survival analysis, cox model, Kaplan Meier curves, evaluation, and validation test, machine learning methods that will be used and among other topics. In chapter 3, the solution methodology of

the present investigation is described. All the methods used and the description of each data set to perform the experiments are detailed. These experiments are described in the next chapter 4, where graphical representations and details of each of the results are presented. Finally, in chapter 5, the results are discussed, conclusions are generated and the possible future work of this thesis is shown.



# Chapter 2

## Background Work

This chapter provides a context for the main topics and definitions necessary to understand the research corresponding to this thesis. For this, documents, books, and research that are relevant in the area are taken into account. First, an introduction is presented for three chronic degenerative diseases selected due to the high mortality and prevalence rate they currently have. These three diseases also have information banks available with a collection of patient data; These data openly and freely allow the study of this disease. The information of each of the data sets used will be described in the next chapter Chapter 3. The first of the diseases, although its prevalence is higher in the female population, can also be found in men. Breast cancer (BRCA) due to a large number of affected annually, means a large public expenditure and especially human losses [110]. The following two diseases have a higher prevalence in the elderly population [76, 6]; osteoarthritis is a degenerative joint disease and Alzheimer's disease is a chronic neurodegenerative disease.

After the context of a doctor who will help to understand the data that will be studied in this thesis, this chapter will describe the technique of statistical analysis that will be used in the available data, Survival Analysis. This statistical technique provides us with a set of methods to analyze data where the objective of the study is the time variable that elapses until the occurrence of an event of interest. From this statistical analysis, many terms and notions are derived that are necessary to understand the proceeding of the analysis. Within this chapter, you will find definitions of Kaplan Meier Curves, LogRank Test, Survival models. Considering the survival models, a wide range of solutions are mentioned in the chapter, but we will pay great attention to Cox Regression being one of the most used models with reasonably good estimates. Subsequently, the computational techniques that will be used are described. A set of diagrams and explanations will allow the reader to understand the operation of the algorithms used, as well as the machine learning strategies that are part of the research. As a penultimate topic, this chapter details the metrics that will be used to present a fair and valid comparison between the models generated by this research. Finally, here it is taken into consideration the techniques used to validate the results, various cross-validation strategies are described.

## 2.1 Chronic degenerative diseases

In order to carry out the research and above all to show the theoretical sustenance of the proposal, it is necessary to know about the survival of chronic-degenerative diseases, events that can occur within the diseases; the methods used to analyze data and extract features from the images. It is also necessary to know information about the methods used to calculate the survival of diseases. Below, all the theoretical bases of the present investigation are detailed.

Chronic-degenerative diseases are an extremely worrying topic in the modern age. They are currently the leading cause of death in most developed countries. Their multiple factors and diversity make them very difficult to control [27]. These types of diseases are characterized by having the following qualities:

- Multiplicity in clinical conditions, covering thousands of nosological entities
- Multiple locations of injuries
- Multistep pathogenesis
- Multifactors that generate the disease. Various types of risk factors

This kind of disease is more common on the longer-lived populations and considering the life expectancy is longer the chronic-degenerative diseases are more common [33]. Breast cancer

### 2.1.1 Breast cancer

It is the most common type of cancer in women, but men can also suffer from it. In the female sex, it affects approximately 10% of its population and in recent years the incidence has not stopped [2, 109]. Breast cancer begins when cells in the breast start to grow in ways that are not normal. These cells that behave differently, tend to form a tumor that can be detected through a mammogram or by a process known as Touch-Look Check (TLC). If the tumor is malignant, it is considered cancer and, in that case, it begins to invade all the tissues of nearby areas of the body [109].

Depending on where it is generated, it is a different type of cancer. The most common types are those that start in the mammary glands (known as lobular cancers) and those that start in the milk ducts (ductal cancer). There are other kinds of cancer, but they are less common; however, when cancer starts in other tissues inside the breast they are not considered BRCA anymore (sarcomas, lymphomas). Of the types of cancer mentioned, you can find variations that change in the way they are shown in the early stages. Many of them do not generate lumps in the breast and even get to have no symptoms. However, all of them can be detected through mammograms. This allows that the use of them can be used for research [109, 20].

#### Breast cancer screening and mammograms

This is the first step to diagnostic the disease to someone; and in cancer, mammograms are a big part of screening [20]. Regarding BRCA screening, scientists are trying to detect cancer before symptoms appear, to do that, certain methods can be used, and one of them is a mammogram.

A mammogram is an x-ray picture of the breast. When the mammogram is used to detect cancer on someone who has no signs of cancer, it is called screening mammograms. On the other hand, if the mammogram is used after a lump or other sign is present, it is called a diagnostic mammogram. The main difference between these two kinds of mammograms is the time to perform the test and the images, views, and angles that are the outcomes expected [20, 116]. Diagnostic mammograms are chosen to make an accurate diagnosis and therefore they might be helpful for prognosis prediction.

### **2.1.2 Osteoarthritis**

Osteoarthritis is a degenerative joint disease, is one of the most common chronic degenerative diseases. It primarily affects the articular cartilage and most of the time is associated with aging. OA will most likely affect the joints that have been frequently used throughout the years including fingers, hips, and knees [121] which is the one that we are going to use in the research.

Osteoarthritis takes importance when its numbers show a great incidence throughout the world. Currently, the disease has already entered the top ten of disabling diseases in the most developed countries; worldwide there is a rate of 18% of women and 9.6% in men over 60 who suffer from the disease. However, the behavior of the disease before age 45 favors women who suffer from it in smaller amounts; after this age, the percentage is reversed again. Of all these people with the disease, 80% have limitations in their movements, consequently, 25% of them cannot carry out their daily activities [121, 76]. Concentrating on the knees is a consequence of the fact that this is the joint that is most commonly affected by this disease. Symptoms may include swelling, stiffness, and pain that causes problems when walking or doing some physical activity. This kind of osteoarthritis can lead to disability [76]. To make a correct diagnosis of the disease, many doctors make use of various methods and tests on patients; including data from his past, physical examination, laboratory tests, and x-rays.

### **2.1.3 Dementia**

Dementia is a syndrome that generates deterioration in cognitive function, in other words, the ability to think [7, 122], which goes beyond the expected deterioration of normal aging. This deterioration commonly affects memory, thinking, and judgment, which affects the action of most daily tasks; however, consciousness is not affected. Dementia is one of the leading causes of disability and dependency among older people around the world [122]. The main problem is that, besides, there is a disease that causes disability, the number of people who suffer from it is gigantic. Around the world, around 50 million people have dementia and almost 10 million new cases are produced every year. And within these cases, Alzheimer's disease is the most common form of dementia and can contribute to 60-70% of cases [7, 122].

#### **Alzheimer's disease**

Alzheimer's disease is a chronic neurodegenerative disease that usually starts slowly and gradually worsens over time. And while cognitive loss is common with aging, Alzheimer's is not normal in aging [6, 7]. The greatest known risk factor is the increase in age, and most people

with Alzheimer's are 65 or older, yet Alzheimer's is not just an old-age disease. Alzheimer's disease is a progressive disease, in its early stages, memory loss is slight, but later, people lose the ability to hold a conversation and respond to their environment. These symptoms are described as dementia, and in the case of this disease, they continue until death. Alzheimer's disease has no current cure yet, the efforts to improve current treatments or seek new solutions have not been stopped. Although current treatments for Alzheimer's can not stop the progress, they can temporarily delay the deterioration of dementia symptoms [6, 7].

**Plaques and Tangles** Despite knowing a lot about the brain and all the advances that have to diagnose the disease, it is still unclear what causes this disease. Most research seems to agree that there are two proteins in the brain that are the main suspects of causing the deterioration. One is beta-amyloid, and the other one is p-tau both reaches abnormal levels in the brain of someone with Alzheimer's [72].

## 2.2 Survival Analysis

Statistical analysis known as survival analysis is a technique that lets analyze the expected duration of time until an event happens. The event could be of any kind such as battery discharging, time a lightbulb will last, or in a clinical context, the time a person that is diagnosed with Cancer, OA or Alzheimer's disease can turn into an event, such as recurrence, total knee replacement or Alzheimer's dementia. [95, 103]. In other words, we can define the term as a collection of statistical procedures for the analysis of data for which the variable of interest is the time until an event occurs. The time measure could be any time unit, such as day, month, week. It is usually called **survival time**. And by event, usually called **failure**, every activity that may happen to an individual or a thing could be considered.

In survival analysis, there are some terms that are the main base on how it works. Such terms are very commonly used so they will be defined as follows:

- **Censored event:** If a subject does not have an event during the observation time, they are described as censored. Nothing is known about the subject, but neither is it known whether or not it had an event at that time or after it. Nothing is known about the subject but neither is it known whether or not it had an event at that time or after it. In the case of health, it usually happens that some patients do not return to the same institution, whether due to death, change of institution or other reasons. And because you can not infer a result the event is censored [103].
- **Event:** Any type of action that may happen depending on the chosen topic. In the case of diseases, you can have: Death, disease occurrence, disease recurrence, recovery or another kind of event depending on the disease [95, 103].
- **Survival function  $S(t)$ :** The probability that the subject will survive in the given time.
- **Time:** The time in which survival will be considered. In the medical area, the time is taken from the beginning of the treatment or the diagnosis of the disease.



### 2.2.1 Hazard Function

Also known as hazard rate or force of mortality, this function is denoted by the lambda symbol ( $\lambda$ ) and is defined as the event rate at a time  $t$  given by the survival time. Returns the probability of an event happening between time  $t$  and  $dt$  [95]. In other words, The hazard function  $h(t)$  gives the immediate potential per unit time for the event to occur, given that the person or object has already survived the time  $t$ . It is denoted by  $h(t)$ , is given by the formula 2.1 [103, 95]:

$$h(t)dt = \lim_{\Delta t \rightarrow 0^+} \frac{P(t < T \leq t + \Delta t) | T > t}{\Delta t} = \frac{f(t)}{S(t)}, \quad (2.1)$$

Where  $h(t)$  equals the limit, as  $\Delta t$  approaches zero, of a probability statement about survival, divided by  $\Delta t$ , where  $\Delta t$  denotes a small interval of time. Integrating  $h(u)$  over  $(0, t)$  gives the cumulative hazard function  $H(t)$  that describes the accumulated risk up to time  $t$ .

$$H(t) = \int_0^t h(u)du. \quad (2.2)$$

The  $h(t)$  is always non-negative and has no upper bound. Different types of hazard functions could lead to different survival models. In contrast to the survival function  $S(t)$ , the hazard function is looking for an event that makes the subject fail or in other words, is looking for an event where the individual not survive. Therefore, the hazard function allows us to obtain information contrary to that provided by the inverse function that the survival function [103, 26].

### 2.2.2 Formulation

Knowing these terms is easier to understand the operation of survival analysis. And therefore, also the formulas that are postulated for its resolution [95, 26, 103]. Take  $T$  as a non-negative random variable that represents the lifetime of individuals in a population. In the case in which  $T$  is continuous, let  $F(\cdot)$  be the distribution function of  $T$  and  $f(\cdot)$  the probability density function. ( $f(t) = 0$ ).

$$F(t) = P(T \leq t) = \int_0^t f(x)dx \quad (2.3)$$

given that, the complement function that means the probability that an individual survives to time  $t$  is given by the survivor function:

$$S(t) = P(T > t) = 1 - F(t) = \int_t^{\infty} f(x)dx. \quad (2.4)$$

The survival function tends to be 0 while the age increases.

Having defined  $S(t)$ ,  $H(t)$  and  $h(t)$ . Therefore, we can get:

$$h(t) = -\frac{dS(t)/dt}{S(t)} = -\frac{d \log(S(t))}{dt}, H(t) = -\log(S(t)), S(t) = \exp(-H(t)), \quad (2.5)$$

because  $t$  is in the interval 0 to  $\infty$  the following properties have to be considered into survival function:

1. Survival function is not increasing.
2. At time 0 the  $S(t) = 1$  and in  $t = \infty$  is 0

### 2.2.3 Censoring

Censoring happens when we only know that there is some information about an individual in a given survival time, but we don't know the survival time exactly. In other words, censoring is referred to as partial observations [103] and this partial information is about a random variable of interest. There are usually three reasons concerning how censored data may occur or why censoring may happen [57]. These reasons will be based on the assumption that the survival analysis will be made on the healthcare area survival:

1. The event does not happen before the study ends.
2. A person is lost to follow-up during the period which the study is in progress
3. A person leaves the study due to death. This always and when the death is not the event of interest and for which the analysis is being developed. Or there is some other reason why the subject leaves the treatment.

There are three types of data that could be censored that will be described and also represented on the figure 2.1 [26]:

#### Censoring types

There are different types of censoring. The definition of each type is important due to the fact that different types lead to a different type of data preparation.

- (a) **Right censoring:** It occurs when the person's true survival time turns into an incomplete at the right side of the follow-up period (described on each analysis) which is occurring when the study ends or when the person is lost to follow-up or is withdrawn of the study. This kind of data is usually known as right-censored data. For these, the complete interval of the survival time is unknown for the analysis, it has been censored at the right side of the observed survival time interval. In other words, right-censored data occurs when just exists the knowledge of the variables that exist, but we do not know anything in the range of the time of survival [57, 103].
- (b) **Left censoring:** On the other side, there is another kind of censored data, Left censoring happens when you can only observe just several random variables instead of information that are inside of the time we need to complete the study. It can occur when a person's true survival time is less than or equal to the person's observed survival time. In the case of this thesis we can have left-censored data when a patient is diagnosed with osteoarthritis disease, but it is not known since when exactly the disease began in his

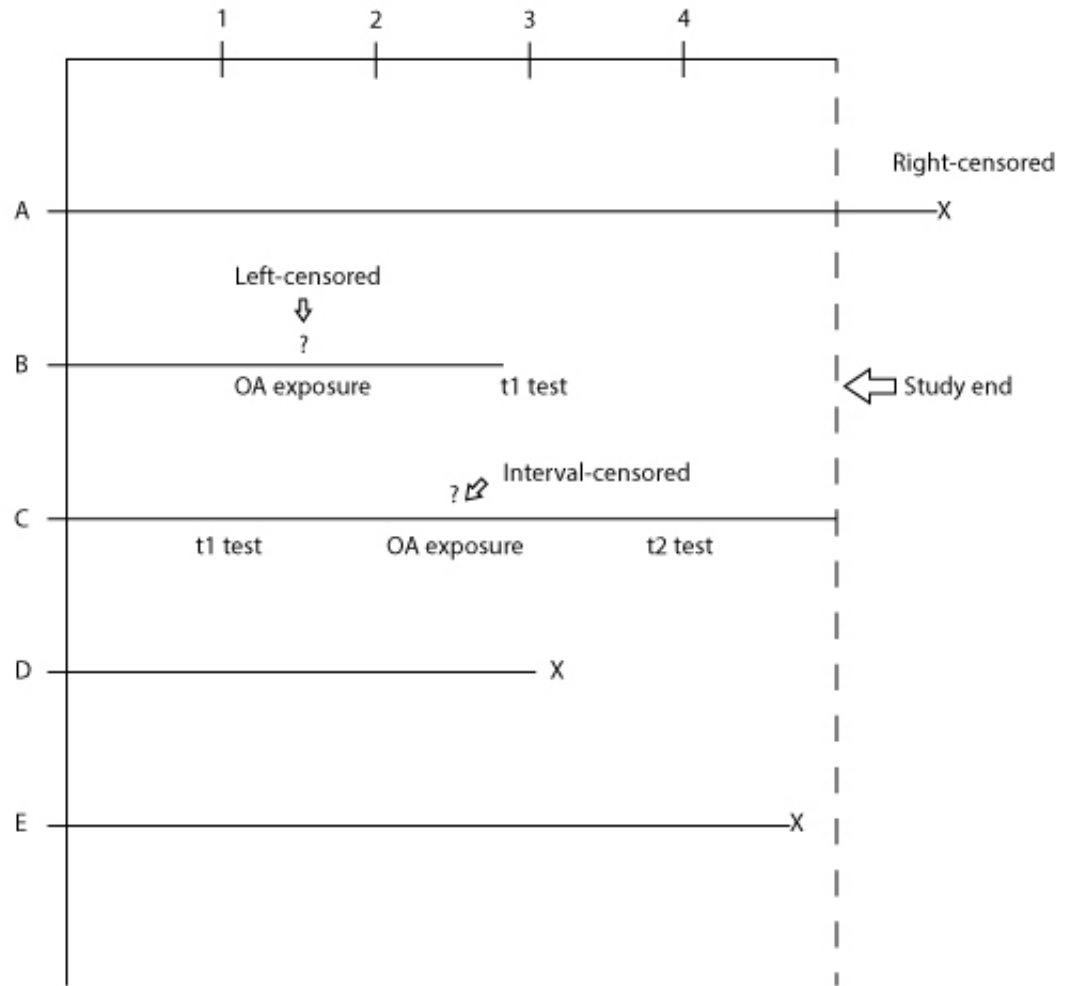


Figure 2.1: Censoring types in a study.

body, he is only known about his current situation. In other words, if a patient is left-censored at time  $t$ , we know they had an event between time 0 and  $t$ , but we do not know the exact time of the aforementioned event.

- (c) Interval censored: Survival analysis data can also be interval-censored, this type of censoring can occur if a person's true and unobserved survival time  $t$  is inside of a fixed interval of time. As an example in this thesis data there are some cases where a patient is in a constant control of their diseases. Interval-censoring actually incorporates both right-censoring and left-censoring as special cases. The data censored on the left occurs when the value of  $t_1$  is 0 and  $t_2$  is a known upper limit on the actual survival time. In contrast, the data censored on the right occur when the value of  $t_2$  is infinite and  $t_1$  is a known lower limit on the actual survival time.

## 2.2.4 Approaches to Survival Models

Depending on the hazard function complexity, different survival models could be constructed. If the  $h(t)$  is constant, then the survival model is exponential. If the hazard function increases over time, the model will be an increasing Weibull model. On the contrary, if it decreases over time the model will be a decreasing Weibull. There are other times that the hazard function increases and later decreases; for those cases, the model will be a lognormal survival model. In all these cases, it is assumed that the survival models follow a known distribution, as the ones mentioned. When a known distribution is assumed the model is called **parametric survival model** [57].

Distribution	Survival function	Hazard function
Exponential	$exp(-\lambda t)$	$\lambda$
Weibull	$exp(-\lambda t^p)$	$\lambda p t^{p-1}$
Log-logistic	$\frac{1}{1+\lambda t^p}$	$\frac{\lambda p t^{p-1}}{1+\lambda t^p}$

Table 2.1: Kinds of survival models

Three are the most common distributions on survival models. In table 2.1, we show the survival and hazard functions of those commonly used distributions. The Exponential distributions as was mentioned before, have a constant hazard function, represent by the  $\lambda$  symbol.

### Exponential and Weibull models

**Exponential model** It is the simplest parametric survival model. If the hazard functions are constant, then the distribution is exponential. The hazard function is represented by the symbol  $\lambda$ . In the case of an exponential model, if the base risk (risk in  $t_0$ ) is a constant and the risk value is doubled or tripled, the new risk remains constant but with a greater value. On the other hand, if the risk changes two or three times faster, the new risk doubles or triples, but is constant over time, so we remain in the exponential family [92].

The baseline risk is constant over time  $\lambda_0(t) = \lambda_0$ . Therefore, the exponential risk function based on a set of  $i$  variables on  $x$

$$\lambda_i(t, x_i) = \lambda_0 exp(x_i \beta). \quad (2.6)$$

**Weibull model** The second most common distribution is Weibull distribution. It is the most widely used parametric survival model. It uses two parameters  $\lambda$  and  $p$ . In this case, Weibull reduces by the exponential if  $p = 1$ .  $p$  is the shape parameter and it determines the shape of the hazard function. If  $p > 1$  the hazard increases with the time. If  $p = 1$  the Weibull model turns into an exponential model. And if  $p < 1$  then the hazard decreases over time. The new parameter gives the model the flexibility that the exponential model does not have,

Using both parameters the survival function on a Weibull distribution is given by

$$S(t) = exp(-(\lambda t)^p). \quad (2.7)$$

Hence, the hazard function is:

$$\lambda(t) = p\lambda(\lambda t)^{p-1}. \quad (2.8)$$

### Log Logistic

Log-logistic is a parametric model in which the risk rate initially increases and then decreases and, sometimes, can be hump-shaped [4]. It is defined by the following equations:

$$S(t) = \frac{1}{1 + \lambda t^p}, \quad (2.9)$$

and the hazard function

$$\lambda(t) = \frac{\lambda p t^{p-1}}{1 + \lambda t^p}. \quad (2.10)$$

Log logistics is a parametric model in which the risk rate initially increases and then decreases and, sometimes, can be hump-shaped [4]. It is defined by the following equations:

## 2.3 Kaplan Meier

The Kaplan Meier (KM) curves are an alternative representation of survival analysis data. The basis of it is part of a layout representation of the information that goes like this: Taking into account the information organized in a table, the first column in the table would have the information referring to the survival times, ordered from lowest to highest. The second column denotes the frequency of failures in each different failure time. The third column provides frequency counts of those people censored in the time interval that begins with the time of failure until the next time of failure, but without including it. The last column provides the set of risks, which denotes the collection of individuals who have survived at least the corresponding time.

To estimate the probability of survival in time  $t$ , we use the risk of that moment to include the information we have about a person censored up to the time of censorship, instead of simply ignoring that information. That survival probability is calculated with the Kaplan Meier method.

### Kaplan Meier Estimator

The Kaplan–Meier (KM) estimator is also known as the product-limit estimator (PL estimator), it is a non-parametric statistic used to estimate the survival function from the collection of the life data of a particular object. Within the area in which this thesis is focused, it is often used to measure the number of patients who survive in a given period of time after treatment, where the treatment and the measuring process is made depending on each disease. The estimator is named like that because of the authors Edward L. Kaplan and Paul Meier, who jointly produced the document that described this estimator. The estimator is given by the following formulas [95, 53, 103].

$$\widehat{S}(t) = \prod_{i: t_i \leq t} \left(1 - \frac{d_i}{n_i}\right), \quad (2.11)$$

with  $t_i$  a time when at least one event happened,  $d_i$  the number of events that happened at time  $t_i$  and  $n_i$  the individuals are known to survive (have not yet had an event or been censored) at time  $t_i$ . One of the definitions needed to understand better this estimator is *esf*. Where to understand that it is required to treat this data like it does not have any censored event. Let  $t_i$  denote an ordered observed value. The empirical survivor function (*esf*), denoted by  $S_n(t)$ , is defined to be [103].

$$S_n(t) = \frac{\# \text{ of observation } > t}{n} = \frac{\{t_i > t\}}{n}. \quad (2.12)$$

The  $S_n(t)$  is the proportion of patients still in remission after  $t$  weeks. Kaplan-Meier estimator adjusts the *esf* to reflect the presence of right-censored observations.

The KM estimator is one of the most frequently used methods of survival analysis. It is also used to the probability of death, examine recovery rate and the treatment quality. It is limited in its ability to estimate survival adjusted for covariates; and that is why to solve this limitation is needed the study of other concepts as parametric survival models and the Cox proportional hazards model that will be useful to estimate covariate-adjusted survival [57, 58].

time (days)	$m_f$	$q_f$	$n_f$	$S(t_f)$
0	0	0	257	1
1000	74	47	257	0.53
2000	25	65	136	0.18
3000	7	20	46	0.07
4000	1	18	19	0

Table 2.2: Group 1 (Males) alternative ordered layout.  $m_f$  is the number of events at time  $t$ .  $q_f$  number of censored subjects at time  $t$ .  $n_f$  set of subjects who are at risk of failure

time (days)	$m_f$	$q_f$	$n_f$	$S(t_f)$
0	0	0	185	1
1000	52	42	91	0.49
2000	24	47	20	0.11
3000	3	7	10	0.05
4000	1	9	0	0

Table 2.3: Group 2 (Females) alternative ordered layout.  $m_f$  is the number of events at time  $t$ .  $q_f$  number of censored subjects at time  $t$ .  $n_f$  set of subjects who are at risk of failure

In order to better understand the operation of the KM curves, it is first necessary to be able to visualize the tabulated data unusually, that is, not the raw data table. This new table-like visualization allows us to understand the operation and the bases under which the Kaplan

Meier curves are generated. To show this visualization, the information that will be used later will be used in one of the experiments of this thesis. On this table we found information regarding ADNI database. Information about those data will be described on the Chapter 3. By using this information we are trying to compare the survival information of two cohorts on the ADNI data, males and females with risk of Alzheimer’s disease. In the Tables 2.2, 2.3 we find the column corresponding to the information about time with a specific unit. The grouping of the other columns depends on the values of this column; so a unit of time is usually selected with which a series is started from the first value until reaching the study completion date (Ex: Week 1 - Week 52, Interval: 1 week). In this case, these times are in the unit of time: days (This is not a limitation for the display, any unit of time can be used to organize the data, depending on the convenience).

### KM Curves ADNI

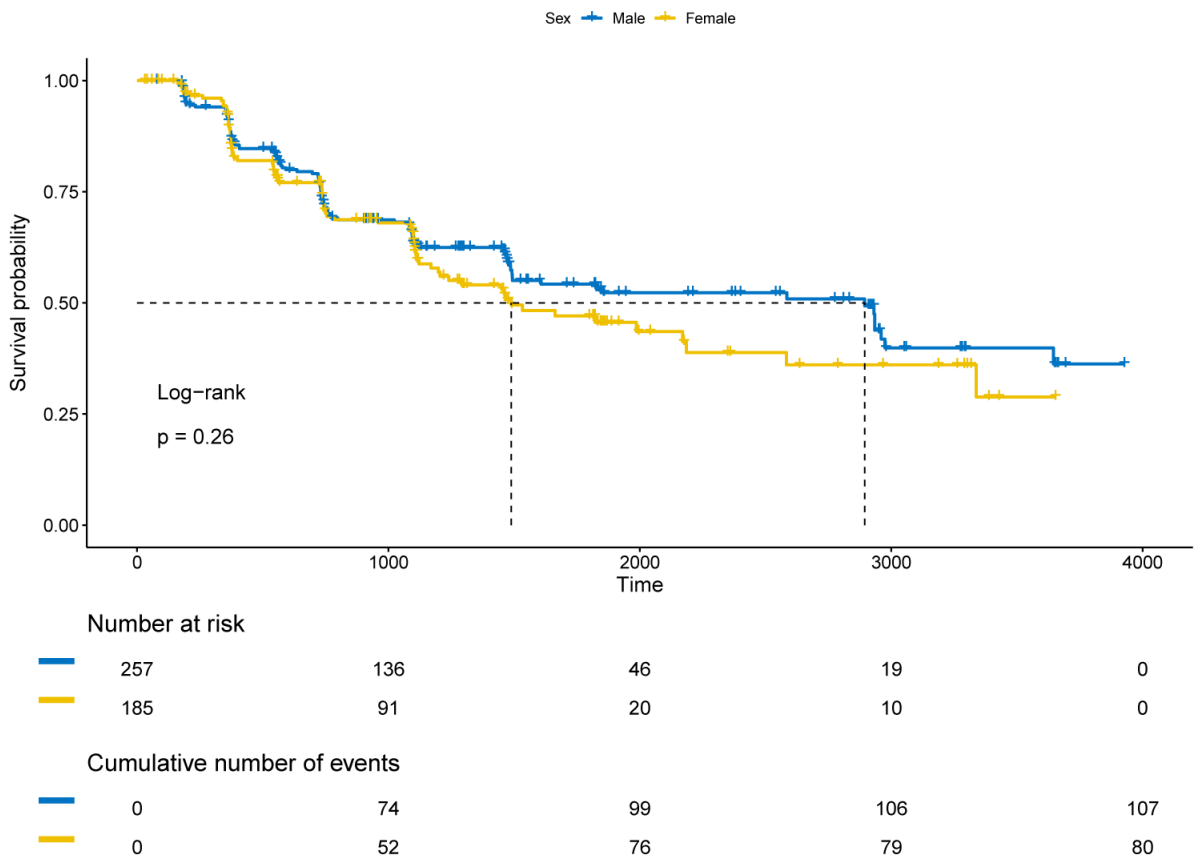


Figure 2.2: KM Curves sex-stratified between 442 Alzheimer’s disease patients

Depending on the interval and the unit selected, the number of rows that the table will have changes. Each row symbolizes a time  $t$  in which the events that happened or were censored at that time were grouped. Continuing to the right, the next column  $m_f$  shows the count of the failures within the detailed time. that is to say, the number of the events occurring up to said time  $t$ . The third column  $q_f$  denotes the frequency of people who have

been censored in the corresponding time. Finally, the following column  $R(t_f)orn_f$  shows the group of individuals who are at risk of failure. This column only counts individuals who survive in time  $t$ . These data allow the calculation of a probability of survival using the group of individuals at risk in time  $t$ . Using the conjunct of at-risk individuals aims to allow the use of censored subject information within the study and not just discard such information. This probability is calculated using the Kaplan Meier method already detailed. A plot of KM survival probability is shown in figure 2.2.

### 2.3.1 Survminer

Survminer is R package for Drawing Survival Curves using 'ggplot2'. The main purpose of the package is Survival Analysis and Visualization. Developed by Alboukadel Kassambara [54]. Using one of its methods `ggsurvplot` we can graph survival curves in a simple way. This method makes use of the `survfit` function of the survival package [112]. This function calculates an estimate of the curve of the data censored using Kaplan-Meier or Fleming-Harrington or calculates the survival function with Cox. By default, Survminer makes use of Kaplan-Meier. Different parameters can be sent to make the graph, risk data can be included in a table, information on the number of subjects that have not yet experienced the event and, above all, the most important thing is that it allows combining different curves in the same graph. This characteristic is what will allow us to have curves of high and low-risk groups for different diseases.

### 2.3.2 Log Rank Test

Once it is possible to represent the survival data in a graphic and orderly way, it is necessary to find metrics that allow comparing different behaviors in the population groups. To evaluate when two KM curves (described above) are statistically equivalent or not, the Log-rank Test metric is used. Conversely, the Log-rank test is a long subset of chi-square tests that provides, through the test criteria, an overall comparison of the KM curves. The method makes use of the observed data against the expected values in the counts on the outcome categories. Once again, using ADNI data will exemplify the use of this technique for the calculation of the comparison metric.

times (days)	$n_{1f}$	$n_{2f}$	$m_{1f}$	$m_{2f}$	$e_{1f}$	$e_{2f}$
0	257	185	0	0	0	0
1000	257	185	74	52	73.26	52.43
2000	136	91	25	24	29.36	21.39
3000	46	20	7	3	6.97	3.02
4000	19	10	1	1	1.31	0.80

Table 2.4: Tables 2.2 and 2.3 combined. Number one 1 in the underscore section of the columns' names denotes group 1 (males) and the number two 2 (females).  $e$  is the expected value  $t$

In this case, two more columns were added to the results of the tables that were used for



KM. Cell counts with the expected value for each group were added. The formula for these values is shown in equations 2.13 and 2.14

$$e_{1f} = \left( \frac{n_{1f}}{n_{1f} + n_{2f}} \right) \times (m_{1f} + m_{2f}), \quad (2.13)$$

$$e_{2f} = \left( \frac{n_{2f}}{n_{1f} + n_{2f}} \right) \times (m_{1f} + m_{2f}). \quad (2.14)$$

The first part of the equation represents the proportion in the risk set and the second part is the number of failures over the both groups.

times (days)	$m_{1f}$	$m_{2f}$	$e_{1f}$	$e_{2f}$	$m_{1f} - e_{1f}$	$m_{2f} - e_{2f}$
0	0	0	0	0	0	0
1000	74	52	73.26244	52.42885	0.737557	-0.42885
2000	25	24	29.35683	21.38974	-4.35683	2.610258
3000	7	3	6.969697	3.02112	0.030303	-0.02112
4000	1	1	1.310345	0.796671	-0.31034	0.203329
TOTALS	107	80			-3.89931	2.363617

Table 2.5: Table with expected values and a column with observed minus expected values

When two or more KM curves are compared, the LogRank test statistic is formed using the sum of the difference between the observed values and the expected values calculated in the table 2.5. In this example, this sum is  $-3.89931$  for group 1 and  $2.363617$  for group 2. We will use the value of group 2 to perform the test.

The Log-rank statistic is computed by dividing the square of the observed minus expected values of one group by the variance of the subtraction of both groups. The equation is show in 2.15. The variance is calculated with the following equation 2.16.

$$Log - rank = \frac{(O_2 - E_2)^2}{Var(O_2 - E_2)}, \quad (2.15)$$

$$Var(O_i - E_i) = \sum \frac{n_{1f}n_{2f}(m_{1f} + m_{2f})(n_{1f} + n_{2f} - m_{1f} - m_{2f})}{(n_{1f} + n_{2f})^2(n_{1f} + n_{2f} - 1)}. \quad (2.16)$$

This test has the null hypothesis that there is no difference between the two survival curves. Under this null hypothesis, the log-rank statistic is approximately chi-square with a degree of freedom. Therefore, a P value for the log-rank test is determined from tables of the chi-square distribution.

## 2.4 Cox Model

This strategy is one of the best-known models for analyzing failure time regression data [11] and it is the most commonly used mathematical modeling technique for estimating survival curves when considering some descriptive variables simultaneously [57, 18]. The Cox Proportional Hazards (CoxPH) was described by Cox in 1972 [24]. CoxPH is essentially a regression model commonly used in the statistical area of medical research to find the association between patient survival time and one or more predictor variables, also allowing the estimation of the hazard (risk) of an event for an individual or prognostic variable [11, 114]. Using CoxPH's main objective is to evaluate the simultaneous effect of several characteristics on the survival of an object against an event. In other words, it allows us to examine how specific factors influence the event rate (e.g. surgery, death, change of medical condition) at a specific time  $t$ . This rate is commonly known as the risk rate, already defined in the previous section 2.2. The Cox model is expressed by the danger function denoted by  $h(t)$ . This function is interpreted as the risk of the event occurring at time  $t$  2.17.

$$h(t) = h_0(t) \times e^{\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}, \quad (2.17)$$

where the hazard function, denoted as  $h(t)$ , is dependent on a set of  $p$  covariates ( $x_1, x_2, \dots, x_p$ ), whose impact is measured by the size of the coefficients represented by the letter  $\beta$  ( $\beta_1, \beta_2, \dots, \beta_p$ ). The coefficients give the proportional change that is found in the covariates. The term  $h_0(t)$  is called the baseline hazard and is the value of the hazard function if all the  $x_i$  are equal to zero, which causes that the exponential 1; now the value  $h(t) = h_0(t)$  that is why it is called baseline function. This function is an unspecified function which makes the Cox model semiparametric [57]. Cox is considered and also called the proportional hazards model, because it assumes that the hazard of the event is constant, i.e. the hazard will remain in the time. If some patient has a risk of event two times greater than another patient, that proportion will remain two times greater in all the times later. This assumption is considered in the equation 2.17, on which the baseline hazard  $h_0(t)$  does not involve any covariate and the second term in the exponential expression does not involve the time  $t$  [57]. The coefficients  $\beta$  are estimated by Maximum Likelihood [11, 98].

**Hazard Ratio** The Cox Model also gives more information about each covariate, one of them is **hazard ration HR**. HR is given by:  $e^{\beta_i}$  [0 · · · p], gives an estimated hazard ratio (HR) for the effect of each variable adjusted for the other variables in a model. A value of  $b_i$  greater than zero will produce an HR greater than one, this result indicates that as the value of the  $i$ th covariate increases the risk increases and the event time decreases. In the opposite case, a value of  $b_i$  less than zero produces an HR less than one, which means that the covariate is inversely related to risk.

$\beta = 0$	$HR = 1$	No effect
$\beta < 0$	$HR < 1$	Reduces $h(t)$
$\beta > 0$	$HR > 1$	Increase $h(t)$

Table 2.6: HR value and  $\beta$  coefficient effect in the Hazard ratio function

Finally, it is important to understand why the Cox model is so important and why it is so commonly used. A key reason for the popularity of the Cox model is that, although the baseline hazard  $h_0$  is not specified, the estimations that can be obtained, based on the regression coefficients, are quite good and allow to generate interest risk indexes together with Survival curves adjusted for a wide variety of data situations. In other words, the Cox PH model is a robust model, so the results of the use of the model will approximate the results for the correct parametric model [57].

## 2.5 Statistical Learning and Machine Learning Methods

Statistical Learning (SL) and Machine learning (ML) approaches have solved the issues of regularization and subset selection. Embedded statistical learning like L1 regularization via LASSO, allows the exploration of multivariate models composed on hundreds of features [98]. On the other hand, subset-selection allows the exploration of realizable Cox models from hundreds of features [117]. Model selection via the Bootstrap Step-Wise Model selection (BSWiMS), and Best Subset Selection (BeSS) are among two of the machine learning options readily available to researchers. Statistical Learning (SL) and Machine learning (ML) approaches have solved the issues of regularization and subset selection. Embedded statistical learning like L1 regularization via LASSO, allows the exploration of multivariate models composed on hundreds of features [98]. On the other hand, subset-selection allows the exploration of realizable Cox models from hundreds of features [117]. Model selection via the Bootstrap Step-Wise Model selection (BSWiMS), and Best Subset Selection (BeSS) are among two of the machine learning options readily available to researchers

### 2.5.1 FRESA.CAD

The first R package available to use is Feature Selection Algorithms for Computer Aided Diagnosis (FRESA.CAD). It is available in the Comprehensive R Archive Network (CRAN) <https://CRAN.R-project.org/package=FRESA.CAD>. The main objective of this tool is to help scientists in the health-related area. Specifically, it was designed to find features not described or to build practical models for computer-aided diagnosis, hoping that the information found by this tool will be supportive of decision-making in the medical area. The package contains methods for data conditioning, data exploration, univariate filters, model building, model diagnostics and model visualization.

### BSWiMS

Bootstrapped Stage-wise Model Selection (BSWiMS). BSWiMS is part of the FRESA.CAD package in the programming language R. It is a supervised model selection method that aims to select the best possible statistical model that predicts a user-specified result. In the case of this investigation, a survival result. The statistical model is constructed by packaging a set of Cox models created by the unique set of statistically significant characteristics in terms of the model [105]. Specifically, the statistical model is constructed by packaging a set of compact linear models (model nuggets), where each model nugget is constructed using a unique set of statistically significant characteristics at the model level.

The workflow of the BSWiMS algorithm is divided into sections. As the author proposed BSWiMS is divided into 5 main stages: Univariate Filter, Bootstrapped Forward Selection, Frequency-based Forward Selection, Bootstrapped Backwards Elimination and Model Bagging. Each stage of the process is designed to select features that are statistically relevant in explaining the desired outcome while trying to keep the false discovery rate (FDR) at the desired level. The summary of this process is detailed in Figure 2.3

1. In the first stage the BSWiMS strategy univariate filters the covariates. It computes the univariate association of each feature to the outcome and by using the Benjamini-Hochberg procedure [13] it selects features that are above the desired q-value to build the models.
2. The second stage uses the user input B (number of bootstrap samples) to build a set of B linear models using a forward selection procedure. Each model tries to add more significant features until there is no more improvement.
3. The third stage is the frequency-based forward selection. To generate a single model from a set of formulas generated in the previous phase, the characteristics of bootstrapped models are ordered by the selected frequency. The forward model is built by stepwise adding the ordered features if the p-value is statistically significant.
4. The fourth stage stands for backward elimination. It uses bootstrapping in the forward model and analyzes the bootstrap distribution of each model feature. If the largest test or train p-value is not significant for a term of the formulae then the feature is removed from the model. The model that results from backward elimination is a compact linear model; Therefore, a nugget-model. All the terms of this model are statistically significant i.e each feature used in the nugget-model adds unique information, which is not redundant with the other features and each term is necessary to improve the model in a statistically significant way. Once this process is finished, the procedures from step two to the fourth are repeated, until no more models can be found or the test performance of the last model is lower than the first aggregate model.
5. The fifth stage is responsible for grouping (bagging) all the models found in a single statistical model. Bagging consists of taking the performance-weighted average of the coefficients of the nugget-model. The final result of this process is a bagged model that is the conclusion of BSWiMS procedure

The main characteristic of a BSWiMS model is that each of the characteristics selected in the final model is described from the average nugget-fitted fitness statistics and the feature selection frequency [104].

## 2.5.2 GLMNET

The second R package that we will use and where the LASSO, RIDGE and ELASTICNET strategies rely on is Regularized Generalized Linear Models (GLMNET), which is available in CRAN at <https://CRAN.R-project.org/package=glmnet>. GLMNET is a

package that aims to provide the tools to adjust a generalized linear model through penalized maximum likelihood. Inside of this package exists the implementation of the Penalized Cox Regression (CoxNet) that can be explored with different parameters and turned into three different algorithms. [98].

### Coxnet

Coxnet is the function belonging to the GLMNET package that makes use of a Cox model regularized by an elastic net penalty. Choose a small number of covariates to build an appropriate model. In this case, the strategy can be divided into two sections, the main algorithm which tries to find beta coefficients by employing a cyclical coordinate descent, and the cross-validation section which tries to find the optimal  $\lambda$  value to use in the regularization.

The cyclical coordinate descent considers the normal survival framework, described in the last section 2.4 and tries to find the  $\beta$  coefficients which maximize the partial likelihood function. By scaling the log-partial likelihood by a factor of  $2/n$  and restricting it with the elastic net penalty the problem becomes [98]:

$$\hat{\beta} = \operatorname{argmax}_{\beta} \left[ \frac{2}{n} \left( \sum_{i=1}^m x_{j(i)}^T \beta - \log(\sum_{j \in R_i} e^{x_j^T \beta}) \right) - \lambda P_a(\beta) \right], \quad (2.18)$$

$$\lambda P_a(\beta) = \lambda \left( \alpha \sum_{i=1}^p |\beta_i| + \frac{1}{2}(1 - \alpha) \sum_{i=1}^p \beta_i^2 \right),$$

$\lambda P_a(\beta)$  is the elastic net penalty which is a mixture of L1 Lasso [114] and L2 Ridge regression [48]. The main advantage of using elastic net comes by combining the robustness of the two strategies where lasso ignores the correlated predictors only by selecting one of them; and on the other hand, ridge regression finds a coefficient greater than zero for all predictors and gives equal weight to the correlated predictors. If the value of alpha is closer to one the algorithm tends to behave as lasso but only removing the extreme correlations. Changing the value of alpha could lead to different behaviors of the coordinate descent. Alpha value 1 will turn the algorithm into LASSO,  $\alpha = 0$  turn the algorithm into RIDGE regression and values between 0-1 will behave as ELASTICNET

The algorithms follow the next steps:

1. Initializes  $k$  folds. By default GLMNET uses 10 folds and the worst case is  $n$  which turn the CV process into leave-one-out cross-validation.
2. Initializes  $\tilde{\beta}$  coefficients and  $\tilde{\eta} = X\tilde{\beta}$
3. Compute Hessian of log-partial likelihood with respect of  $\tilde{\eta}$   $\ell''(\tilde{\eta})$  and the value of the function  $z(\tilde{\eta})$  which is defined by:

$$z(\tilde{\eta}) = \tilde{\eta} - \ell''(\tilde{\eta})^{-1} \ell'(\tilde{\eta},) \quad (2.19)$$

$\ell'(\tilde{\eta})$  stands for gradient of the log-partial likelihood with respect of  $\tilde{\eta}$

4. Find new  $\tilde{\beta}$  by minimizing the function

$$\frac{1}{n} \sum_{i=1}^n w(\tilde{\eta})_i (z(\tilde{\eta})_i - x_i^T \beta)^2 + \lambda P_a(\beta). \quad (2.20)$$

The diagonal of the Hessian with respect of  $\tilde{\eta}$  is denoted as  $\omega(\tilde{\eta})$

5. Update values of  $\tilde{\beta}$  and recalculate  $\tilde{\eta}$
6. Repeat steps from 2-4 until convergence of  $\beta$
7. Find the  $\lambda$  value which maximizes the goodness of fit estimate, defined by the equation:

$$CV_i(\lambda) = \ell(\beta_{-i}(\lambda)) - \ell_{-i}(\beta_{-i}(\lambda)), \quad (2.21)$$

$\ell_{-i}$  is the log-partial likelihood without the test part of the CV, and  $\beta_{-i}(\lambda)$  is the optimal  $\beta$  in the train process found by maximizing  $\ell_{-i} + \lambda \|\beta\|_1$ . Repeat the process form 1-5 with different lambda values until all the folds are used.

The algorithm of the cyclical coordinate descent section and the cross-validation section is summarized in the figure 2.4.

### 2.5.3 BeSS

BeSS (Best subset selection) is an R package available from the CRAN at <https://cran.r-project.org/package=BeSS> for Best Subset selection in linear, logistic and CoxPH models [117]. This strategy takes into consideration the subset selection problem which in simple words means the selection of a set with  $k$  out of  $p$  predictors. The number of possible combinations turns the problem into NP-hard combinatorial optimization problem.

BeSS tries to solve this problem with the Primal-dual formulation of the problem. The best subset selection problem with size  $k$  turns into the following optimization problem.

$$\min_{\beta \in \mathbb{R}} l(\beta) \quad \text{s.t.} \quad \|\beta\|_0 = k, \quad (2.22)$$

where the loss function  $l(\beta)$  is a convex function. In the case of CoxPH regression, that is the regression that we are going to use, the loss function is the partial likelihood. BeSS uses Newton-Raphson algorithm to estimate the values with the predictors in the active set. Like GLMNET, BeSS also replace the hessian matrix with its diagonal, reducing the computational complexity.

The BeSS package proposed an Active set Algorithm to solve the problem and they named it Primal-dual active set (PDAS). The best subset problem define an Active set  $A$  with  $k$  elements and its complement  $I$  with  $p - k$  elements. For a detailed description of the algorithm look at [117]. The determination of the optimal  $k$  is other problem to be solved. To confront that, BeSS proposed two strategies, the first one uses a sequential procedure and takes its name from that, Sequential primal-dual active set (SPDAS). This algorithm specifies the maximum value of  $k$  and iterates from 0 to  $k$  in the PDAS algorithm. Then, select the optimal  $k$  by comparing the model with the minimum Akaike information criterion (AIC) [3], Bayesian information criterion (BIC) [97] or Extended Bayesian information criterion (EBIC) [22]. And the second procedure is the Golden section primal-dual active set (GPDAS) which is created to avoid to run the PDAS algorithm extensively for a whole sequential list  $0 \rightarrow k$ ; the detailed explanation of this algorithm is in [117].

## 2.6 Model evaluation metrics

Once described the methods of machine learning and statistical learning that will be used, it is necessary to take into account which set of metrics will be used to be able to compare the models in a fair and adequate way. The word metric can be used in different contexts. However, in almost all of them, it is used to measure something with a specific unit. For this reason, metrics that have been accepted and tested in the literature will be taken into account in this Thesis. Two types of specific metrics will be taken into account to report, compare and study the results of this investigation. First, the ability to predict event risk through survival analysis with Cox regression; and second, the ability of the same model to classify the individuals that belong to the study. For the comparison of survival analysis, the Log-Rank Test, concordance index (c-index) metrics will be described and for the classification comparison the Receiver operating characteristic (ROC), Accuracy (ACC), Sensitivity (SEN), and Specificity (SPE) will be described. The first metrics are aimed at evaluating survival methods, which in the end are those that allow to build disease predictions. On the other hand, we will use the same models to be able to find patients with low and high risk although the main objective of this technique is not classification, with the aim of being able to compare these models with others found in the literature.

### 2.6.1 Jaccard Index

The Jaccard index allows us to compute the average similarity between the selected features. It is also known as Intersection on the Union and the Jaccard similarity coefficient was described by Paul Jaccard [51]. Jaccard index is a statistic used to measure the similarity and diversity between the selected samples of a set. It is defined as the size of the intersection divided by the joint size of the sample sets. The equation 2.23 illustrates the formula used to calculate it.

$$J = \frac{2}{(R^2 - 2R)} \sum_{i=(j+1)}^R \sum_{j=1}^{R-1} \frac{|A_i \cap A_j|}{|A_i \cup A_j|}, \quad (2.23)$$

where  $R$  is the number of elements that are part of the test set, and  $A_j$  is the set of the  $k$  selected features for of the  $j$  holdout training sample. The range of the index varies from 0 to 1, where 1 represents that the feature selection method always selects the same set of features on each repetition.

### 2.6.2 Concordance index

Within the literature it has not been possible to find a standardized metric to compare survival models that use a multivariate cox regression. However, it is well known that one of the most popular techniques for evaluating these methods is the Concordance Index (c-index) [89]. The concordance index, also known as c-index or its acronym CI, is one of the most commonly used performance measures of survival models. It is the probability of concordance between the predicted and the observed survival [89]. You can write through the following formula

$$c(D, G, f) = \frac{1}{|\epsilon|} \sum_{\epsilon_{ij}} 1_{f(x_i) < f(x_j)}, \quad (2.24)$$

where  $D$  is the training data,  $G$  is the graph of survival function,  $f$  us the function.  $G$  is composed by  $V$  vertex and  $\epsilon$  edges. With the indicator function  $1_{a < b}$ , and 0 otherwise;  $\epsilon$  denotes the number of edges in the order graph.  $f(x_i)$  is the predicted survival time for the subject  $i$  by the model  $f$ . Is because of this reason that, the concordance index can also be written explicitly as:

$$c = \frac{1}{|\epsilon|} \sum_{T_i \text{ uncensored}} \sum_{T_j > T_i} 1_{f(x_i) < f(x_j)}. \quad (2.25)$$

This index is a generalization of the area under the receiver operating characteristic curve to regression problems, since it can be applied to the variables of continuous output and consider the censorship of the data. Similar to the case of the area under the curve, the concordance index  $c = 1$  indicates perfect prediction accuracy and  $c = 0.5$  is as good as a random predictor [89].

### 2.6.3 Log Rank

This metric was already described in the section of survival analysis. Even so, we will try to summarize how to calculate this metric and what the objective is when evaluating these models in this way. Log-rank test is a long subset of chi-square tests that provides, through the test criteria, an overall comparison of the KM curves [68]. The method makes use of the observed data against the expected values in the counts on the outcome categories. The Log-rank statistic is computed by dividing the square of the observed minus expected values of one group by the variance of the subtraction of both groups. The equation is show in 2.26. The variance is calculated with the following equation 2.27.

$$Log - rank = \frac{(O_2 - E_2)^2}{Var(O_2 - E_2)}, \quad (2.26)$$

$$Var(O_i - E_i) = \sum \frac{n_{1f} n_{2f} (m_{1f} + m_{2f}) (n_{1f} + n_{2f} - m_{1f} - m_{2f})}{(n_{1f} + n_{2f})^2 (n_{1f} + n_{2f} - 1)}. \quad (2.27)$$

### 2.6.4 Classification results

Before defining the metrics to be used for the classification section, it is necessary to define the different types of results that we can have when predicting an outcome. The use of four words is essential when classifying an object and knowing what the result was. These words are true or false and positive or negative. True or false, refers to whether the classification assigned by the model is correct or not. On the other hand, positive or negative, refers to the assignment of one class or another.

Considering the ADNI data that will be used in this study, we can find patients who have the conversion status between MCI to AD. If the conversion exists (status=1), it can be said that the result is positive (P) and otherwise it is negative (N) (status=0). Using this table Table



2.7, we can define the combinations that will be the possible results of our classifier. The definitions will be stated in the survival analysis context.

#	RID	Time to event	Status	APOE
1	4	1106	0	0
2	33	1127	0	0
3	38	357	0	0
4	42	364	1	0
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
256	5007	741	0	0
257	5066	1104	0	1

Table 2.7: ADNI/TADPOLE data of male patients that will be used on one of this experiments of this thesis. Time to event is in days, status=1 represents that the patient suffered the conversion of MCI to AD

The first case occurs when the classifier identifies a patient has uncensored survival information as someone who will undergo the conversion, and in fact, the said event happened; This case is known as **True-Positive (TP)** (correctly identified). In the second case of **False-Positive (FP)** (incorrectly identified), patients who did not have conversion, that is, who maintain MCI, have been identified as patients who will suffer from AD. The next case is **True-Negative (TN)** (correctly rejected) which is people with censored time to event that was identified as not converters. Lastly, **False-Negative (FN)** (incorrectly rejected) that refers to patients that were incorrectly identified as not converters and they suffered the evolution between MCI and AD, namely the patient was uncensored.

$$TP = \left| \left( \tilde{f} \geq 0 \right) \cap \text{uncensored} \right|, \quad (2.28)$$

$$TN = \left| \left( \tilde{f} < 0 \right) \cap \text{censored} \right|. \quad (2.29)$$

### Confusion matrix

Considering a group with positive instances and negative instances of some condition. The four results can be formulated in a  $2 \times 2$  contingency table or confusion matrix. Also known as an error matrix, it is a specific table layout that allows the display of the performance of a classifier. Each row of the matrix represents the instances with its predicted class, while each column represents the instances in a real class. It is a special type of contingency table, with two dimensions one real and one prediction [100].

In context of this thesis we will be using the confusion matrix provided by the *plotROC* function of the FRESA.CAD package [104]. The confusion matrix as shown in the Figure 2.5

is a graphical representation of the number of cases that belong to each group. The size of the rectangle is determined by the number of cases in each group. The first rectangle, in the upper left, represents the **True-Positives** results. The upper right rectangle stands for the **False-Positives**. Then, the following rectangle (down left) is the region of the **False-Negatives**; and lastly, the representation of **True-Negatives**.

### 2.6.5 Accuracy

In simple words, we can refer to the Accuracy (ACC) to the fraction of correct cases [73]. In other words it is the proximity degree of the estimated measurements versus the actual value of that same measurement [16]. This term is linked to the term precision, which refers to the degree to which repeated measurements with equal conditions return to have the same result [16, 107]. Accuracy can be applied to any type of measurement, but in the case of this investigation we will apply the term accuracy as a statistical measure to evaluate the performance of classifiers. In this context, we can refer to accuracy as the proportion of true results (both TP and TN) among the total number of studied cases [73]. The formula for calculating accuracy in binary classification is shown in the equation 2.30 and the formula for precision in the same context is shown in 2.31

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{\text{Total population}}, \quad (2.30)$$

$$Precision = \frac{TP}{TP + FP}. \quad (2.31)$$

Although accuracy provides a simple way to compare the classification or measurement performances, its simplicity often allows too many interpretations; that is why it should be interpreted with considerable caution. These limitations force us to add complexity to our evaluation scheme. To do this, we will use complementary metrics that will give more information about performance and help us draw more relevant conclusions.

### 2.6.6 Sensitivity and Specificity

To face accuracy limitations, there are other metrics that give more information about the performance of the classification model. To describe these terms in simple words, Sensitivity (SEN) and Specificity (SPE) represent two kinds of accuracy; SEN is the accuracy for the positive cases and SPE for negative cases [73]. However, they go far beyond these simple words. Each measurement has its specific objective and its way of quantifying. Sensitivity and specificity are proportions, so confidence intervals can be calculated using standard methods of calculating intervals for proportions.

**Sensitivity** is a statistical measure that measures the performance of a binary classification test, better known as classification, in the statistical context. This term is widely used in the medical area and it refers to the proportion of true positives that were correctly identified in the test [5, 17]. It is also known as the true-positive rate (TPR), recall or probability of detection in machine learning. As was aforementioned, the probability is calculated by the

use of true-positive results that are the correct predictions of the presence of a condition. The formula goes as follows:

$$SEN = \frac{TP}{TP + FN} = \frac{TP}{\text{Observed positives}}. \quad (2.32)$$

Although it would seem that the important thing is to have a complete sensitivity test, you have to be careful with a  $SEN = 100\%$ . Although this can happen and be successful, it can also be a mistake because it does not take false positives into account and it can happen that a test generates 100% correct results, but also a 100% False positive rate (FPR).

**Specificity** is the the true proportion contrary to sensitivity. It is also known as the true-negative rate (TNR) and it refers to the true negatives that are correctly identified by the test. It can be calculated as shown in the equation 2.33

$$SPE = \frac{TN}{TN + FP} = \frac{TN}{\text{Observed negatives}}. \quad (2.33)$$

This measure helps to calculate the false-positive rate also known as the fall-out or probability of false alarm. It can be calculated as  $(1 - SPECIFICITY)$ .

### 2.6.7 Receiver operating characteristic

SEN and SPE, in addition to confronting the limitations of accuracy, allow us to obtain the Receiver operating characteristic (ROC). ROC Curve is basically an illustration that shows the diagnostic capability of a binary classifier system as its discrimination threshold varies. It is created by plotting the true positive rate (TPR) or sensitivity against the false positive rate (FPR) calculated with SPE, at various threshold settings [73]. ROC analysis is used as a tool to select the optimal models and discard those that are not. Regardless of specifying the context of class distribution.

FPR and TPR define the space of the illustration ROC SEN is used on the x axis and TPR on the y axis, respectively. This represents a cost-benefit comparison between TP results with lower number of RPF. The best possible prediction method would produce a point in the upper left corner or coordinate (0,1) of the ROC space, which represents 100% sensitivity (no false negatives) and 100% specificity (no false positives). Point (0,1) is also called perfect classification, which is not normal in the real world. Within the representation, there is a diagonal that divides the ROC space. The points above the diagonal represent good classification results; The points below the line represent bad results [34].

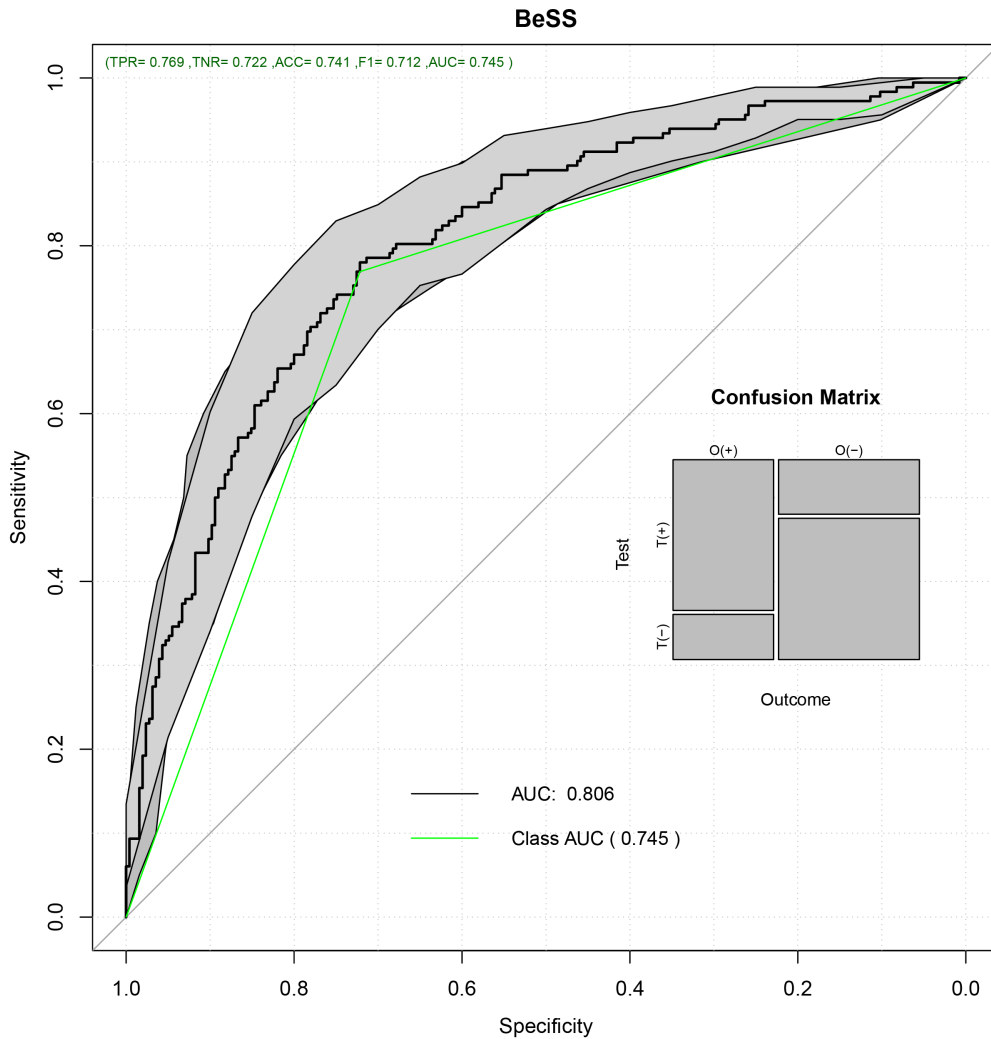


Figure 2.6: ROC Curve plotted by FRESA.CAD R package with some data classification analysis

## 2.7 Validation

### 2.7.1 Cross-Validation

There are model validation techniques, which allow evaluating how the results of a statistical analysis will be generalized to an independent data set, different from those used to generate the model, one of them, is the Cross-Validation (CV) [60, 75, 19]. The objective of a CV is to test the ability of the selected model to predict new data that was not used to estimate them, to avoid problems in the model such as overfitting or selection bias. To test the model with new data, a cross-validation iteration divides the data sample into complementary subsets. The analysis is divided into two, using the training subset first and validating the analysis in the other test set subset. The process is repeated on several occasions, making different partitions in each iteration, to eliminate the variability. Once the validation results have been calculated

with the test sets, there are methods to combine them, for example, the average. There are several two types of cross-validation, exhaustive and non-exhaustive cross-validation.

**Exhaustive Cross-validation** The first of the types of cross-validation makes use of all possible ways to divide the original sample into a test and validation set. Depending on the amount of data, the number of combinations grows by leaps and bounds. This method allows configuring the number of elements  $p$  that will be used to test the model, the name of the method is Leave- $p$ -out Cross-Validation, the rest of the observations will belong to the train set. So the strategy will be repeated until on all the possible ways to divide the data into a set of  $p$  observations are used as a training set.

**Non-exhaustive Cross-Validation** Non-exhaustive cross-validation methods do not compute all the possible combinations of splitting the original sample.

There are different ways to divide the data for cross-validation. One of them takes into account the proportions of classes within the observations. That is to say, the CV uses the process of rearranging the data to ensure each set of data has a good proportion of the whole. For example, in a binary classification problem, it tries to select the same proportion of both classes in the sets [60].

## 2.7.2 Leave-one-out Cross-Validation

Leave-one-out Cross-Validation (LOOCV) is a special case of Leave- $p$ -out Cross-Validation where  $p = 1$ . Just one observation is left to the validation set and all the  $n - 1$  observations are part of the training set. Choosing  $p = 1$  allows the computational time required to find all combinations to be shorter. The advantage of this method lies in its simplicity and the strategy of using all observations as well as tests or training. In some cases, although this type of cross-validation is much easier computationally,  $n$  remains a very large number.

---

### Algorithm 1 Leave-one-out Cross-Validation algorithm

---

```

1: procedure LOOCV(Data)
2:    $Error \leftarrow 0$ 
3:    $N \leftarrow Rows(Data)$ 
4:   for  $i \leftarrow 1, n$  do
5:      $Test \leftarrow Data[i, :]$ 
6:      $Training \leftarrow Data[-i, :]$ 
7:      $fit \leftarrow Fit(Training)$ 
8:      $Error \leftarrow Error + fit.validate(Test)$ 
9:   end for
10:   $Error = Error/N$ 
11: end procedure

```

---

### 2.7.3 $k$ -fold Cross-Validation

This method is part of the Non-exhaustive type of cross-validation. In this strategy, the original sample is randomly divided into  $k$  subsamples of equal size. Of these  $k$  subsampled sets, a single subsample is taken for the validation stage and the following  $k - 1$  remaining sets are used as training data. The process concludes when each of the  $k$  sets have been used as test sets, in total  $k$  repetitions.

The advantage of this method lies in the use of all observations for training and validation. This happens because the  $k$  sets are maintained during all iterations and only the test set changes. In literature, it is common to find cross-validation 10 [71], but in general,  $k$  remains as a non-fixed parameter. If  $k$  is equal to the number of observations  $n$ , the cross-validation of  $k$  is exactly the cross-validation of leave-one-out [46].

---

#### Algorithm 2 $k$ -fold Cross-Validation algorithm

---

```

1: procedure kFOLDCV(Data)
2:    $Error \leftarrow 0$ 
3:    $N \leftarrow Rows(Data)$ ;
4:    $SampleSize \leftarrow \lceil N/K \rceil$ 
5:    $Samples \leftarrow RandomSplit(Data, SampleSize)$ 
6:   for  $i \leftarrow 1, k$  do
7:      $Test \leftarrow Samples[i]$ 
8:      $Training \leftarrow Samples[-i]$ 
9:      $fit \leftarrow Fit(Training)$ 
10:     $Error \leftarrow Error + fit.validate(Test)$ 
11:  end for
12:   $Error = Error/N$ 
13: end procedure

```

---

### 2.7.4 Holdout and Repeated Holdout CV

This method Holdout Cross-Validation (HOCV) is considered as the simplest form of cross-validation. In the method, a set of observations are randomly selected to form two sets called test set and training set, respectively. Although the method does not have determined the amount belonging to each set; However, the most normal is that the test set is smaller than the training set [60, 9].

On the other side, Repeated Holdout Cross-Validation (RHOCV) creates  $r$  random divisions of the data set to divide them between training and validation data with a given fraction [88]. It is also known as Monte Carlo CV [31]. For each training set, a model is generated that is valid with the respective test set, the final result is calculated with the average of each of the validations performed on the test set. The advantage of this method lies in the constant number of sets that are formed, regardless of the number  $k$  used in  $k$ -fold. On the other hand, the disadvantage of this method is that some observations can never be selected in the validation subsample, while others can be selected more than once. However, this disadvantage can be treated with the number of repetitions of the method.

**Algorithm 3** Repeated Holdout Cross-Validation algorithm

---

```

1: procedure RHOCV(Data,Repetitions,TrainFraction)
2:    $Error \leftarrow 0$ 
3:    $N \leftarrow Rows(Data)$ ;
4:   for  $i \leftarrow 1, Repetitions$  do
5:      $TrainSize \leftarrow \lceil N * TrainFraction \rceil$ 
6:      $TestSize \leftarrow \lceil N * (1 - TrainFraction) \rceil$ 
7:      $Test \leftarrow RandomSample(Data, TestSize)$ 
8:      $Training \leftarrow RandomSample(Data, TrainSize)$ 
9:      $fit \leftarrow Fit(Training)$ 
10:     $Error \leftarrow Error + fit.validate(Test)$ 
11:  end for
12:   $Error = Error / N$ 
13: end procedure

```

---

## 2.8 Summary

This chapter details all the theory necessary for the compression of the methods that will be detailed in the next chapter. Regarding the main objective of this thesis, which is the evaluation of different machine learning techniques for the study of survival in different chronic degenerative diseases; it was defined which are the diseases with which, the implementation of code will be tested for fulfilling the objective of the thesis. Therefore, first in this chapter, we found the details of the problem for the three chronic degenerative diseases to be treated.

### 2.8.1 Chronic-degenerative disease

The first chronic-degenerative disease to be part of this study is the most common type of cancer in women, breast cancer. This cancer affects approximately 10% of the female population. The commonly used radiological images, mammograms; are chosen to make an accurate diagnosis and therefore they might be helpful for prognosis prediction. The second disease to study is Osteoarthritis. It is a degenerative joint disease and one of the most common chronic degenerative diseases. It takes importance because of its big incidence throughout the world. To make a correct diagnosis of the disease, many doctors make use of various methods and tests on patients; including data from his past, physical examination, laboratory tests, and x-rays. The third and last disease to study is the syndrome that generates deterioration in cognitive function, Dementia, specifically dementia caused by Alzheimer's disease. AD is the most common form of dementia and can contribute to 60-70% of cases. It has no current cure yet and the efforts to improve current treatments or seek new solutions have not been stopped. Later this chapter explains the main statistical tool used in this thesis, Survival analysis. It is a statistical analysis that let analyze the expected duration of time until an event happens and this time it will be the power that supports all the machine learning techniques. The probability that the subject will survive in the given time is given by the Survival Function  $S(t)$ . The time is delimited by the observation time. If a subject does not have an event during that observation time, they are described as censored, and its survival information is censored. The

immediate potential per unit time for the event to occur, given that the person or object has already survived the time  $t$  is given by the hazard function  $h(t)$ . Depending on the hazard function complexity, different survival models could be constructed. If the  $h(t)$  is constant, then the survival model is exponential. If the hazard function increases over time, the model will be an increasing Weibull model

## 2.8.2 Survival analysis

Next, inside this chapter is shown some survival analysis tools, such as Kaplan Meier curves, LogRank Test and the Cox Model. The Kaplan–Meier (KM) estimator is also known as the product-limit estimator (PL estimator), it is a non-parametric statistic used to estimate the survival function from the collection of the life data of a particular object. To evaluate when two KM curves (described above) are statistically equivalent or not, the Log-rank Test metric is used. Conversely, the Log-rank test is a long subset of chi-square tests that provides, through the test criteria, an overall comparison of the KM curves. The Log-rank statistic is computed by dividing the square of the observed minus expected values of one group by the variance of the subtraction of both groups. The equation is show in  $Log - rank = \frac{(O_2 - E_2)^2}{Var(O_2 - E_2)}$ . One of the most important survival models parts of the exponential family is the Cox Model. CoxPH is essentially a regression model commonly used in the statistical area of medical research to find the association between patient survival time and one or more predictor variables, also allowing the estimation of the hazard (risk) of an event for an individual or prognostic variable. The Cox model is expressed by the danger function denoted by  $h(t) = h_0(t) \times e^{\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}$

## 2.8.3 Machine Learning techniques

Finally, machine learning techniques to be used are introduced. Statistical Learning (SL) and Machine learning (ML) approaches have solved the issues of regularization and subset selection. The first R package available to use is Feature Selection Algorithms for Computer-Aided Diagnosis (FRESA.CAD). The principal algorithm to be used in this thesis is the Bootstrapped Stage-wise Model Selection (BSWiMS). It is a supervised model selection method that aims to select the best possible statistical model that predicts a user-specified result.

The second R package that we will use and where the LASSO, RIDGE and ELASTICNET strategies rely on is Regularized Generalized Linear Models (GLMNET). Inside GLMNET the principal algorithm is Coxnet which is the function that makes use of a Cox model regularized by an elastic net penalty. It chooses a small number of covariates to build an appropriate model and depending on the parameter  $\alpha$  value, the algorithm can turn into different strategies.

The third package is BeSS (Best subset selection). This strategy takes into consideration the subset selection problem which in simple words means the selection of a set with  $k$  out of  $p$  predictors. The number of possible combinations turns the problem into an NP-hard combinatorial optimization problem. BeSS tries to solve this problem with the Primal-dual formulation of the problem. The package has two main algorithms GSPDAS (default parameters) and SPDAS with GIC.



### 2.8.4 Metrics

Finally, in this chapter we take into account which set of metrics will be used to be able to compare the models in a fair and adequate way. We divided the metrics into two types, the survival status, and the classification stats.

On the survival metrics side, We use the Jaccard index which allows us to compute the average similarity between the selected features. It is also known as Intersection on the Union and the Jaccard similarity coefficient was described by Paul Jaccard. Then, the concordance index, also known as c-index or its acronym CI, is one of the most commonly used performance measures of survival models. It is the probability of concordance between the predicted and the observed survival. Once again, the Log-rank test which is a long subset of chi-square tests that provides, through the test criteria, an overall comparison of the KM curves. On the classification stats, we can summarize all the metrics to be used in the Table 2.8 and Figure 2.7.

	O(+)	O(-)	
T(+)	$TP$	$FP$	$ACC = \frac{TP+TN}{\text{Total population}}$
T(-)	$FN$	$TN$	$PRECISION = \frac{TP}{TP+FP}$
$SEN = \frac{TP}{\sum O(+)}$		$SPE = \frac{TN}{\sum O(-)}$	

Table 2.8: Review

Some terms are really important. True-Positive (TP) means correctly identified). False-Positive (FP) means incorrectly identified. True-Negative (TN) means correctly rejected and False-Negative (FN) means incorrectly rejected. Accuracy refers to the fraction of the correct cases in a classification test. Sensitivity and specificity are metrics that give more information about the performance of the classification. SEN is the accuracy of the positive cases and SPE for negative cases. SEN and SPE, in addition to confronting the limitations of accuracy, allow us to obtain the Receiver operating characteristic (ROC). ROC Curve is basically an illustration that shows the diagnostic capability of a binary classifier system as its discrimination threshold varies.

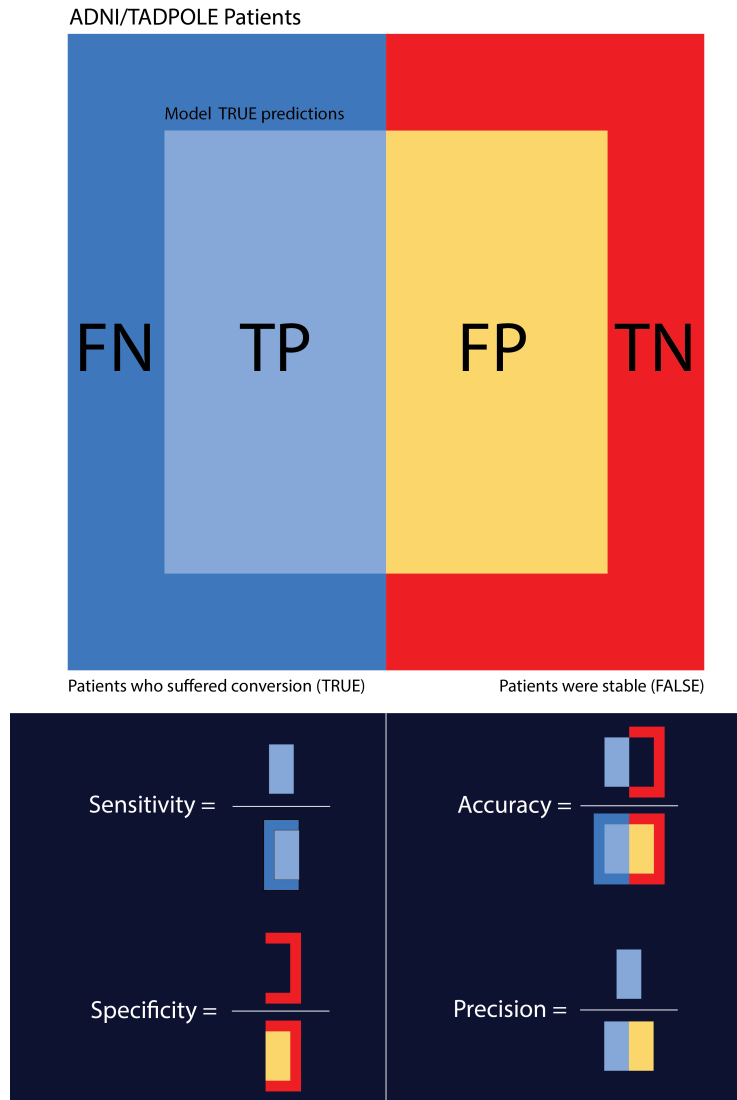


Figure 2.7: Confusion matrix plotted by FRESA.CAD R package in the ROC curve of random classifier with some data

The validation techniques are part of the Cross-validation family. Repeated Holdout Cross-Validation (RHOCV) creates  $r$  random divisions of the data set to divide them between training and validation data with a given fraction.

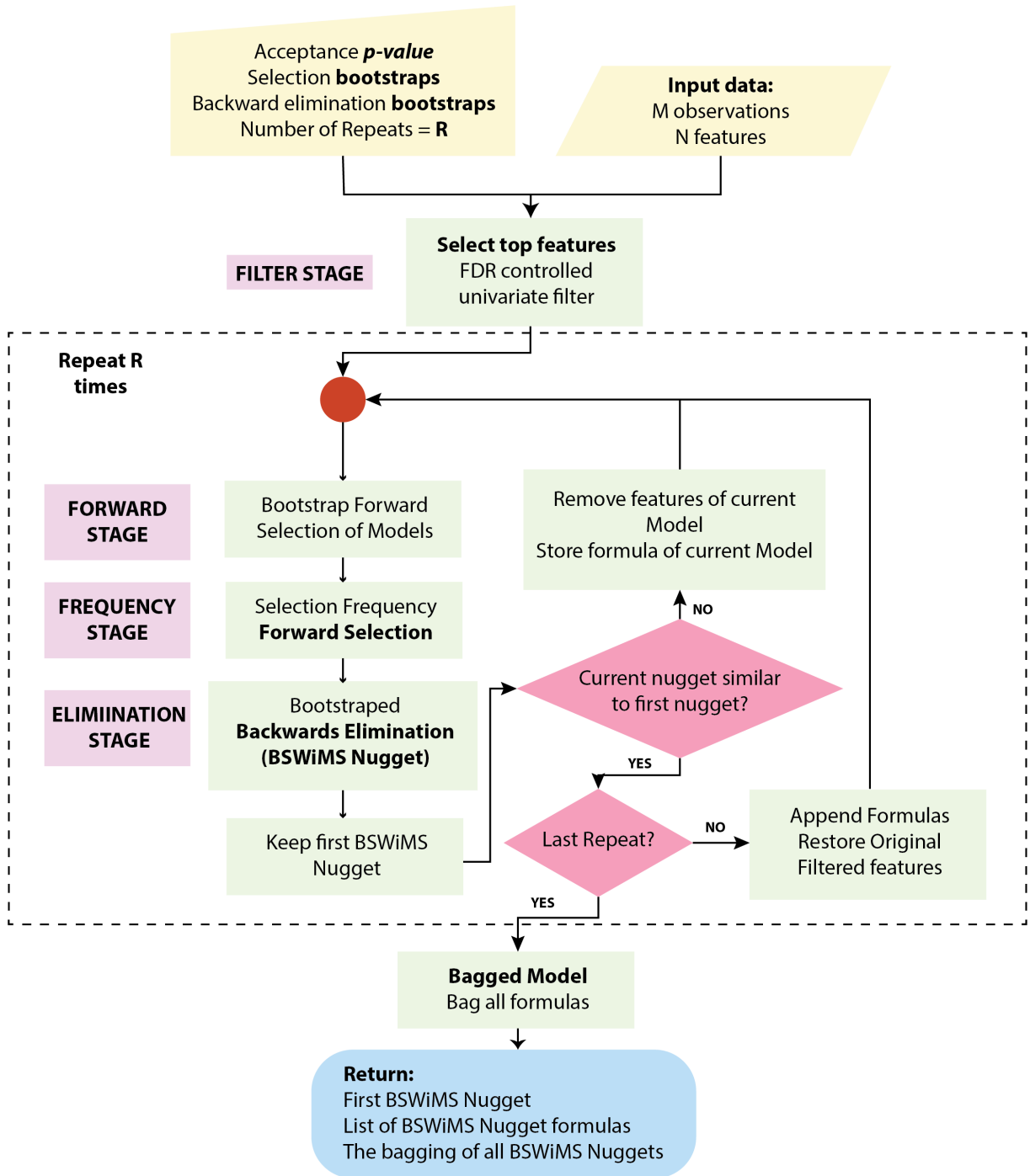


Figure 2.3: Stages of BSWiMS procedure based on Figure 2(a) in [104]

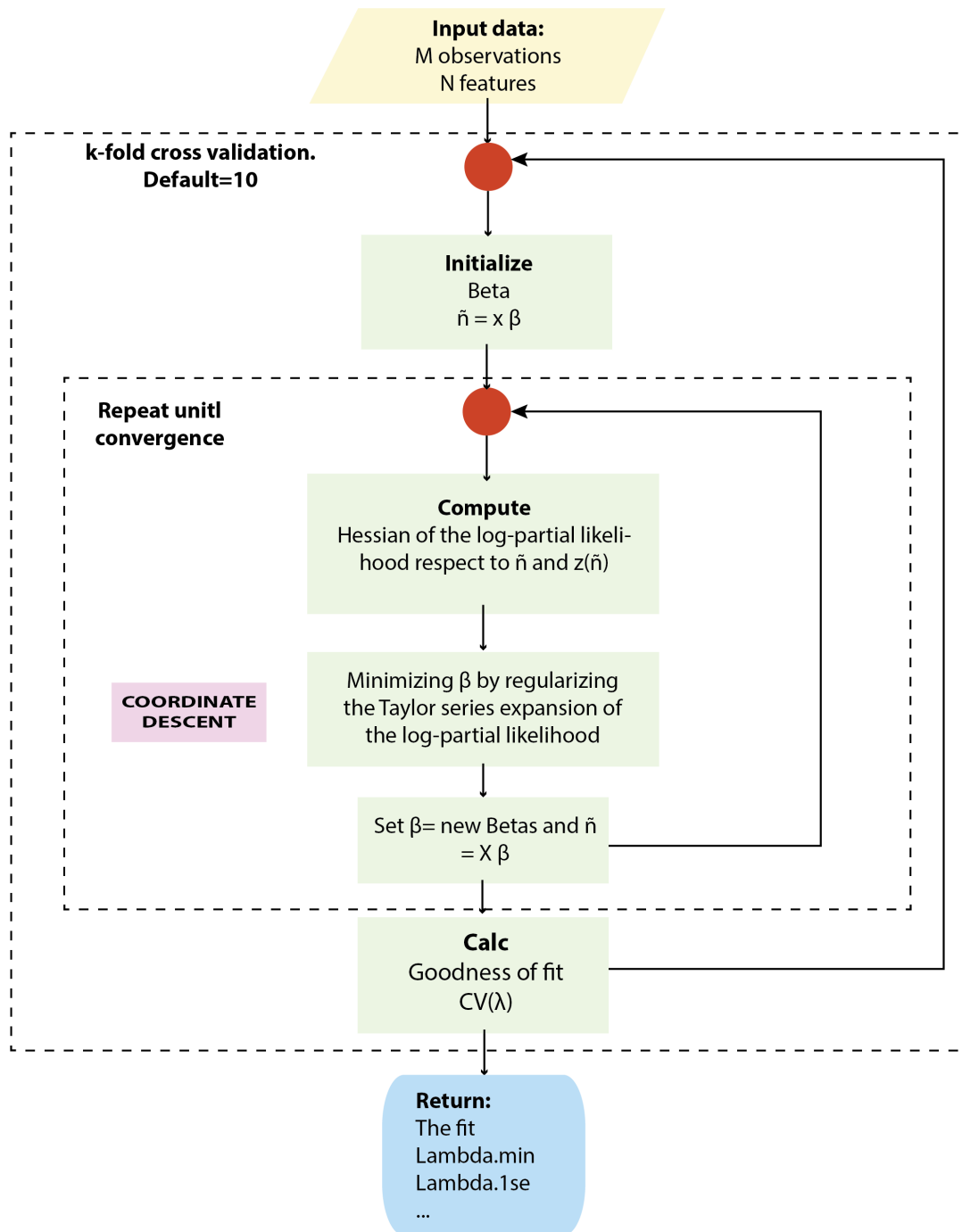


Figure 2.4: Stages of Coxnet procedure

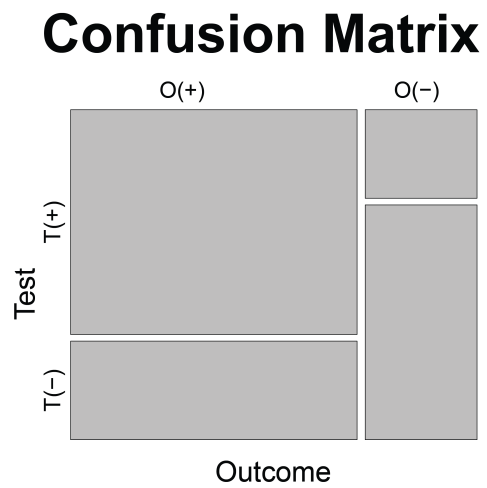


Figure 2.5: Confusion matrix plotted by FRESA.CAD R package in the ROC curve of random classifier with some data



# Chapter 3

## Methodology

The main purpose of this thesis is the evaluation of some different machine learning alternatives for the analysis of survival characteristics of patients suffering from chronic degenerative diseases. This analysis will be only possible with the information on time-to-event data along with the clinical and personal characteristics that describe each subject, the machine learning and statistical methods implemented, and the strategy of validation and benchmarking or the strategies; things that will be described here. In the previous chapters, the scope of the thesis was limited and the topics used to fulfill the mentioned objectives were taken into context. Now is time to take into consideration all the materials and methods that will be implemented to test our hypothesis. First of all, the data sets that will be used are described; With this, a complete description of the characteristics of each of the sets, the explanation of the origin of the data and the process of acquiring the imaging information of the observations in each of the initiatives will be detailed. Two types of data will be used to test the methods.

The first of the types will be data created solely for this study. The simulated data, when created with simulated information as absolute truth, are aimed at checking the operation of the techniques and knowing how reliable the subsequent results will be. The second type of information to be used will be clinical data of real patients suffering from Breast Cancer, Alzheimer's disease and Osteoarthritis with data from the analysis of images used in the normal process of diagnosis or treatment of the disease and clinical information. collected from each subject. Consequently, all this information has to be processed and transformed for use in the different computational techniques we are going to use in our analysis; This data preparation process is the next section in this chapter.

Subsequently, and taking into account that the data is ready, we will describe the implementation of the cross-validation process, in which the statistical analysis of survival of the data sets through different machine learning methods was taken into account. The implementation of various techniques requires the use of strategies that allow a fair comparison between the methods. For this, the cross-validation technique that will be detailed in the 3.2.3 point was considered. As a consequence, this process generates a large amount of information that in turn can be collected, analyzed and displayed graphically for better understanding. This process is combined in a single method of comparison called: Cox Benchmarking. We will contextualize the implementation of the comparative evaluation of Cox models, with the description of the use of graphic functions and statistical calculations that will allow the fair evaluation of the strategies. Finally, chapter 3 explains the objectives and the problem to be

solved for each of the experiments with the different data sets. The process for obtaining the results shown in the next chapter will be described.

## 3.1 Experiments: Data acquisition and preparation

Within the methodology of this research and after having placed into context all the topics necessary to understand the study, the first phase to be carried out is the acquisition, preparation, and analysis of the data to be used. In the next pages, we will detail each of the characteristics of the data sets, their origin, their context, and their demography.

### 3.1.1 Simulation data

One of the best known and used methods to verify the operation of computational strategies and statistical analysis is to make use of simulated data, where the author is the only one who knows the absolute truth. In the context of this study, it is necessary to carry out a data set with time-to-event information that is influenced to some extent by various variables that change the risk of suffering the event. In our case, the data simulation will be carried out in the field of basketball players who are about to start their professional careers. This theme was selected simply by the author's affinity and knowledge of this topic. The problem could be detailed as below.

Basketball professional teams, especially NBA teams, have a very specific event known as a draft. This special event is the place where a group of amateur players that have been previously filtered as possible players in the league, can be selected to sign a professional player contract. Each of the teams has access to the player's historical information and also their specific health data. That information is used to make a decision about whether the player worth it or not. This process (although it is more complex in real life) tries to find players that can be useful for the franchise and especially players that provide real results to the team, economically and in the sport context. Of the players selected, there are players that can last a long time in the organization, others that just retire some seasons after they began. One of the main problems with selected players in the draft is the time they remain in the league. Whether for sports, health or just that they do not meet the expectation, some players stop playing professionally in the NBA. Taking this into account, we are going to simulate historical information of 1000 players who have entered the league and relate their characteristics before entering the league (amateur stats) with the time of their NBA careers.

Once we know the problem and what we need for this data simulation, we decided just to use a small number of variables that have information about the players. These variables, in the real-life, could be or not risk factors for the withdrawal time of the prospects. Here, the characteristics were selected because of the ease of generating random numbers that follow a predefined distribution function data. The data generated involves information from 1000 former players and some current players with their simulated data from their amateur experience. The variables to use are detailed in the table 3.1 and the paragraphs below.

Eight of the ten features that will be part of this simulation follow a normal distribution and the two remaining follow a binomial distribution. For the simulation of these data, averages and standard deviations were estimated by the author's empirical knowledge and some



Covariate	Code	Distribution	Mean	Standard deviation
Body Mass Index	bmi	Normal	24	2
Age	age	Normal	21	1.5
Games played	games	Normal	40	8
Average minutes	minutes	Normal	22	5
Assists	AST	Normal	5	2
Field goal percentage	FGP	Normal	35	3
Block per game	BLK	Normal	1.8	0.5
Offensive rating	ORtg	Normal	105	3
Defensive Rating	DRtg	Normal	100	3
Injuries	I	Binomial	NA	0.5*
Surgeries	S	Binomial	NA	0.5*

Table 3.1: Features to be related with the NBA careers of 1000 NBA players simulated information. \* the probability of success in binomial distribution

research about the players' data before entering the draft. All the numbers were generated with Excel specifically with the provided data analysis tool. The only input that the tool required was the mean and standard deviation for the first eight features, and the probability for the remaining ones. Once these data have been generated, we can calculate the probability of survival of each subject, based on the calculated risk coefficients that we will assign to each variable according to our empirical knowledge and the simulated absolute truth. To calculate the coefficients, the cox model will be used. In the model, each variable affects the Hazard function with a different weight. Positive coefficients will mean that the effect size of that specific covariate will make the risk growth and negative ones will decrease the value. The Cox PH's formula let to overcome to this calculation and a new equation was derived for the calculation of the effect size ( $\beta$ ) of each variable per unit of change of each feature. In this formula, an initial risk value is considered with the first value of the variable and the second risk value when the variable changes in the unit of change determined. The formula 3.1 details the calculation of the covariates effect size.

$$\beta_{x_i} = \frac{\log[\log(1 - h_0(x_{i1})) - \log(1 - h_0(x_{i2}))]}{x_{i2} - x_{i1}} \quad (3.1)$$

For each covariate, a different unit of change and added risk was considered. These data were determined according to the simulated truth. In the following lines, the meaning of each of the variables and the values used to calculate the risk coefficients will be detailed. These effect sizes calculated when submitted to the cox model will let us know the probability of survival of each subject and thus also the probability of occurrence of the event for each unit of time selected.

We used Body max index (BMI) because it was the only way to use the height and the weight of the players without the need to generate random variables related to each other. The BMI is a person's weight in kilograms divided by the square of height in meters [55]. In this variable, we consider that 1 unit of change would increase the risk by 0.9%. The

Covariate $x_i$	$H_0(x_{i1})$	$H_0(x_{i2})$	$x_{i2} - x_{i1}$	$\beta$ Effect size
Body Mass Index	0.001	0.0011	1	0.041
Age	0.01	0.04	5	0.122
Games played	0.01	0.02	-51	-0.006
Average minutes	0.009	0.011	-35	-0.003
Assists	0.009	0.02	-10	-0.035
Field goal percentage	0.004	0.03	-20	-0.044
Block per game	0.001	0.035	-3	-0.517
Offensive rating	0.009	0.09	-20	-0.051
Defensive Rating	0.011	0.09	18	0.052
Injuries	0.005	0.05	1	1.010
Surgeries	0.005	0.15	1	1.511

Table 3.2: Features to be related with the NBA careers of 1000 NBA players simulated information.

resulting coefficient for the BMI was  $\beta = 0.041414$ . The next features were age, which we consider as an important variable to define the time the race can last. Each exchange unit is defined as the five-year difference, which means that if the difference between the two players is 5 years the risk of the major is 3% greater than that of the other; The calculated effect size is  $\beta = 0.121743$ . The games that the player had in the season before his postulation for the league, are described in the third variable. In this case, this variable has a negative relationship, which means that the risk difference between the player with the most games and the player with the minimum of games is 1% of the risk; the highest risk being that of the player with the least amount of games. The number of games ended with a negative impact measure of value  $\beta = -0.00595$ . The next feature is the average number of minutes each applicant played in each game last season. This characteristic also has a negative association where the 35 minutes difference represents a 0.2% risk change and its resulting coefficient was  $\beta = -0.0025$ .

Then we take into account the average number of assists per game a player has. Like the previous two, its relationship with the probability is also negative and the resulting coefficient was -0.03492. The difference between a player with 0 assists and one with 10 is 1.1 %. We continue with the percentage of successful shots, Field goal percentage (FGP) with a coefficient of  $\beta = -0.04404$  which refers to the difference of 2.6 % risk when the percentage difference is 20 units. The following is the number of blocks per game (BPG), the resulting coefficient is  $\beta = -0.51719$ . The difference of 3 blocks defines a 3.4% change in risk. The last variable with a negative relationship with risk is the offensive rating of each player (ORtg). This statistic gives us the amount of points that a player averages every 100 times he has the ball in his hands. Within our simulation, we will consider this characteristic as an important impact factor, so players with a difference of 20 points have an 8 % difference in risk, with a resulting coefficient of  $\beta = -0.05092$ . Quite the contrary in the value of the effect size  $\beta = 0.051709$ , is the defensive rating of each player (DRtg). This rating is the amount of points that a player allows per 100 possessions. The change of 18 points between players makes the

risk increase 7.9 %. Finally, there are the two variables that will have more weight among the simulated characteristics; First, there is the Injuries variable (I). The value 1 indicates that the player has suffered from injuries that have left him relegated from the courts and 0 that he could have had superficial injuries or not suffered them. This characteristic has a 4.5% risk increase if you have suffered the injury and the calculated coefficient is  $\beta = 1.010003$ . Second, there is the variable surgery (S), which with the value 1 indicates that the player has suffered an injury surgery. This variable increases the risk by 14.5% and the measure of impact is  $\beta = 1.510846$ .

A summary of the coefficients with the risk values considered for the calculations is found in table 3.2. The first column shows the risk with the first value of the variable, the second with the other variable value. The third column shows the difference of units between the values considered for the risk. Finally, there is the value of the calculated coefficient.

### 3.1.2 TADPOLE/ADNI

Considering the great problem caused by one of the best known chronic degenerative diseases, Alzheimer's disease, different initiatives have been created around the world to be able to control, diagnose and treat the disease. Many of these have allowed organizations or groups of institutions to join their purpose and through economic and academic incentives, to take advantage of all the information available in the initiatives. One of the most recent and well received challenges in the context of this disease was the TADPOLE Challenge. The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) Challenge born to compare different techniques to predict the future evolution of people at risk of Alzheimer's disease. All participants in this challenge were provided with historical data from patients belonging to the Alzheimer's Disease Neuroimaging Initiative (ADNI) [69]. The challenge was responsible for delivering a ADNI-derived set available via the Laboratory Of NeuroImaging: LONI; With that, they eliminated the need for data preprocessing to join patient information into a single spreadsheet.

#### ADNI data

The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary objective of ADNI has been to test whether MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see [www.adni-info.org](http://www.adni-info.org). The initial objective of this initiative was to recruit 800 subjects, however, its successful start and above all the support of different organizations allowed the continuation of new protocols ADNIGO, ADNI-2, and ADNI-3. Currently, and according to TADPOLE, the data set delivered with the first three protocols have recruited more than 1500 adults, aged between 55 and 90 years. These people participate in the research and mostly consist of cognitively normal people, people with early or late MCI and people with early AD. ADNI provides its inclusion criteria in [84]. Data used in the final spreadsheet provided by the TADPOLE Challenge has ADNI information about: (1) CSF markers of amyloid-beta and tau deposition; (2) various imaging modalities such as magnetic resonance imaging (MRI), positron emission tomography (PET) using several

tracers: Fluorodeoxyglucose (FDG, hypometabolism), AV45 (amyloid), AV1451 (tau) as well as diffusion tensor imaging (DTI); (3) cognitive assessments acquired in the presence of a clinical expert; (4) genetic information such as apolipoprotein E4 (APOE4) status extracted from DNA samples; and (5) general demographic information.

### Tadpole datasets

TADPOLE provides three types of datasets: (1) *training data set* which refers to the measurements with associated outcomes that will be used to train the algorithms. (2) *Prediction data set* that contains only baseline longitudinal measurements without associated outcomes. This data is provided to be used as input to make the forecast process in the challenge. (3) *Test data set* which contains the real outcomes of each patient to compare with the calculated forecast. Regarding the two first types of the datasets, TADPOLE prepared three standard datasets:

- **D1:** TADPOLE Standard training set based on longitudinal data across ADNI1, ADNI GO and ADNI2. The information is a set of measurements for every patient that at least two separate visits (different dates) in the process. D1 contains information of 1667 patients.
- **D2:** TADPOLE longitudinal prediction set contains information from ADNI rollover individuals whom data has to be used for the forecast in the challenge. D2 dataset includes all the time related information of the patients. It contains information of 896 patients.
- **D3:** TADPOLE cross sectional prediction set contains the most recent time point and a limited set of variables for each rollover patient in D2. D3 shows the information typically available when selecting a cohort for a clinical trial.

In this thesis we will consider just D1 and D2 combined information. The process of the material selection will be described later.

### Image pre-processing

ADNI manages its own protocols for obtaining information from medical images. Imaging information has been pre-processed with standard ADNI pipelines [124]. Specifically, for these thesis we will concentrate in three-dimensional T1-weighted magnetic resonance imaging which provides extensive information to develop and test analysis techniques to study the conversion of MCI to AD. ADNI MRI Core created standardized analyzes that include scans that meet the minimum quality control requirements, the dataset included correction for nonlinearity of gradient, correction of non-uniformity B1 and sharpness of spikes [69]. Significant regional features such as volume and cortical thickness were extracted using Freesurfer transverse and longitudinal pipelines [90].

### APOE

There are some useful clinical information such as Apolipoprotein E (APOE), that is a protein involved in the metabolism of fats in the body with a polymorphic structure that has

three major alleles [101]. The fourth allele (APOE4) had been validated several times as a biomarker indicative of the risk of suffer Alzheimer’s disease [23]. Screening laboratories were obtained as well as blood for DNA for APOE testing [84]. ADNI uses information of APOE4 biomarker as APOE status that was treated as a categorical variable with three levels (Noncarriers < Heterozygotes < Homozygotes) [113].

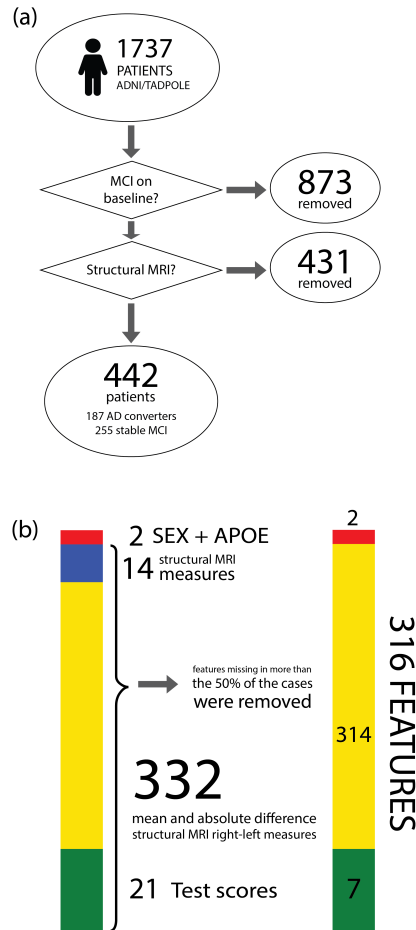


Figure 3.1: ADNI/TADPOLE (a) Patient selection process. (b) Feature types used in this study.

### Cognitive Assessments

Within the information available from ADNI. Data from neuropsychological tests to patients are found. In this thesis they will be mentioned as Cognitive Assessments or test scores. The objective of the ADNI neuropsychological tests is to make use of objective and reliable procedures to measure the cognitive abilities of a patient. There are several problems within the tests that the examiner may encounter, the problems can be emotional and physical interfering with the test results. To prevent this, the examiner must perform several tests while determining the patient’s condition. Protocols and guidelines indicated as the aforementioned, are found in the ADNI data usage manual. These guidelines ensure in a certain way the generation of valid and accurate measurements with a minimum of stress and discomfort for the

participants.

### Material

The ADNI/TADPOLE challenge datasets considered for this study were: “D1 - a comprehensive longitudinal data set for training”, and “D2 - a comprehensive longitudinal data set on rollover subjects for forecasting”. The challenge included 1737 individuals from the ADNI database with longitudinal observations. Each subjects’ data included the diagnosis status, neurocognitive evaluations, qMRI longitudinal observations, PET studies, APOE4 status [69]. For this study, we selected some features: sex, APOE4, 21 Test-scores and the 346 longitudinal qMRI measurements provided by the University of California San Francisco (UCSF). UCSF used FreeSurfer Version 4.4, for the analysis of the MRI data sets [90].

We divided the dataset into groups depending on their condition. First of all, the provided dataset had the information about all the patient’s visits. Since this investigation considers the survival analysis, we are going to use the information of the baseline visit, which is the first visit regarding this situation. The dataset included 864 MCI diagnosed subjects at baseline. Of them, 431 MCI subjects did not have the structural qMRI data leaving just 442 MCI patients remaining with longitudinal qMRI. The 442 patient’s information at the baseline were studied in this thesis. Among the studied subjects, just 187 patients demonstrated MCI to AD conversion and 255 maintained the MCI diagnosis during the observation period. Furthermore, we used normal control patients from the TADPOLE/D1-D2 dataset with qMRI information (n=233) as reference controls. Figure 3.1 shows the selection process and the main features considered in this paper. Table 3.3 shows the demographics of the three groups.

	Sex		Mean age (s.d.)	Mean Time-to-event (s.d.)	APOE*		
	M	F			1	2	3
MCI to AD	107	80	73.41( 7.13)	848.56(678.96)	64	98	25
No Event	150	105	73.1(7.60)	1470.22(967.97)	149	81	25
Normal Control	113	120	74.58(5.27)	NA	167	62	4

Table 3.3: Characteristics of tadpole challenge subjects used in this study. 187 patients presented the MCI to AD conversion event and 255 maintained the MCI diagnosis during the observation period. The normal control patients (n=233) were used as reference controls. APOE status (1: Noncarriers 2: Heterozygotes 3:Homozygotes)

### Data conditioning

We extended the information provided by the TADPOLE Challenge by computing the time to MCI-to-AD conversion. The event time for stable MCI subjects consisted of the difference in days between the date of the baseline and the date of the last recorded follow-up visit. The event time for subjects that suffer the MCI-to-AD conversion consisted of the difference in days between the date of first AD diagnosis and the baseline date. MCI stable subjects were labeled as censored. After computing the event time, we explored the 346 baseline-qMRI measurements. 332 of these correspond to measures of the left and the right side of

the same brain region. Because AD affects both sides of the brain, we described the left-right paired measurements as the mean and absolute differences between them. After that, all the measurements were z-normalized using the 233 normal subjects as reference controls. Finally, qMRI features that were not measured in more than half of the subjects were removed ( $n=28$ ). After that, the non-reported values of the 314 qMRI features that had majority representation were imputed by the nearest neighbor strategy [104]. A complete graphical summary of the data conditioning process can be found in Figure 3.1(b). Figure 3.2 shows an overall heatmap representation of the analyzed data.

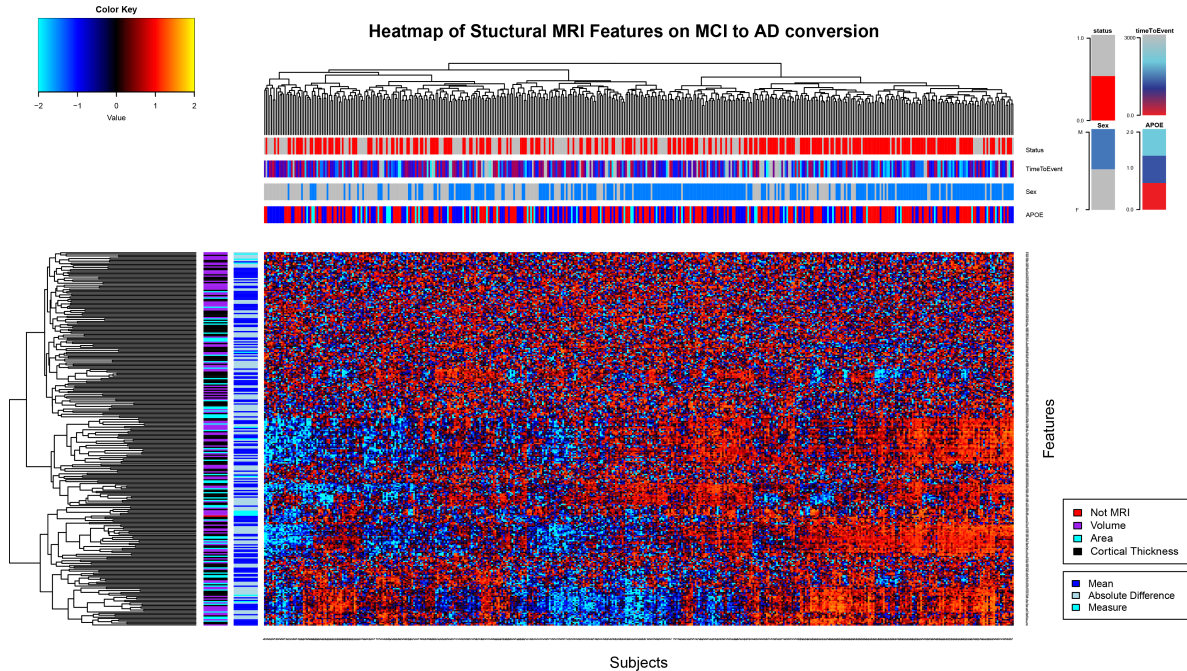


Figure 3.2: Heat map with 301 features selected by all the Machine Learning Methods. On the top section, patients dendrogram and 4 bars with the subjects' information about conversion, time to event, sex and APOE. On the left section, dendrogram of features and the information about the type of feature. Subject identification x-axis, features on y-axis.

### 3.1.3 Osteoarthritis Initiative: OAI

Osteoarthritis (OA) is the most common form of arthritis; it causes considerable disability in the elderly populations. Osteoarthritis does not have a consistent technique that can be used for its early diagnosis and is more common than expected. In Mexico, the prevalence of osteoarthritis was 10.5% [82] and despite a high prevalence, there is no treatment or medication that can cure it. And within this disease, when it attacks the knee it becomes the most common cause of disability in adults (CITA). The "Osteoarthritis Initiative (OAI): a study of knee health" is a multicenter, longitudinal, prospective observational study of knee osteoarthritis that allows researchers to gather information about physical changes in the development of the disease. This Initiative is a public-private partnership between the NIH and private industry that seeks to develop a public domain. The main purpose of this study is to examine people

who have knee arthritis or who have a high risk of knee arthritis; assessing biomarkers that will give us more information to better understand how to prevent and treat OA. In this Thesis will be using just part of the available information in OAI. Osteoarthritis Initiative has been collecting large amounts of clinical data in patients with OA; Although all these characteristics provide important information about the stages of pain and other characteristics of the subject, we will concentrate on raw and derived X-ray measurements, demographic information, and scores of standardized questionnaires. The study is comprised of three subgroups: 1) those with clinically significant knee OA who are at risk of disease progression, 2) individuals who are at high risk of developing clinically significant knee OA, and 3) a normal control group.

In the literature, considering the nature of the image itself, it is obvious the implication of possible better results using Magnetic Resonance Imaging (MRI) than X-ray imaging to study different outcomes in OAI. Nevertheless, some studies are not entirely complete, such as the rate of incidence of Total knee replacement related to X-ray measurements. In the baseline visit, Joint imaging biomarkers (magnetic resonance imaging and radiography) and biochemical and genetic markers (from blood and urine) are collected, this section is concentrated in the radiographies. The OAI data gives us the necessary information to explore this relationship and that is why we chose it.

### **X-rays acquisition**

OAI owns a protocol for X-ray acquisition that considers all the requirements for participant inclusion or the x-ray technologist and investigator roles. In our case, here, we are going to explain how the baseline x-ray measurements were acquired. Depending on the cohort on which each patient was classified, the schedule of the radiographic examination was set. The radiographs acquired for a participant will differ depending on their sub-cohort assignment and visit. Some of the patients due to the lack of quality in the radiographs taken could need a procedure of repeated exam. Following the standard radiology process, each study center uses a single x-ray unit for each acquisition protocol to avoid variability in the data. Measurements were extracted from those radiographs with different techniques. All the quantitative data set contains measurements of longitudinal center joint space width and related parameters of serial OAI knee x-rays were performed in the laboratory of Dr. Jeff Duryea at Brigham and Women's Hospital in Boston, MA [32].

These measures are taken directly on the x-ray image, and are presented in millimeters, because of this is considered quantitative scores.

On the other hand, the semi-quantitative dataset contains the readings for the central longitudinal cohort of serial Entire OAI knee x-rays. These measurements were extracted at the Boston University Clinical Epidemiology Research and Training Unit, under the direction of Dr. David Felson, MD. [35]. These measures are taken directly from the x-ray image; the image is compared with an atlas, and is scored from 0 to 3, where 0 is no evidence of radiological OA.

### **WOMAC**

Besides the use of imaging features within this dataset, we use also add one on of the most important assessments for the OA screening, the Western Ontario and McMaster Universities Arthritis Index (WOMAC). The WOMAC is a widely used evaluation of Hip and Knee



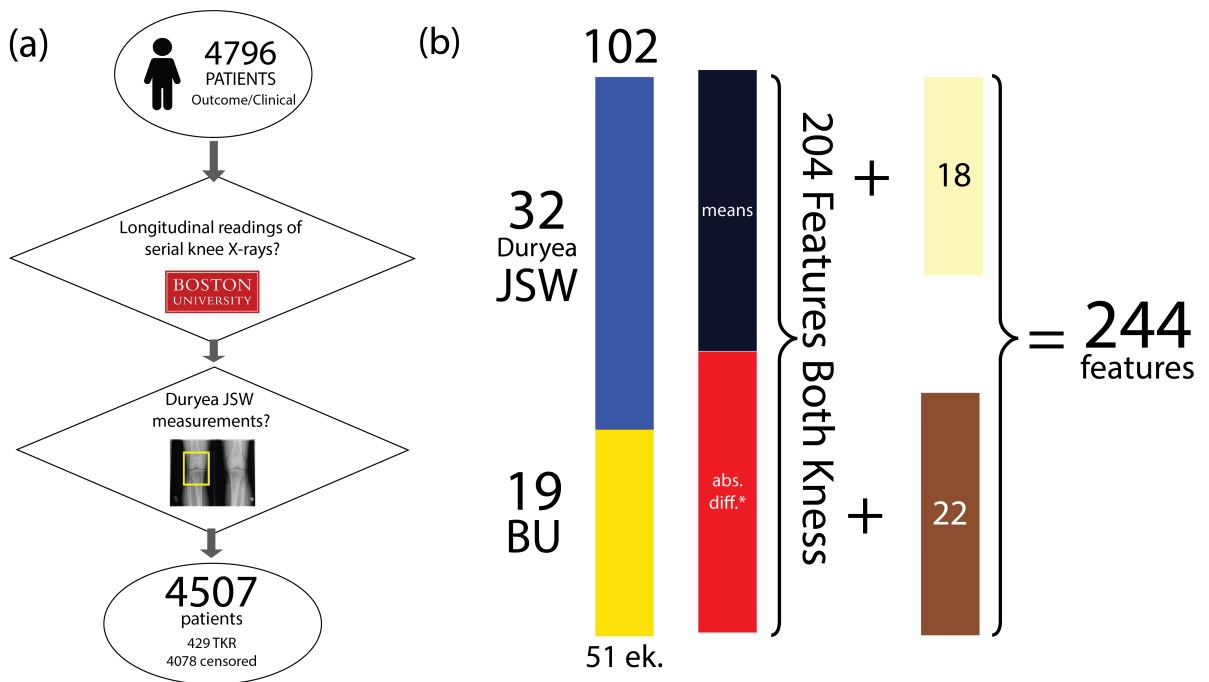


Figure 3.3: Summary of the selection of Participants for the OAI experiment and data conditioning process. (a) Participants selection (b) Feature types and data conditioning process. \*Absolute difference of X-ray measurements

Osteoarthritis. It is a self-administered questionnaire which consists in 24 items divided into 3 sections. It was developed in 1982 at Western Ontario and McMaster Universities and is available in over 65 languages which has been linguistically validated [12].

The sections consider all the possible characteristics that can explain the symptoms of the patients. The first scale is Pain. To this belongs 5 items. So, the patient has to evaluate the degree of pain suffer during walking, using stairs, in bed, sitting or lying, and standing upright. The second section describes level of Stiffness and 2 items belong to this. Stiffness after first waking and later in the day. The third section considers the Physical Function of 17 actions. Using stairs, rising from sitting, standing, bending, walking, getting in-out of a car, shopping, putting on / taking off socks, rising from bed, lying in bed, getting in-out of bath, sitting, getting on-off toilet, heavy domestic duties, light domestic duties. ] [https://www.physio-pedia.com/WOMAC\\_Osteoarthritis\\_Index](https://www.physio-pedia.com/WOMAC_Osteoarthritis_Index)

## KOOS

In addition to the WOMAC form, the Knee injury and Osteoarthritis Outcome Score (KOOS) was developed. The main objective of this form is to evaluate short and long term symptoms and function in subjects with a knee injury and osteoarthritis. Like WOMAC, KOOS divides the scores depending on the context of the form questions. In this case, it is divided into 5 subscales. Pain, other Symptoms, Function in daily living (ADL), Function in Sport and Recreation (Sport / Rec), and knee-related Quality of Life (QOL), we included all of them.

## Material and Data conditioning

“Data used in the preparation of this article were obtained from the Osteoarthritis Initiative (OAI) database, which is available for public access at <http://www.oai.ucsf.edu/>. The process of preparing data for OAI required a set of more extensive steps than those performed for the experiment with ADNI information. OAI has its data sets available through public text files that divide the information by origin and by knee of each patient. In addition, they include the patient’s clinical and demographic information in separate files. Each scheduled visit of the patients is a separate file with the information acquired in it and in the end, a file of patient outcomes is collected. In our case, we used all the information from the baseline visit; that is, all the text files that had the visiting code as 00. All the clinical, forms and demographic information of the patient was collected within a single data set. The entire process require different steps to be completed. So, in the next paragraphs we will detail the process. A graphical summary of the material used and the participants selections is shown in the Figure 3.3

**Clinical and demographic information** patients participating in its initiative. In our case, we take into account patients whose information is available from the first visit. The demographic information we consider is age, sex, weight, height and BMI data. On the other hand, we added 13 variables that belong to the scores generated from the clinical assessments used for OA screening. The scores included are part of KOOS and WOMAC. In total, this dataset includes patient information with 18 informative variables. Together with these variables, this data set provides the outcome of the patients, which in our case is the Total Knee Replacement (TKR). For the calculation of time-to-event, the first TKR of the patient was taken, regardless of the knee that is placed to include the information of both knees. Patients who do not have TKR are censored 3500 days after the start of the follow-up time.

**Join Knee X-rays** OAI provides Knee X-Rays measurements in different datasets depending on its origin. In this experiment, we are going to use two main datasets. The first one contains longitudinal readings of serial knee X-rays for tibiofemoral radiographic OA done in the Boston University Clinical Epidemiology Research and Training Unit by Dr. Piran Aliabadi and other researchers [35]. The second dataset contains central longitudinal measurements of joint space width (JSW) and related parameters of serial OAI knee x-rays. This study was performed by Jeff Duryea at Brigham and Women’s Hospital in Boston [32]. OAI provides these datasets in a text format divided into the visits during the 48-month follow-up time. In our case, as we already mentioned, we used just the baseline visit.

In the first case, the file contains information of 12813 longitudinal readings. Of them, 4799 are repeated samples for some patients. Considering that, we use the first appearance of each patient and we left just 8014 rows on that file. These measurements belong to 4507 patients and there is an observation for each knee. In the case of Duryea data, the same amount of information was not found. 3090 rows corresponding to left knees and 3088 right knees were found. Considering that Boston University data has a greater number of observations, we take that number into account in order to create a complete dataset. Taking into account that we have information by knee and by the patient, we start with the 4507 data that we have corresponding to the BU data. We separate the data sets by knee and add the corresponding

information to the patient's row Duryea data. Both data sets contain 19 variables belonging to BU and 32 from Duryea, a total of 51 X-ray characteristics.

To increase the amount of information we are going to analyze, we use the information we have corresponding to patients. First, we will make simple calculations with the available information. In this case, we have the information of 4507 patients per knee. We grouped by the patient in a single data set the information of both knees, leaving us with a data set of 102 variables. We increase the data, adding the sums and the absolute differences between the measurements of each knee. In total, we will obtain a set with 204 x-ray characteristics.

#	Medial Compartment	Lateral Compartment
1	150	850
2	175	825
3	200	800
4	225	775
5	250	750
6	275	725
7	300	700

Table 3.4: The relation between positions in the x-axis. The difference between the first column which belongs to the Medial compartment and the second column belonging to the Lateral compartment is calculated (Position 150 - Position 850)

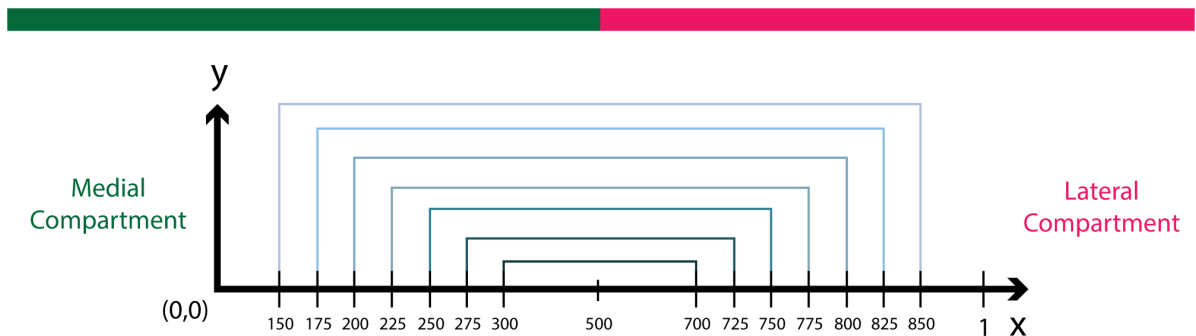


Figure 3.4: Approximation of positions on x-axis that Duryea provides in the dataset. Green section is the representation of the Medial Compartment and the red section is the representation for Lateral Compartment. Lines between the positions show the relation between them.

**Data augmentation (derived information)** After simple calculations with available data, we will explore measurements that have a clinical basis. In this case, we will use measurements that are part of the Joint Space Width (JSW). Considering Duryea's measurements and how to analyze the images of your software [DURYEA], there are two sections, the Medial and Lateral compartment. Duryea takes into account the medial section as the origin for both axis and the lateral compartment being the limit 1. The measurements corresponding to JSW will be part of the y-axis, on the other axis, measurements are made in sections that allow you to have an idea Better how the knee is. The measurements found on the x-axis and are greater

than 0.5 are part of the lateral compartment. The set of positions on the x-axis from which the medial compartment measurements are taken is as follows: 150,176,200,225,250,275,300. The lateral section positions are: 700,725,750,775,800,825,850,875,900.

The first measurement derived from these compartments is the difference between the opposite measurements of each compartment. That is, the first measurement of the medial side that corresponds to position 150 will be used to calculate the difference with position 850 that corresponds to the first measurement of the lateral section. The table 3.4 shows the differences that will be calculated from the positions. A total of fourteen features are added to the dataset.

Features	Source	Description	Quantity	Total
X-ray measurements	<i>BU</i>	Longitudinal readings of serial knee X-rays for tibiofemoral radiographic OA.	19 each knee	38
X-ray measurements	<i>Jeff Duryea</i>	Central longitudinal measurements of JSW.	32 each knee	64
Mean of X-ray measurements	<i>Derived</i>	Sum of the corresponding measurements for each knee. 51 measurements of each knee are used in the sum.	51	51
Absolute Difference of X-ray measurements	<i>Derived</i>	The absolute difference in the corresponding measurements for each knee. 51 measurements of each knee are used in the sum.	51	51
Clinical Data	<i>OAI</i>	Demographic and clinical information of the patient. Scores of OA assessments for screening.	18	18
Difference on JSW	<i>Derived</i>	Differences of measurements in the corresponding positions of each compartment on each knee.	7 each knee	14
Slopes of JSW	<i>Derived</i>	Slopes of measurements of each compartment on each knee.	2 each knee	4
SD of JSW	<i>Derived</i>	SD of measurements of each compartment on each knee.	2 each knee	4
<b>Total features:</b>				244

Table 3.5: Summary of the features used in this Experiment. Detailed information of the features can be found in the previous paragraphs.

After the differences were calculated, we take advantage of the fact that the data was already grouped by the medial and lateral compartment of each knee. Two more groups of variables will be calculated from them. The first is the slope generated from the y-axis values (measurements) of each of the zones. In other words, a total of 4 slopes will be added, 2 from each knee. The first one corresponding to the medial compartment and the other to the

lateral one. The calculation of the slope will be made from a linear model for the values of the measurements. The second measure derived from this grouping of data will be the standard deviation of the measurements per compartment. Which gives a total of 4 deviations, one for each compartment. Being a total of 8 variables that joining them with the 14 differences mentioned above, they form a set of 22 variables derived from JSW. Table 3.5 summaries all the features, raw and derived features will be listed.

### 3.1.4 Prognostic Wisconsin Breast Cancer Database

Breast Cancer disease has always been a problem and its prevalence rate is still high today. Different studies have tried to find patterns that help their diagnosis and clinical treatment. This is the case of the information available in the Machine Learning Repository of the University of Wisconsin. On which, Dr. Wolberg managed to collect a set of 569 patients since 1984 [120]. The data set has two study options, one data set was focused on the diagnosis of the disease and another for the prognostic. The characteristics are different depending on the dataset chosen. In the context of this thesis and considering that our study handles information from time to event, we will use the prognostic data set available online: <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Prognostic%29>. This set just has information about 198 patients. The data set includes 198 patients with 34 characteristics. Among the characteristics is the patient's clinical information, the condition of recurrence (status) the time to said event and 30 characteristics calculated from a digitalized image of a fine needle aspirate (FNA) of a breast mass. These characteristics describe the cell nuclei present in the image. Table 3.6 summarizes the measurements taken for each of the cells; The method [14] explains their extraction for the set.

<b>Feature</b>	<b>Description</b>	<b>Type</b>
<i>Outcome</i>	Recur (1) or nonrecur (0)	Status
<i>Time</i>	Time in days	Time to Event
<i>Radius</i>	mean of distances to the center	Cell nucli
<i>Texture</i>	standard deviation of gray scale values	Cell nucli
<i>Perimeter</i>		Cell nucli
<i>Area</i>		Cell nucli
<i>Smoothness</i>	local variation of radius	Cell nucli
<i>Compactness</i>	$\frac{perimeter^2}{(area-1)}$	Cell nucli
<i>Concavity</i>	severity of concave portions of the contour	Cell nucli
<i>Concave</i>	number of concave portions	Cell nucli
<i>Symmetry</i>		Cell nucli
<i>Fractal</i>		Cell nucli
<i>Tumor size</i>	diameter in centimeters	Measure
<i>Lymph node status</i>	number of positive axillary lymph nodes	Clinical

Table 3.6: Summary of the features of Prognostic Wisconsin BRCA Database. A total of 198 patients and 34 features are used.

### 3.1.5 Prognostic San Jose Hospital Breast Cancer Database

As was mentioned in the Introduction of this thesis, BRCA is the most commonly occurring cancer in women and the second most common cancer overall. Specifically, in Mexico the incidence and mortality of this type of cancer have risen in the past years which lead to changes in health-care policies to treat this disease and now they focus on early detection and treatment [2]. Considering this, there are some literature about BCRA in Mexican Population [38, 65, 105]. In this experiment we investigated the relationship of tumor image data with gene expression signatures of 73 subjects with BRCA who underwent digital mammograms and tumor biopsies prior to treatment. These patients were part of the prospective study “Exploratory study for image-based biomarker discovery of breast cancer and its biological validation” (THSJ-BC) which was approved by the institutional review board years ago. We recruited patients with breast cancer identified by mammography or by clinical examination. All patients were pathologically confirmed. A complete description of patient inclusion criteria, image acquisition and feature extraction, gene signatures estimations are detailed in the original publication [105]. Table shows the summary of the features used in the previous works.

#### Data conditioning

In this experiment, we add the real recurrence information to the image raw features and gene signatures data. The Follow-up information was collected from the enrollment date to the last section of the year 2019.

The recurrence date is set on the date of the visit in which the patient change his condition. The patients survival information is summarized in Table 3.7. The remaining features were used as they were in the Experiment of Tamez et al. We used absolute differences and means of each raw features. also we used gene exam scores. The following subsection explains more about the features.

<b>Condition</b>	<b>Age</b>	<b>Age &gt;50</b>	<b>Mean(age)</b>	<b>Mean time (censored,event)</b>
<i>Recurrence</i>	17	8	49.65	1004
<i>No Event</i>	56	28	50.59	1849

Table 3.7: Survival groups of patients suffering of BRCA in San Jose experiment

#### Features

Table 3.1.5 summarizes the features used in this experiment. First feature type “Image Features” are the same used in [105]. Means and absolute differences on Mammograms measures are used. The features were extracted in different groups by using some approaches. The description of the technique used for the feature extraction process is found in the original paper. The final data used is part of the Additional Resources that the paper provides. In this work, we used 25 signal distribution features, 25 fractals, 24 gray level co-occurrence texture features, 40 features derived by Minkowski functionals, 400 features coming from the Wavelet functionals decomposition and 25 local standard deviations.

<b>Feature Type</b>	<b>Group</b>	<b>Features</b>	<b>Description</b>
<b>Image Features</b>	<i>Signal distribution</i>	25	Specify the pixel-wise variations of the ROI signal intensities. Particular importance was considered in the smaller and larger quantiles
	<i>Fractals</i>	25	A set of texture features based on signal distribution of the local fractal dimension
	<i>Gray Level Co-Occurrence</i>	24	Haralick's gray-level co-occurrence matrix texture features: energy, entropy, correlation, difference moment, inertia, cluster shade, prominence and Haralick correlation at 3 distances (1,2 and 4 pixels).
	<i>Minkowski functionals</i>	40	Ten values (curvature energy, kurtosis, largest derivate, mean, half of maximum height, skewness, energy of the derivate, order of largest derivate, signal standard deviation, functional standard deviation) of the area, contour length, Euler number, and compactness of the image resulting from the binarization of the ROI at uniformly spaced threshold values
	<i>Wavelet functionals</i>	400	Decomposition of the images using the Daubechies-4 mother wavelet in 2D at multi-resolution levels. The resulting images were subject to the signal distribution and fractal features extraction process (4x4x25)
	<i>Local standard deviation</i>	25	Signal distribution of the standard deviation from a 3x3 pixel image
<b>Gene Expressions</b>	<i>Oncotype and PAM50</i>	4	Oncotype, PAM50 scores and calculated average Risk
<b>Patient Information</b>	<i>Clinical</i>	7	Clinical patient information
	<i>Survival</i>	2	Recurrence and time to event variables.

Table 3.8: San Jose BRCA information summary of features. First column describe the feature type, the second details the group of the features. Third column shows the number of features on that group and the last column shows the description of the group

Besides, we include the gene expressions scores used in the same work (Oncotype and PAM50) which adds 4 variables. The Patient information is also used with the survival information of each patient which adds 9 features. A total of 1091 features were analyzed in this experiment.

## 3.2 Cox Benchmarking implementation

Within the FRESA.CAD package [104], there are different benchmarking methods to compare machine learning algorithms behavior for different outcomes. Benchmarking for binary, ordinal and regression classification outcomes are fully available for use in that package. Each of the benchmarking methods considers many widely and commonly used algorithms to explain the outcome. FRESA.CAD details the output of these techniques and returns a model, which quantitatively and graphically, allows comparing the performance of these techniques in a cross-validation technique. On the other hand, other outcomes such as the survival analysis outcome, do not have methods that allow a simple comparison between different strategies. Specifically, ML/SL algorithms that make use of Cox models do not have a known benchmarking method for use. For this reason, this thesis details the implementation of Cox Benchmarking to be included within FRESA.CAD and later be used for the comparison of performances in the different experiments detailed in this chapter in section 3.1.

For the Cox Benchmarking implementation we are going to use Cox model. Proportional hazard model (PH) or Cox models explore the relationship between the time to the event and the possible explanatory variables. The model estimates the hazard  $\lambda_i$  of the subject  $i$  given the observed feature vector  $X_i = \{X_{i1}, \dots, X_{ip}\}$ , and the unknown baseline hazard  $\lambda(t_0)$ . i.e,

$$\lambda_i(t|X_i) = \lambda(t_0)e^{(X_i \cdot \beta)}, \quad (3.2)$$

where  $\beta = \{\beta_1, \dots, \beta_P\}$  is the vector of coefficients. Hence, the Cox model provided an estimate of total hazard (risk) of conversion, for an individual, given the observed features. Due to the large set of possible qMRI features to be considered in some of the Cox models, machine learning (ML) methods were used to find the “optimal” set of features and their corresponding coefficients that mimiced the observed rate of conversion.

Statistical techniques such as the PH model require technical implementations and complex calculations that are better optimized through machine learning techniques. The implementation of this Cox Benchmarking method has 11 different Machine Learning and Learning Statistics strategies that make use of the Cox technique for the evaluation of survival with time data at events. These strategies are included in subsection 3.2.1 of this section. The ML/SL approaches use feature selection (FS) as a common method used to construct Cox [15, 1] models. The wide variety of methods available to researchers can turn the discovery of biomarkers into a complex effort, especially when there is no clear choice of methodology to build/explore survival models. Thus, as a simpler model in which the result can be better explained through its predictors. Fully linked is the model selection, which is also an important step when conducting experiments with machine learning techniques. Different characteristics can generate different models, which, as appropriate, need a set of statistics that can evaluate the performance of each model and detect the model that maximizes or minimizes the statistics or the



set of them for the result and specific methods. In the specific case of this implementation, of the 11 methods, just 7 present the techniques of FS and MS for the study of survival of time-to-event data. These methods will be treated as Wrappers throughout the Cox Benchmarking implementation and the remaining 4 will be labeled as filters.

We propose a unified approach for the study of Cox models in an ML setting. The approach is based on repeated cross-validating ML/SL methods using exactly the same training-testing sets across all the methods. The ML implementation evaluates LASSO, RIDGE, BSWiMS, GPDAS, SPDAS, SPDAS adjusted with Bayesian information criterion (BIC) (SPDAS.BIC) [97] and Univariate Filtering for building suitable survival models. Thus, at the end of the repeated CV, a fair method comparison and a comprehensive evaluation of the role of each potential biomarker inside a Cox survival model is provided. The Cox Benchmarking implementation is available in the following link. The most important points of the implementation are detailed below.

### 3.2.1 Default ML/SL algorithms

Statistical Learning (SL) and Machine learning (ML) approaches provide efficient and highly competitive solutions to the issues of regularization and subset selection. Embedded statistical learning like L1 regularization via LASSO or L2 regularization RIDGE, allows the exploration of multivariate models composed on hundreds of features [98]. Also, this technique allows subset-selection with the exploration of realizable Cox models from a big number of features [117]. Model selection via the Bootstrap Step-Wise Model selection (BSWiMS), and two algorithms on Best subset selection package (BeSS) (Golden section primal-dual active set (GPDAS) , Sequential primal-dual active set (SPDAS)) are among three of the machine learning options readily available to researchers [117, 105]. This implementation of Cox Benchmarking allows the comparison of Machine Learning techniques through the calculation of the graphic and quantitative organization of the statistics of each of the models. Specifically, it allows the evaluation of the 11 pre-defined strategies by default. All the 11 strategies are part of the aforementioned R packages plus the Survival Package [111]. All the methods were modeled into the FRESA.CAD environment for its best coupling between them. The modeling process were implemented for this thesis.

Besides, the benchmarking process divides the techniques to be compared into two main groups. The Wrappers section, that uses algorithms that allow the selection of features, model construction and subsequently the selection of the best model. In contrary, the second section only considers the selection of characteristics of some methods and later construct a single Cox model.

**FRESA.CAD.** For the default methods, three freely available packages are considered, which use the proportional model for the study of survival information. The first method included was the Bootstrapped Stage-wise Model Selection (BSWiMS). BSWiMS is part of the FRESA.CAD R package and is a supervised model-selection method aimed to select a unique statistical model that predicts a user-specified outcome, in this case, a survival outcome. The statistical model is constructed by bagging a set of Cox models built by the unique set of model-wise statistically-significant features [105].

**GLMNET.** The second package was the Penalized Cox Regression (CoxNet) part of the glmnet. CoxNet algorithm fits the Cox Model regularized by an elastic net penalty [98]. Different parameters on the Elastic net penalty lead to different SL methods. Therefore, by changing the value of the alpha parameter we manage to control the type of regularization that will be used to find the model. All possible regularizations were executed with internal cross-validation, to determine the optimal lambda value. In this case, we executed Coxnet with three different values of alpha ( $\alpha = 1$ ,  $\alpha = 0.95$  and  $\alpha = 0$ ) which resulted in two different approaches. The first method was executed in the LASSO penalty (alpha=1). Briefly, the LASSO considers the L1 regularization only, which decreases the coefficients by a constant (lambda) to perform feature selection removing those coefficients lower than lambda. With alpha=0, the method regularizes the model with a RIDGE penalty. RIDGE considers L2 regularization which scales all the coefficients towards 0 but sets none to exactly zero. Further, with an alpha value between 0 and 1, we get the ELASTICNET approach which is a mixture of L1 regularization and ridge regression. We used  $\alpha = 0.95$  because the model will work like lasso and only deleting the degenerate behavior due to extreme correlations. In the CoxBenchmarking tool, the user can specify an alpha value for the regularization. In this thesis, we used the most common alpha values among literature and recommended author values for each of the techniques. Formula 3.3 summarizes how the alpha value works in the regularization process. With  $\alpha = 0$  it turns into ridge regression, with  $\alpha = 1$  it turns into Lasso. Values between 0-1 turn the regularization into ElasticNet penalty.

$$\alpha \sum |\beta_i| + (1 + \alpha) \sum \beta_i^2 \leq c. \quad (3.3)$$

Thus, the last two methods better handle correlated predictors but do not select features. To overcome this limitation, we proposed a threshold value to select a limited number of features. Features that do not exceed the coefficient threshold were considered to build the model.

**BeSS.** The third strategy used was the Primal-dual Active set. This is part of the BeSS (Best subset selection) R package. This method uses an efficient active set algorithm to choose the best possible Cox model. As Coxnet, BeSS can also turn into different strategies; this time by using different algorithms. The default configuration proposed by BeSS authors uses the Golden Section primal-dual active set (GPDAS) algorithm [117], our fourth method. The second algorithm derived from BeSS is the Sequential primal-dual active set (SPDAS) that attains the minimum Generalized Information Criterion [77]. SPDAS turns out to be our fifth method. Finally, the third BeSS derived strategy uses the same SPDAS algorithm but this time adjusted by BIC.

**Filters.** Of these three packages, the three default algorithms were selected as a feature selection algorithm for the filters section. BeSS uses GSPDAS, GLMNET uses LASSO and FRESA.CAD uses BSWiMS to find the features that will build a unique Cox model for each model. These three filters are combined with an extra filtering method, Univariate Cox Analysis. UniCox uses a certain threshold to choose the characteristics with a p-value lower than the configured one (Default value:  $p < 0.2$ ). Table 3.9 details the default Cox Benchmarking algorithms.

#	Algorithm	Source Package	Type
1	BSWiMS	<i>FRESA.CAD</i>	Wrapper
2	Cox with BSWiMS	<i>Survival</i> <i>and FRESA.CAD</i>	Filter
3	LASSO	<i>GLMNET (Coxnet)</i>	Wrapper
4	RIDGE	<i>GLMNET (Coxnet)</i>	Wrapper
5	ELASTICNET	<i>GLMNET (Coxnet)</i>	Wrapper
6	Cox with LASSO	<i>GLMNET (Coxnet)</i>	-Filter
7	GSPDAS	<i>BeSS</i>	Wrapper
8	SPDAS	<i>BeSS</i>	Wrapper
9	SPDAS.BIC	<i>BeSS</i>	Wrapper
10	Cox with GSPDAS	<i>BeSS,</i> <i>Survival</i>	Filter
11	Univariate Cox	<i>Survival</i>	Filter

Table 3.9: Default algorithms used in CoxBenchmarking method.

**Algorithm 4** Cox Benchmarking Algorithm

---

```

1: procedure COXBENCHMARKING(Data, Outcome, Reps, TrainFraction, Reference)
2:   if Reference is null then
3:     Reference  $\leftarrow$  RHOCV(BSWiMS)
4:     referenceTrainSampleSets  $\leftarrow$  Reference.trainSampleSets
5:   else
6:     referenceTrainSampleSets  $\leftarrow$  CalculateStatsForReference(Reference)
7:   end if
8:   SurvivalStatsTable  $\leftarrow$  []
9:   ClassificationStatsTable  $\leftarrow$  []
10:  DefaultWrappers  $\leftarrow$  [Reference, "LASSO", "RIDGE", "ELASTICNET",
11:  "GSPDAS", "SPDAS", "SPDAS.BIC"]
12:  DefaultFilters  $\leftarrow$  ["Cox.Reference", "Cox.LASSO", "Cox.GSPDAS",
13:  "Cox.UnivariteCox"]
14:  for  $i \leftarrow 1, \text{len}(\text{Wrappers})$  do
15:    model  $\leftarrow$  RHOCV(Wrappers[i])
16:    SurvivalStatsTable[i]  $\leftarrow$  CalculateSurvivalStats(model)
17:    ClassificationStatsTable[i]  $\leftarrow$  CalculateClassificationStats(model)
18:  end for
19:  for  $i \leftarrow 1, \text{len}(\text{Filters})$  do
20:    model  $\leftarrow$  RHOCV(Filters[i])
21:    SurvivalStatsTable[i]  $\leftarrow$  CalculateSurvivalStats(model)
22:    ClassificationStatsTable[i]  $\leftarrow$  CalculateClassificationStats(model)
23:  end for
24:  return  $\leftarrow$  (SurvivalStatsTable, ClassificationStatsTable)
25: end procedure

```

---

### 3.2.2 Cox Benchmarking Algorithm

This method follows a specific flow for each of the strategies that will be compared. The method allows comparing a set of methods or a single one as a reference, compared to the default methods described above. If a set or reference element is not defined, BSWiMS will be used by default as the reference method for comparison. Cox Benchmarking follows the following algorithm:

---

#### Algorithm 5 Stats Algorithms for CoxBenchmarking

---

```

1: procedure CALCULATESURVIVALSTATS(Predictions, Plotname = "")
2:   CIRisks ← CalculateC-IndexRisks(Predictions["Outcome","RisksMedian"])
3:   CIFollowUp ← CalculateC-IndexFollowUp (Predictions [ "Outcome", "FollowUpTimesMedian"])
4:   LogRankTest ← CalculateLogRank(Predictions["Outcome","TimeToEvent", "Risks"])
5:   if plotname is not null then Curves ← PlotKM(preds >
   median(Predictions["Median"]))
6:   end if
7:   Return ← CIRisks, CIFollowUp, LogRankTest, Curves
8: end procedure
9: procedure CALCULATECLASSIFICATIONSTATS(Predictions, Outcome)
10:  PredsForBinary ← Predictions["Outcome", "LinearPredictors"]
11:  Return ← FRESA.CAD.predictionStatsbinary(PredsForBinary)
12: end procedure

```

---

The Cox Benchmarking method is responsible for executing an RHOCV for each of the algorithms that need to be compared. The reference method or methods define the data set to be used as a train and test for all models. The training fraction is what defines the percentage of patients that will be used for each set. The fact of allowing a list of algorithms as a reference gives the user the freedom to call more RHOCV with other methods to include them in the Cox Benchmarking process. The medians of the different survival and classification predictions generated in each iteration of the cross-validation of each method are calculated and returned by the standardized output of the RHOCV from FRESA.CAD. The predictions for each of the test subjects are used to calculate statistics in both contexts and are accompanied by 95% confidence intervals.

### 3.2.3 Random Holdout Cross Validation implementation

All these strategies also require the use of processes that allow fair evaluation among them and provide statistics that are reliable in order to draw conclusions. Taking this situation into account, the implementation makes use of RHOCV. The RHOCV strategy was implemented as an extension of FRESA.CAD R package <https://github.com/joseTamezPena/FRESA.CAD>. The current development version of this package with the CoxBenchmarking

and RandomCV implementation can be found at:<sup>1</sup> The main aim of this thesis is the comprehensive evaluation of different ML approaches that return or select “optimal” Cox models. We used repeated holdout cross-validation (RHOCV), for the evaluation of different ML strategies. The test results of the RHOCV were used to compare and explore the performance of the machine learning alternatives.

The strategy divided the data into random training and testing sets with a user-supplied train fraction. The training set was used for model selection, while the holdout set was used to validate the trained method [105]. The RHOCV for survival analysis requires the calculation of corresponding statistics for survival models, in this case: the RHOCV implementation used the R package Survival to calculate the final Cox predictions of each selected model. Cox predictions returned the linear predictions, the risk, and the expected follow-up times (FT) and implemented execution times, the Jaccard index, the model size, and the training and testing samples of every single method. A detailed summary of the stats that RHOCV calculate is found below. Stats The Jaccard index (JI) computed the average similarity between the selected features between models, and can be written as:

$$J = \frac{2}{(R^2 - 2R)} \sum_{i=(j+1)}^R \sum_{j=1}^{R-1} \frac{|A_i \cap A_j|}{|A_i \cup A_j|}, \quad (3.4)$$

where  $R$  is the number of holdout repeats, and  $A_j$  is the set of the  $k$  selected features for the Cox model of the  $j$  holdout training sample. The range of the index varies from 0 to 1, where 1 represents that the feature selection method always selects the same set of features on each repetition. The R implementation also reported summary statistics of the test results. The Cox-fitted coefficients  $\beta^j$  on each training set  $T_j$  were used to get the linear predictions  $f_i^j$  of the holdout set  $T_j^c$  at each repetition:

$$f_i^j = X_i \cdot \beta^j, \quad \forall i \in T_j^c \text{ and } \forall X_i \in A_j, \quad (3.5)$$

Once all the test predictions were obtained for each repetition, the testing results were summarized by computing the median prediction of each subject:  $\tilde{f}_i = \text{median}(\{f_i^1, \dots, f_i^R\})$ . The median prediction was used to divide the groups into: High-risk (HR:  $\tilde{f} \geq 0$ ) vs Low-risk (LR:  $\tilde{f} < 0$ ). The receiver operating characteristic (ROC) plots and their area under the curve (AUC) with their corresponding 95% confidence intervals (95%CI) were computed for the median prediction using the pROC package [91]. Accuracy (ACC), sensitivity (SEN), and specificity (SPE) describing the ability of the Cox models to predict censored vs uncensored subjects were computed based on the number of true positives (TP), and true negatives (TN).

$$TP = \left| (\tilde{f} \geq 0) \cap \text{uncensored} \right|, \quad (3.6)$$

$$TN = \left| (\tilde{f} < 0) \cap \text{censored} \right|, \quad (3.7)$$

$$ACC = \frac{TP + TN}{|\text{uncensored} + \text{censored}|}, \quad (3.8)$$

<sup>1</sup><https://github.com/joseTamezPena/FRESA.CAD>

$$SEN = \frac{TP}{|uncensored|}, \quad (3.9)$$

$$SPE = \frac{TN}{|censored|}, \quad (3.10)$$

<b>Object</b>	<b>Definition</b>
<i>errorciTable</i>	the matrix of the balanced error with the 95 CI
<i>accciTable</i>	the matrix of the classification accuracy with the 95 CI
<i>aucTable</i>	the matrix of the ROC AUC with the 95 CI
<i>senTable</i>	the matrix of the sensitivity with the 95 CI
<i>speTable</i>	the matrix of the specificity with the 95 CI
<i>errorciTable_filter</i>	the matrix of the balanced error with the 95 CI for filter methods
<i>accciTable_filter</i>	the matrix of the classification accuracy with the 95 CI for filter methods
<i>senciTable_filter</i>	the matrix of the classification sensitivity with the 95 CI for filter methods
<i>speciTable_filter</i>	the matrix of the classification specificity with the 95 CI for filter methods
<i>aucTable_filter</i>	the matrix of the ROC AUC with the 95 CI for filter methods
<i>CIRiskTable</i>	the matrix of the concordance index on Risk with the 95 CI
<i>CIFollowUpTable</i>	the matrix of the concordance index on Follow-up times with the 95 CI
<i>LogRankTable</i>	the matrix of the LogRank Test with the 95 CI
<i>CIRisksTable_filter</i>	the matrix of the concordance index on Risk with the 95 CI for the filter methods
<i>CIFollowUpTable_filter</i>	the matrix of the concordance index on Follow-up times with the 95 CI for the filter methods
<i>LogRankTable_filter</i>	the matrix of the LogRank Test with the 95 CI for the filter methods
<i>times</i>	The average CPU time used by the method
<i>jaccard_filter</i>	The average Jaccard Index of the feature selection methods
<i>TheCVEvaluations</i>	The output of the randomCV (randomCV) evaluations of the different methods
<i>testPredictions</i>	A matrix with all the test predictions
<i>featureSelectionFrequency</i>	The frequency of feature selection
<i>cpuElapsedTimes</i>	The mean elapsed times

Table 3.10: Output of Cox Benchmarking Model. Left side lists the name of the objects in the model and Right side describes it.

### 3.2.4 Cox Benchmarking Model

The Cox Benchmarking model tries to summarize each of the cross-validations in a single process. To do this, it defines a model with a certain group of characteristics that allow comparison and subsequently the graphic sample of the process. Table 3.10 defines each of the outputs of the Cox Benchmarking model and the definition of each of them. The Cox function between its outputs is responsible for generating the KM curves and the ROC curves for each method to be compared.

#### CoxBenchmarking Plot

To summarize the output of this model, a function that generates graphs for the different statistics was implemented. The function tries to graphically compare each of the methods together with their confidence intervals, and thus be able to conclude statistical differences between each of the models. All the barplots were implemented in R for this thesis. The implementation is based on the Benchmarking techniques of FRESA.CAD. This function provides the following graphs:

- Barplot classification stats: SEN, SPE, ACC, AUC.
- Barplot survival stats: C-Index Follow up times, C-Index Risks, Log-Rank p-value.

The figure 3.5 shows an example of the CoxBenchmarking plot function result.

## 3.3 Summary

In this chapter we described all the tools and information used to accomplish the thesis objective. Mainly focused on two things, the implementation of the Benchmarking algorithm for the analysis of methods that work with the Cox Model; and second, the clinical and simulated information used to evaluate the behavior of the CoxBenchmarking function and its importance. In the following paragraphs, we will present a brief summary of each of the sections, which will allow us to remember each of the important points before detailing the results of each experiment and understanding the conclusions and discussions of them.

### 3.3.1 Experiments: Data acquisition, preparation and analysis

This section detailed all the points corresponding to the data used, the origin, the version, the necessary changes in the data to be able to use them with our strategy, the calculations of derived data to be able to find information that serves to make clinical decisions, among others issues. In this summary, the origin of the data and the main objective of each of the experiments will be briefly explained. In addition, the main changes and numbers of characteristics and subjects used will be put into context. For more information on each experiment, each subsection itself can be found in this chapter.

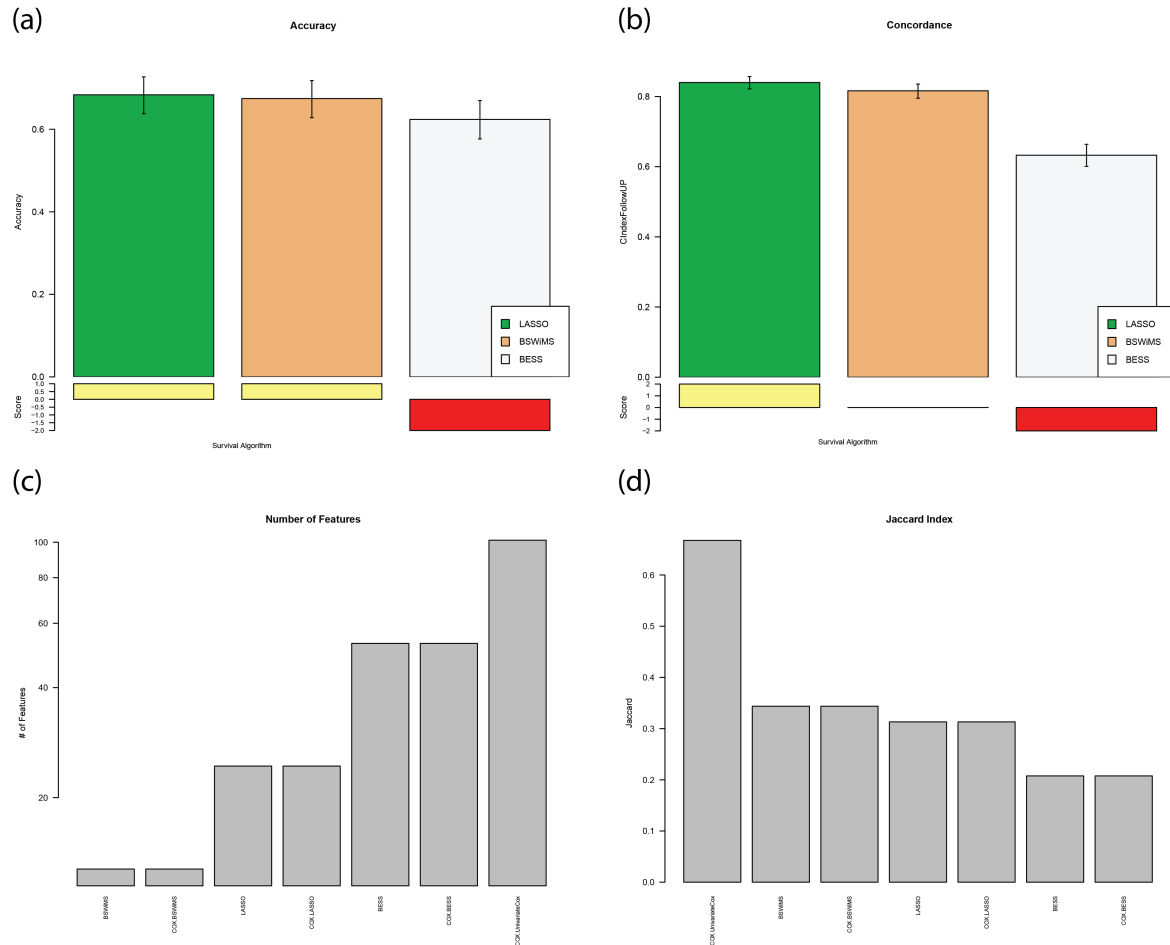


Figure 3.5: Example of CoxBenchmarking plot result. In the figure we can see (a) Accuracy barplot for accuracy of some model (b) Concordance Index Follow-up Times barplot (c) Number of features selected by the default Cox Benchmarking methods (d) Jaccard Index selected by the default Cox Benchmarking methods

**Simulation data** The experiment with simulated information aims to make use of the ground truth to test the results of the CoxBenchmarking method. In this case, it was decided to take the simulation of data from 1000 possible NBA players as true. Descriptive data of his physique and statistics of past games were generated, which would allow predicting the number of seasons that will be part of the league. A total of 11 characteristics were created and each of them was assigned a weight within the survival equation. Depending on the experiment, that weight changed and affected the status of the game. The time to event was generated randomly, taking into account the probability of survival calculated from the weights of the characteristics and their measurements for each subject.

The experiment was divided into two. The first of them uses only 4 simulated characteristics to generate the probability of survival and the second of 11. The two experiments have three sub-experiments, giving a total of 6 mini-experiments within the simulation experiment. The first mini-experiment adds the number of random characteristics equal or similar to that of real characteristics (4/11 real features - 4/10 random features). The second one,



tries to complete the characteristics to the 100 available ones, that is to say in the experiment of 4 characteristics it adds 96 random columns and to the one of 11, it adds 90. The last mini-experiment simulates 1000 characteristics so it consequently adds 996 features and 990.

**TADPOLE/ADNI** The ADNI experiment consists of two sub experiments that make use of the same subjects. With information from the ADNI initiative specifically that which comes from the TADPOLE challenge, we use information from 442 patients who underwent MRI screening and have their available. From them, we extracted information corresponding to CSF Measures, qMRI, APOE, clinical and demographical information, and lastly Cognitive Assessments. The first experiment's main objective is to test the capacity of qMRI features combined with the APOE factor to predict the conversion between MCI and AD. This conversion is currently explained through some factors we want to know how much information will these radiomics features provide to the conversion. Regarding the results, we decided to assess once more how the other available information which is already used to screen the conversion, so with that compare and measure the importance of the qMRI feature versus the clinical real used factors. For this, 7 subsections are developed, each section builds models for each type of existing data and their combination with the image information.

**Osteoarthritis Initiative** With information from the OAI initiative, the experiment with Osteoarthritis data aims to make survival analysis on the total knee replacement event (TKR). For this, all the information of x-rays images available in OAI is used and it is combined with the information of screening forms of each knee. A total of 4507 patients are used for the study. The set was divided into 1000 patients for a train set, and the remaining for the test. Of those 1000, 70% of the events were used within the 1000 patients and the remaining 30% belong to the test set. Data were derived from the image information to have more informative features of the outcome.

**Prognostic Wisconsin Breast Cancer Database** The study with breast cancer information from Wisconsin tries to find the prediction of the recurrence time of patients suffering from the disease. For this, 198 patients with 34 clinical characteristics that come from cell measurements and tumor features were taken into account. This experiment was performed with the information as it is found publicly.

**Prognostic San Jose Hospital Breast Cancer Database** Finally, within the summary of the data are the BRCA data of the San José hospital. As an extra study for the one that was made a few years ago by Tamez et al. This experiment adds medical tracking information (survival) to the image information extracted in that work. The information and time at which the patient's recurrence was determined are added to each patient. The CoxBenchmarking test is carried out with information of 73 and 1091 variables.

### 3.3.2 Cox benchmarking

This section describes the CoxBenchmarking algorithm and the functions used within it. The default methods used in each benchmarking use are described and the characteristics of both

input and output are detailed. In its description are the details of the statistics calculated for comparison, as well as the implementation of the graphical functions that allow the summary of the results in a simpler way.

First of all, describe the benchmarking algorithm that begins with one or more reference methods. The reference must be an object or several that are compatible with the Random Holdout Cross-Validation output detailed below. The CoxBenchmarking method is responsible for creating the structures where you will organize and then report each of the statistics. After having the statistics, proceed to perform the same process for each of the methods that the method has by default.

**Random Holdout Cross Validation** All these strategies also require the use of processes that allow a fair evaluation between them and provide reliable statistics to draw conclusions. Given this situation, the implementation makes use of RHOCV. The RHOCV strategy was implemented as an extension of the FRESA.CAD R. The main objective of this thesis is the comprehensive evaluation of different LD approaches that return or select "optimal" Cox models. We use Cross-validation of repeated waiting (RHOCV), for the evaluation of different LD strategies. The results of the RHOCV test were used to compare and explore the performance of machine learning alternatives.

# Chapter 4

## Experiment Results

This chapter states all the experiment results and tries to summarize all the stats generated by showing data numerically and graphically. Taking into consideration the data explained in the last chapter, the first experiment made was the analysis of survival time of the NBA data simulation. In all the internal experiments belonging to Experiment I, all the methods found the correct features, but some selected some random variables together with the 4 or 10 generated variables, which really do have relation to the outcome. Following, results on two experiments with TADPOLE/ADNI data are detailed. These two experiments were the first to use the CoxBenchmarking method clinically, Experiment II and III. Experiment II just took into consideration the image data information with descriptive clinical information of each patient to evaluate the capacity of qMRI features to predict the conversion between two stages of Alzheimer's disease. Experiment III unlike the previous experiment studied the improvement of these variables against the information provided by the characteristics of CSF Measures and the numerical information obtained through clinical forms that are used to diagnose the disease. Both results were presented at international conferences and the clinical information obtained was discussed with the help of health professionals. Then, another clinical analysis was developed, this time with information concerning Osteoarthritis data. This experiment is detailed as Experiment IV. On this occasion, the results obtained by submitting the information processed by OAI are shown. Finally, two data sets with information about breast cancer are reported. Experiment V analyzed the BRCA prognostic information provided by the University of Wisconsin. previous experiments with this data found that Cox analysis fails to determine a separation between high and low-risk patients, so our experiment got the same result. The last of experiments, Experiment VI, is performed with breast cancer information from the San José Hospital. All these experiments used information treated for the development of this thesis. Each experiment used a similar methodology; however, different tools are used for each experiment to report the results as appropriate.

### 4.1 Simulation data

For the simulated data experiment, 6 different experiments were run in two sections. The first considered only 4 real variables related to the result and the second 10 of them. First we add the same number of real characteristics to each dataframe. 4 in the first section and 10

in the second one. In both cases, the necessary characteristics were added to complete 100 and 1000 variables in the dataframe were used as the next experiments. The results for these experiments are detailed below. In all the cases, the number of iterations were 20 and the train fraction was 0.7. For the generation of these survival times, a censorship rate of 20% was used. If the random number of censorship generated was lower than the rate (0.2), then the subject was censored at that simulated time.

### 4.1.1 4 variables

In this case just 4 real variables were considered for the survival function estimation. Offensive Rating (ORTg), Defensive Rating (DRtg), Injuries (I) and surgeries (S) were used, two Normal distributed variables (ORTg, DRtg) and two binomial distributed variables (I,S). The status and the event time were simulated with an R script that considered the hazard and survival probability of each subject in each period of time. A set of probabilities for each subject were simulated with a Normal distribution. If that probability was lower than the survival probability the event happened and the period of time was assigned as the event time. In this case, the data simulation set had 573 subjects with an event and 427 censored throughout the study.

### 8 features (4 real - 4 random)

This was the most simple experiment with a data simulation dataset. A total of 8 characteristics were analyzed by the CoxBenchmarking tool. As expected, all the methods work very well on this kind of data. Seven model built with wrapper methods and four models with filters. Table 4.1 summaries the stats of the wrapper methods. The filter section found the same stats on all the methods, differences were found in the marginal decimal places; so, their results will be reported as one stat. Almost all the methods found the same stats in survival and classification stats. This is the case of the area under the curve which in all the models was the same  $AUC = 0.74(0.71,0.77)$ . The main difference between them was the number and the features selected.

The model I built with BSWiMS selected a mean of 2.90 features with a Jaccard index of 0.93. Two of the Four real features were selected all the times, Injuries and Surgeries. The offensive rating was selected in 18 iterations. Defensive Rating was never selected such as all the four random variables. BSWiMS got  $ACC = 0.68(0.65,0.71)$ ,  $SEN = 0.68(0.64,0.72)$ ,  $SPE = 0.68(0.64,0.72)$ ,  $C\text{-Index Risks} = 0.7(0.69,0.72)$ . The second model uses LASSO strategy. It found an average of 6.55 features by a model with a Jaccard index of 6.55. The same three features were selected all the time, and the Defensive rating was selected in 90% of the iterations, but this time all the random variables were used in the model in at least half of the iterations. Model II got  $ACC = 0.68(0.65,0.71)$ ,  $SEN = 0.68(0.64,0.72)$ ,  $SPE = 0.68(0.63,0.72)$ ,  $C\text{-Index Risks} = 0.71(0.69,0.72)$ .

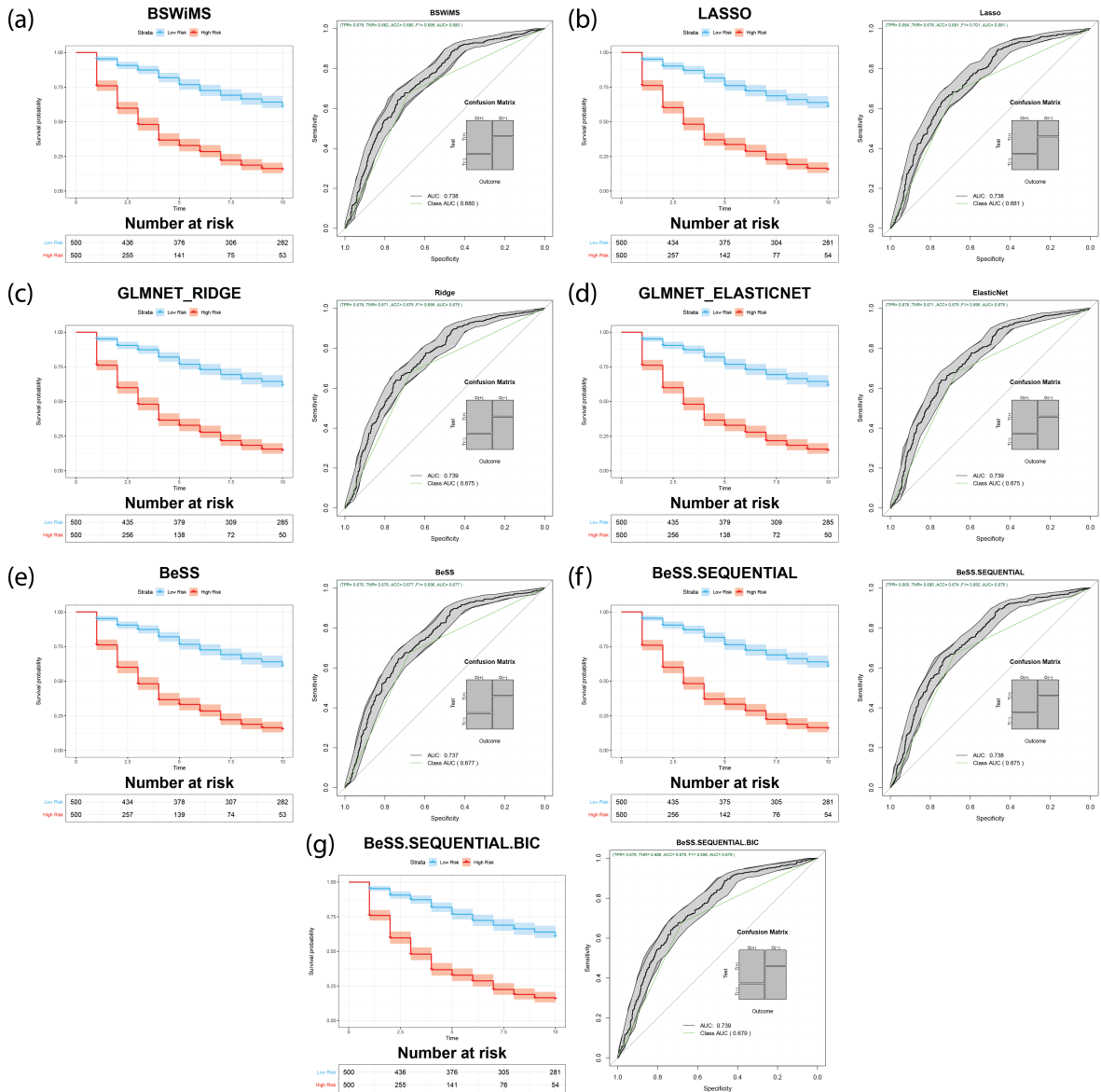


Figure 4.1: KM and ROC Curves for the Simulation experiment of 8 features (4 real - 4 random features) (a) BSWiMS (b) LASSO (c) RIDGE (d) ELASTICNET (e) GSPDAS (BeSS) (f) SPDAS (g) SPDAS with BIC

Method	ACC (95% CI)	SEN (95% CI)	SPE (95% CI)	C-index Risks (95% CI)	C-Index FU (95% CI)
I	0.68 (0.65,0.71)	0.68 (0.64,0.72)	0.68 (0.64,0.72)	<b>0.7</b> <b>(0.69,0.72)</b>	0.58 (0.56,0.61)
II	0.68 (0.65,0.71)	0.68 (0.64,0.72)	0.68 (0.63,0.72)	0.71 (0.69,0.72)	0.62 (0.6,0.64)
III	0.68 (0.64,0.7)	0.68 (0.64,0.72)	<b>0.67</b> <b>(0.63,0.71)</b>	0.71 (0.69,0.72)	0.64 (0.62,0.66)
IV	0.68 (0.64,0.7)	0.68 (0.64,0.72)	<b>0.67</b> <b>(0.63,0.71)</b>	0.71 (0.69,0.72)	0.64 (0.62,0.66)
V	0.68 (0.65,0.71)	0.68 (0.64,0.72)	0.68 (0.63,0.72)	<b>0.7</b> <b>(0.69,0.72)</b>	<b>0.58</b> <b>(0.56,0.6)</b>
VI	<b>0.67</b> <b>(0.64,0.7)</b>	<b>0.67</b> <b>(0.63,0.71)</b>	0.68 (0.63,0.72)	<b>0.7</b> <b>(0.69,0.72)</b>	<b>0.58</b> <b>(0.56,0.6)</b>
VII	0.68 (0.65,0.71)	0.68 (0.64,0.72)	0.68 (0.63,0.72)	<b>0.7</b> <b>(0.69,0.72)</b>	<b>0.58</b> <b>(0.56,0.6)</b>

Table 4.1: Classification and survival stats for wrapper methods in the Simulation experiment of 8 features (4 real - 4 random). I = BSWiMS, II = LASSO, III = RIDGE, IV = ELASTICNET, V = GSPDAS (BESS), VI = SPDAS (BESS.SEQUENTIAL), VII = SPDAS.BIC (BESS.SEQUENTIAL.BIC). Worst scores for each stat are bolded.

The third model was constructed with Ridge strategy. As is well-known the RIDGE strategy selects a big number of features. This time, it chose a mean of 7.90 features with a Jaccard index of 0.97. Those numbers indicate that all the features were used, just a random variable was ignored in two iterations. Model III got ACC =0.68(0.64,0.7), SEN =0.68(0.64,0.72), SPE=0.67(0.63,0.71), C-Index Risks=0.71(0.69,0.72). Model IV built with the ELASTICNET method found the same results as the last method.

Model V uses the GSPDAS algorithm. It used an average of 5.30 features in the models and its Jaccard index is 0.74. The four real features were selected every time. One of the random columns were used in the 50% of the iterations and the other three is less than 10 iterations. GSPDAS resulted in an ACC =0.68(0.65,0.71), SEN =0.68(0.64,0.72), SPE=0.68(0.63,0.72), C-Index Risks=0.7(0.69,0.72). SPDAS algorithm developed Model VI. It used a mean of 3.40 features with a Jaccard index of 0.85. As happened with BSWiMS, of the four real features, just the defensive rating was not selected in all the iterations. But this time, it was selected in more than 5 iterations. A random variable was used just in 10% of the models. Model VI got ACC =0.67(0.64,0.7), SEN =0.67(0.63,0.71), SPE=0.68(0.63,0.72), C-Index Risks=0.7(0.69,0.72). Finally, Model VII built with SPDAS with Bayesian Information Criterion used an average of 3.05 features to build the models. Its Jaccard index is 0.975. In this case, a Defensive rating was just used in one model out of 20 possible. The other real features were selected always. This Model VII reported ACC =0.68(0.65,0.71), SEN =0.68(0.64,0.72), SPE=0.68(0.63,0.72), C-Index Risks=0.7(0.69,0.72).

**100 features (4 real - 96 random)**

The second experiment uses the same 4 real variables and added 96 random variables. 46 random variables are binomially distributed and the remaining ones are normally distributed. The first model uses BSWiMS. Unlike the first experiment, BSWiMS selected less number of features on average (2.55 vs. 2.90). Its Jaccard Index is also lower at 0.83. This model chose Lesions and Surgeries in all the iterations and Offensive rating just in the half of the iterations. All the 96 random variables were ignored. Model I got AUC = 0.73(0.7, 0.77), ACC = 0.68 (0.65, 0.7), SEN = 0.67 (0.63, 0.71), SPE = 0.69 (0.64, 0.73), CI-Risks = 0.7 (0.69, 0.72) and CIFU = 0.58 (0.56, 0.61).

Method	ACC (95% CI)	SEN (95% CI)	SPE (95% CI)	C-index Risks (95% CI)	C-Index FU (95% CI)
I	0.68 (0.65,0.7)	0.67 (0.63,0.71)	0.69 (0.64,0.73)	0.7 (0.69,0.72)	0.58 (0.56,0.61)
II	0.68 (0.65,0.71)	0.68 (0.64,0.72)	0.68 (0.63,0.72)	0.7 (0.69,0.72)	0.7 (0.68,0.72)
III	<b>0.66</b> <b>(0.63,0.69)</b>	<b>0.65</b> <b>(0.61,0.69)</b>	<b>0.66</b> <b>(0.62,0.71)</b>	<b>0.69</b> <b>(0.67,0.7)</b>	0.77 (0.75,0.79)
IV	<b>0.66</b> <b>(0.63,0.69)</b>	<b>0.65</b> <b>(0.61,0.69)</b>	<b>0.66</b> <b>(0.62,0.71)</b>	0.69 (0.67,0.71)	0.78 (0.76,0.79)
V	0.67 (0.64,0.7)	0.66 (0.62,0.7)	0.67 (0.62,0.71)	0.69 (0.68,0.71)	<b>0.57</b> <b>(0.55,0.59)</b>
VI	0.68 (0.65,0.71)	0.68 (0.64,0.72)	0.68 (0.63,0.72)	0.7 (0.69,0.72)	0.58 (0.56,0.6)
VII	0.68 (0.65,0.7)	0.68 (0.64,0.72)	0.67 (0.62,0.71)	0.71 (0.69,0.72)	0.58 (0.56,0.6)

Table 4.2: Classification and survival stats for wrapper methods in the Simulation experiment of 100 features (4 real - 96 random). I = BSWiMS, II = LASSO, III = RIDGE, IV = ELASTICNET, V = GSPDAS (BESS), VI = SPDAS (BESS.SEQUENTIAL), VII = SPDAS.BIC (BESS.SEQUENTIAL.BIC). Worst scores for each stat are bolded.

Method	ACC (95% CI)	SEN (95% CI)	SPE (95% CI)	C-index Risks (95% CI)	C-Index FU (95% CI)
I	0.68 (0.65,0.7)	0.67 (0.63,0.71)	<b>0.69</b> <b>(0.64,0.73)</b>	0.58 (0.56,0.61)	0.58 (0.56,0.61)
II	0.67 (0.64,0.7)	<b>0.68</b> <b>(0.64,0.72)</b>	0.66 (0.61,0.7)	0.58 (0.56,0.61)	0.58 (0.56,0.61)
III	0.67 (0.64,0.7)	0.66 (0.62,0.7)	0.67 (0.62,0.71)	0.57 (0.55,0.59)	0.57 (0.55,0.59)
IV	<b>0.68</b> <b>(0.65,0.71)</b>	0.67 (0.63,0.71)	0.68 (0.64,0.73)	<b>0.59</b> <b>(0.56,0.61)</b>	<b>0.59</b> <b>(0.56,0.61)</b>

Table 4.3: Classification and survival stats for filter methods with 4 real features and 96 random variables. I = Cox with BSWiMS, II = Cox with LASSO, III = Cox with BESS IV = Univariate Cox. Best scores for each stat are bolded.

Model II was developed with LASSO. LASSO selected a mean of 9.90 features with a Jaccard Index 0.37. Lesions, surgeries and Offensive Rating were selected in every iteration, 2 random variables were selected in more than half of iterations and Defensive Rating was just used in two models. A total of 47 features were used to build the model, but just 4 were used in at least half of the models. The second model got AUC =0.73(0.7,0.77), ACC =0.68(0.65,0.71), SEN =0.68(0.64,0.72), SPE=0.68(0.63,0.72), CI-Risks=0.7(0.69,0.72) and CIFU =0.7(0.68,0.72). The model III and IV got the same results in the stats once again. But the number of features selected is different. RIDGE method selected 96.90 features and ELASTICNET 97.10. Their Jaccard Index is the same as 0.98. All the features were used but just 65 were used in all the iterations. Model V used the BeSS method with the GSPDAS algorithm. BeSS selected 56.25 features with 0.48 of the Jaccard Index. The four real features were selected in all the iterations with the other 5 random variables. GSPDAS also used all the features but just 59 were used in more than half of the models. GSPDAS reported an AUC = 0.72(0.69,0.75), ACC = 0.67(0.64,0.7), SEN = 0.66(0.62,0.7), SPE = 0.67 (0.62,0.71), CI-Risks = 0.69(0.68,0.71) and CIFU = 0.57 (0.55,0.59). Model VI uses SPDAS and selected 3.25 features. Its Jaccard Index is 0.48. Lesions and Surgeries were chosen in all the models, the Offensive Rating was selected 18 times. 4 random features were also selected but all in less than 3 iterations. The defensive rating was never selected. Model VI got AUC = 0.74 (0.7,0.77), ACC =0.68(0.65,0.71), SEN = 0.68 (0.64,0.72), SPE = 0.68 (0.63,0.72), CI-Risks = 0.7(0.69,0.72) and CIFU = 0.58 (0.56,0.6). Model VII uses SPDAS with BIC. It selected 4.40 features on average with a Jaccard Index of 0.68. It chose the three real features in all the models, defensive rating in one iteration and other nine random features selected ranging between 0.45%-0.05%.



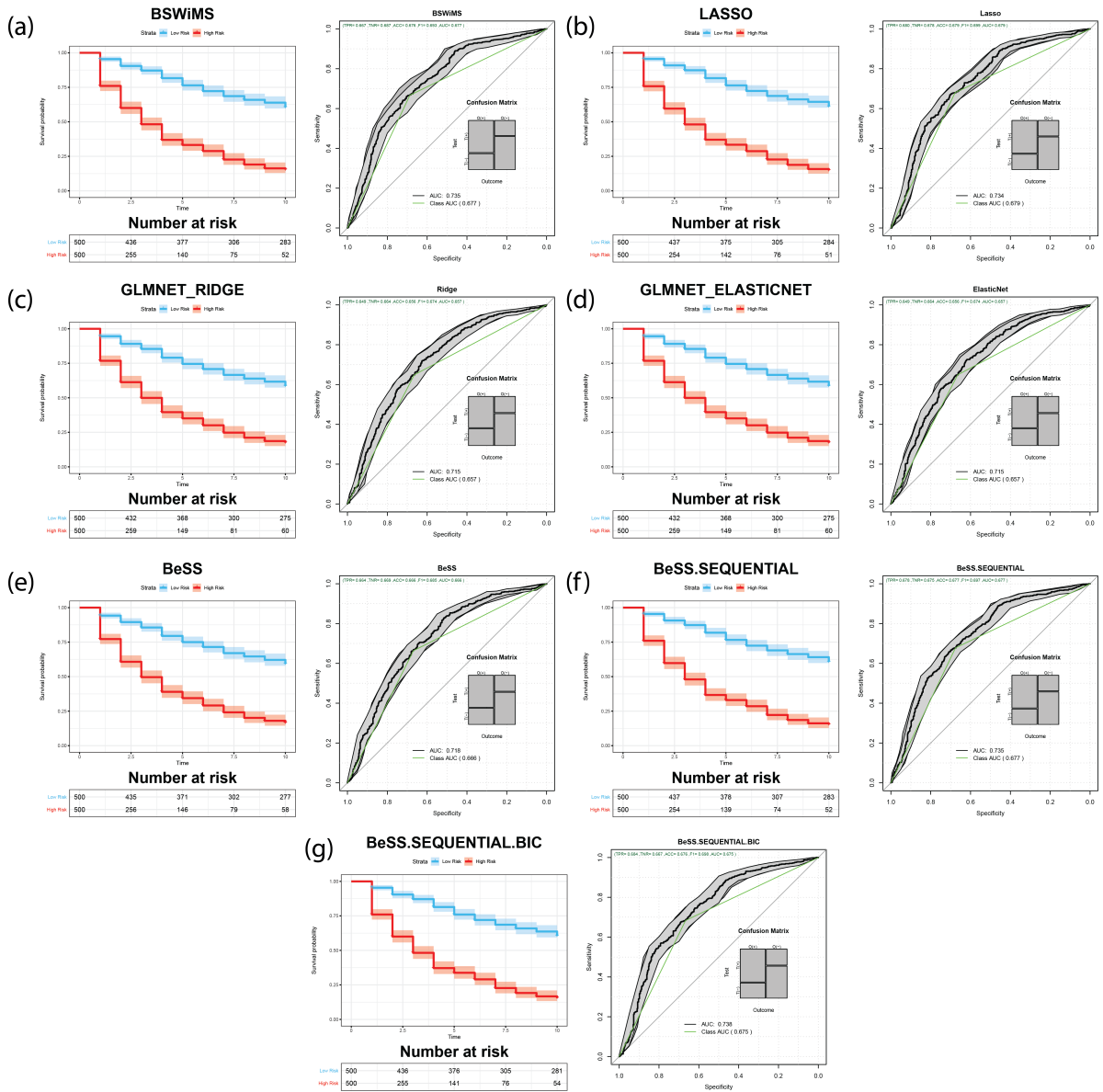


Figure 4.2: KM and ROC Curves for the Simulation experiment of 100 features (4 real - 96 random features) (a) BSWiMS (b) LASSO (c) RIDGE (d) ELASTICNET (e) GSPDAS (BeSS) (f) SPDAS (g) SPDAS with BIC

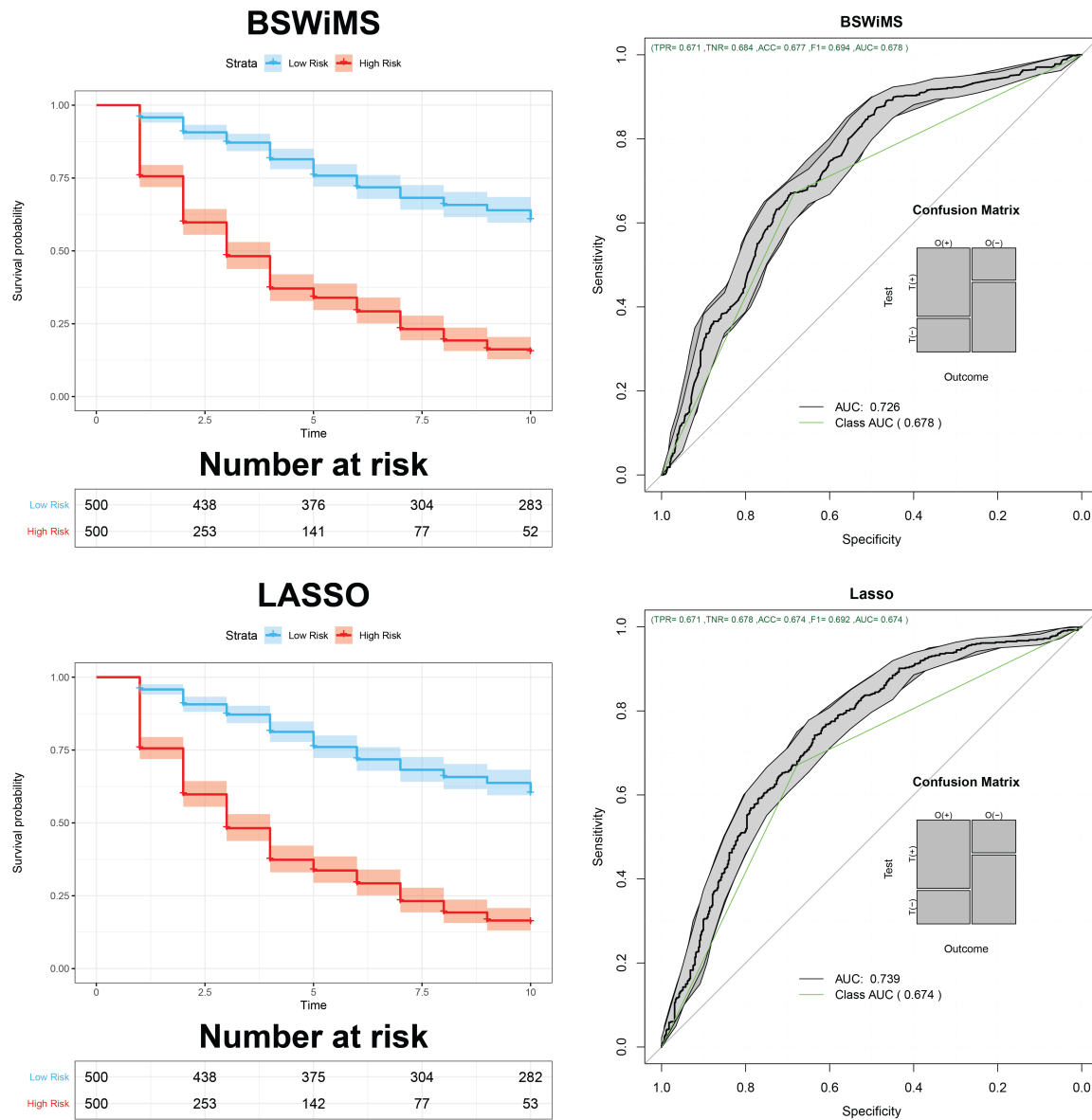


Figure 4.3: KM and ROC Curves for wrappers methods with the 4 real features and 996 random created with (a) BSWiMS (b) LASSO (c) RIDGE (d) ELASTICNET (e) GSPDAS (BeSS) (f) SPDAS (g) SPDAS with BIC

### 1000 features (4 real - 996 random)

In this case, the number of significant features is too low. Just two methods were able to fit a model. BSWiMS and LASSO. The other methods could not finish all the iterations, which makes it impossible to determine a final model through RHOCV. We will only report the results of the models that could finish the iterations.

BSWiMS selected just 3 real features, but just two in all the cases. Lesions and surgeries were the only features used in all the iterations. The offensive rating was used just in two iterations. Two random variables were also used but just 3 times accumulated. The other real

features were ignored. On average it used just 2.25 features to build the model with a Jaccard Index of 0.84. BSWiMS take a mean of 0.96635 seconds to create the model on each iteration. Model I got ACC = 0.677 (0.647, 0.706), AUC = 0.73 (0.69, 0.76), SEN = 0.67 (0.63, 0.71), SPE = 0.68 (0.64, 0.73), CIRisks = 0.70 (0.68, 0.71) and CIFU = 0.58 (0.56, 0.60).

The other method that works was LASSO. In this case, it selected a mean of 25.5 features and its Jaccard Index was very low at 0.24. The only features selected in all the iterations were real: Lesions, Offensive rating and Surgeries. Ten random features were used in more than half of the models. This time, LASSO took 38.86 seconds on average to build the model. LASSO got ACC = 0.67 (0.64, 0.70), AUC = 0.74 (0.71, 0.77), SEN = 0.68 (0.63, 0.71), SPE = 0.68 (0.63, 0.72), CIRisks = 0.71 (0.69, 0.72) and CIFU = 0.75 (0.73, 0.76).

#### 4.1.2 11 variables

The second section simulate the survival information with 11 outcome related features. 7 Normal distributed features were added to the four already described: Body mass index (BMI), Age (age), Games played (games), Average minutes played in the season before (minutes), Mean of assists (AST), Field Goal Percentage (FGP) and Blocks per game (BPG). The status and the event time were simulated with an R script that considered the hazard and survival probability of each subject in each period of time. Using the 11 real features we create the probability of each player to survive a period of time. The simulation tuns a random number to decide if the player retire or keeps playing in the season or if he had to be censored. 450 players have the event and the 550 remaining are censored.

Method	ACC (95% CI)	AUC (95% CI)	SEN (95% CI)	SPE (95% CI)	C-index Risks (95% CI)	C-Index FU (95% CI)
I	0.71 (0.68,0.74)	0.79 (0.76,0.82)	0.71 (0.67,0.75)	0.71 (0.66,0.75)	0.67 (0.65,0.69)	0.47 (0.44,0.5)
II	0.72 (0.69,0.74)	0.79 (0.77,0.82)	<b>0.72</b> <b>(0.68,0.76)</b>	0.71 (0.66,0.75)	0.67 (0.66,0.69)	<b>0.53</b> <b>(0.5,0.56)</b>
III	0.71 (0.68,0.74)	0.79 (0.76,0.82)	0.71 (0.67,0.75)	0.71 (0.66,0.75)	0.67 (0.66,0.69)	0.53 (0.5,0.55)
IV	0.71 (0.68,0.74)	0.79 (0.76,0.82)	0.71 (0.67,0.75)	0.71 (0.66,0.75)	0.67 (0.66,0.69)	<b>0.53</b> <b>(0.5,0.56)</b>
V	0.71 (0.68,0.74)	0.79 (0.77,0.82)	0.71 (0.67,0.75)	0.7 (0.66,0.75)	0.67 (0.66,0.69)	0.45 (0.43,0.48)
VI	<b>0.72</b> <b>(0.69,0.75)</b>	<b>0.8</b> <b>(0.77,0.82)</b>	<b>0.72</b> <b>(0.68,0.76)</b>	<b>0.71</b> <b>(0.67,0.76)</b>	0.67 (0.66,0.69)	0.46 (0.43,0.49)
VII	0.72 (0.69,0.74)	<b>0.8</b> <b>(0.77,0.82)</b>	0.72 (0.68,0.75)	<b>0.71</b> <b>(0.67,0.76)</b>	0.67 (0.66,0.69)	0.46 (0.43,0.49)

Table 4.4: Classification and survival stats for wrapper methods in the Simulation experiment of 21 features (11 real - 10 random). I = BSWiMS, II = LASSO, III = RIDGE, IV = ELASTICNET, V = GSPDAS (BESS), VI = SPDAS (BESS.SEQUENTIAL), VII = SPDAS.BIC (BESS.SEQUENTIAL.BIC). Best scores for each stat are bolded.

**21 features (11 real - 10 random)**

We added 10 random features (50% of normal and 50% binomial random distributed variables) to test the CoxBenchmarking process. The first model was developed with BSWiMS which selected a mean of 4.50 features with a Jaccard Index of 0.67. Of those selected features 3 real features were selected in all the iterations, Lesions, surgeries and Block per game. Four other real features Field goal percentage, Defensive Rating, Offensive Rating and age were selected in less than the half of the models. Games and minutes were completely ignored in the models that BSWiMS made. Model I finished with an ACC = 0.71 (0.68,0.74), AUC = 0.79 (0.76,0.82), SEN = 0.71 (0.67,0.75), SPE = 0.71 (0.66,0.75), CIRisks = 0.67 (0.65,0.69) and CIFU = 0.47 (0.44,0.5). Model II uses LASSO and it selected a mean of 12.35 features with a Jaccard Index of 0.62. 7 of the eleven real features were selected the 100% of the times. Block per game, Lesions, Surgeries, FGP, DRtg, ORtg, AGE. BMI was selected a fraction of the iterations (0.90). Assists per game were selected just 5 iterations. Games and minutes were selected in less than 4 models. All the random variables were selected ranging from (0.6-0.15). LASSO got ACC =0.72 (0.69, 0.74), AUC =0.79 (0.77, 0.82), SEN =0.72 (0.68, 0.76), SPE = 0.71 (0.66, 0.75), CIRisks=0.67 (0.66, 0.69) and CIFU = 0.53 (0.5, 0.56). Model 3 and 4 are the same models but built with different methods. Model 3 uses RIDGE and Model 4 ELASTICNET. Both models selected a mean of 20.95 (all) with a Jaccard Index of 0.995. As the numbers tell, all the features were used and just the minutes variable were not selected in one of the models. Both methods finished with ACC =0.71(0.68,0.74), AUC =0.79(0.76,0.82), SEN =0.71(0.67,0.75), SPE=0.71(0.66,0.75), CIRisks=0.67(0.66,0.69) and CIFU = 0.53 (0.5,0.55).

Method	ACC (95% CI)	SEN (95% CI)	SPE (95% CI)	C-index Risks (95% CI)	C-Index FU (95% CI)
I	0.71 (0.68,0.74)	0.71 (0.67,0.75)	<b>0.71</b> <b>(0.66,0.75)</b>	<b>0.47</b> <b>(0.44,0.5)</b>	<b>0.47</b> <b>(0.44,0.5)</b>
II	0.7 (0.68,0.73)	0.71 (0.67,0.75)	0.7 (0.65,0.74)	0.46 (0.43,0.49)	0.46 (0.43,0.49)
III	0.71 (0.68,0.74)	0.71 (0.67,0.75)	0.7 (0.66,0.75)	0.45 (0.43,0.48)	0.45 (0.43,0.48)
IV	<b>0.72</b> <b>(0.69,0.74)</b>	<b>0.72</b> <b>(0.68,0.76)</b>	<b>0.71</b> <b>(0.66,0.75)</b>	0.46 (0.43,0.48)	0.46 (0.43,0.48)

Table 4.5: Classification and survival stats for filter methods with 11 real features and 10 random variables. I = Cox with BSWiMS, II = Cox with LASSO, III = Cox with BESS IV = Univariate Cox. Best scores for each stat are bolded.

Model V which uses BeSS's GSPDAS selected a mean of 11.30 features with a Jaccard Index of 0.65. BeSS selected the same real features in the 100% of the iterations. Assists just in 25% of the models and games and minutes just in two models. GSPDAS resulted in ACC = 0.71 (0.68, 0.74), AUC = 0.79 (0.77, 0.82), SEN = 0.71 (0.67, 0.75), SPE = 0.7 (0.66, 0.75), CIRisks = 0.67 (0.66, 0.69) and CIFU = 0.45 (0.43, 0.48). SPDAS is also named as Model

VI. It selected 6.40 features on average with a Jaccard Index 0.79. Blocks per game, Lesions, Surgeries and Defensive Rating were selected in the 20 iterations. Offensive rating, field goals percentage, and age were used in more than half of the iterations and the last used feature was BMI with 5 selections. Model VI got ACC = 0.72 (0.69, 0.75), AUC = 0.8 (0.77, 0.82), SEN = 0.72 (0.68, 0.76), SPE = 0.71 (0.67, 0.76), CIRisks = 0.67 (0.66, 0.69) and CIFU = 0.46 (0.43, 0.49). Model VII uses SPDAS with BIC criterion. It selected a mean of 6.20 features on average with a Jaccard Index of 0.76. Three real features were selected in all the iterations. All the other real features were used in more than half of the iterations excluding games and minutes which were ignored.

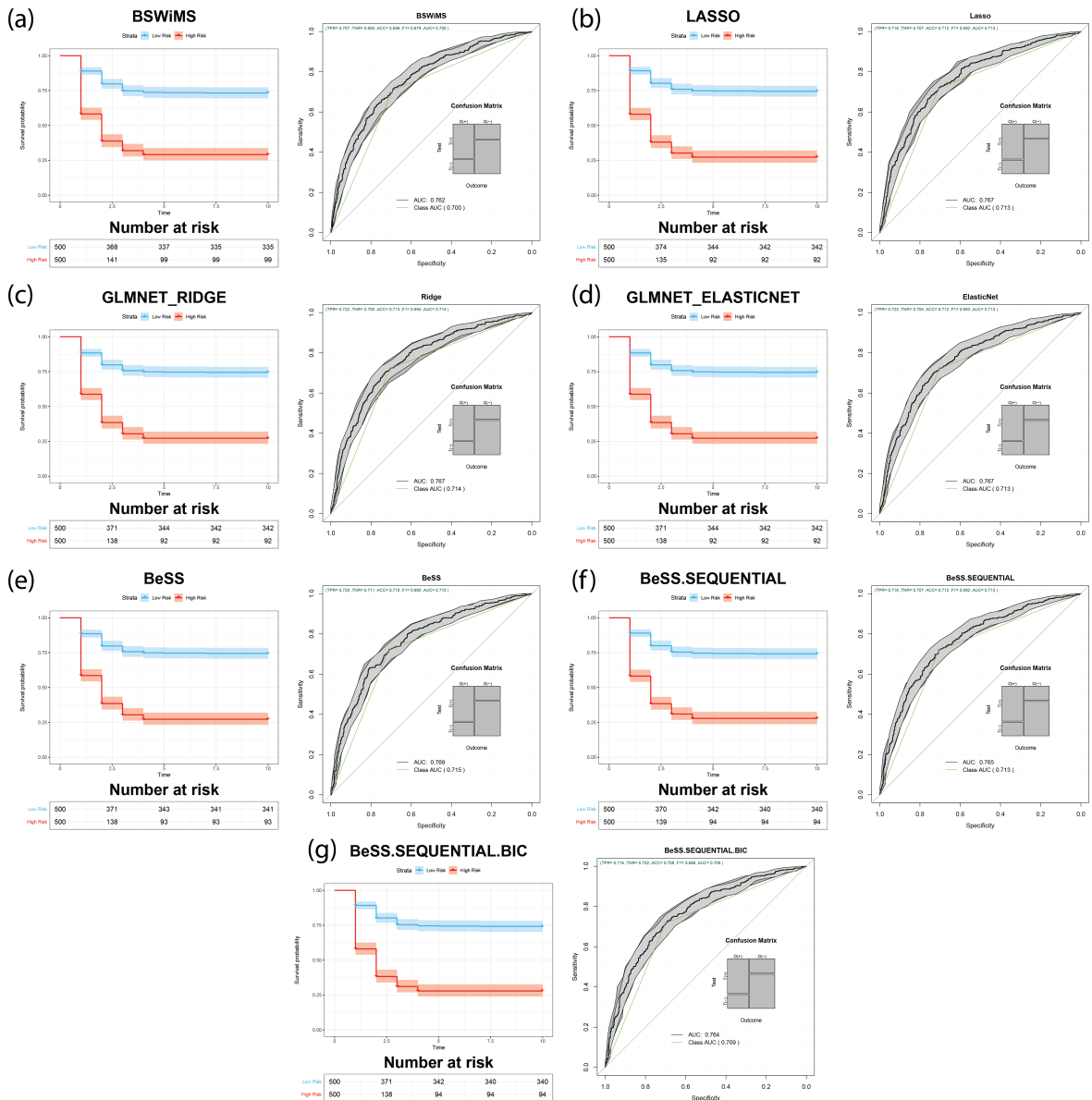


Figure 4.4: KM and ROC Curves for wrappers methods in the Simulation experiment with 21 features. (a) BSWiMS (b) LASSO (c) RIDGE (d) ELASTICNET (e) GSPDAS (BeSS) (f) SPDAS (g) SPDAS with BIC

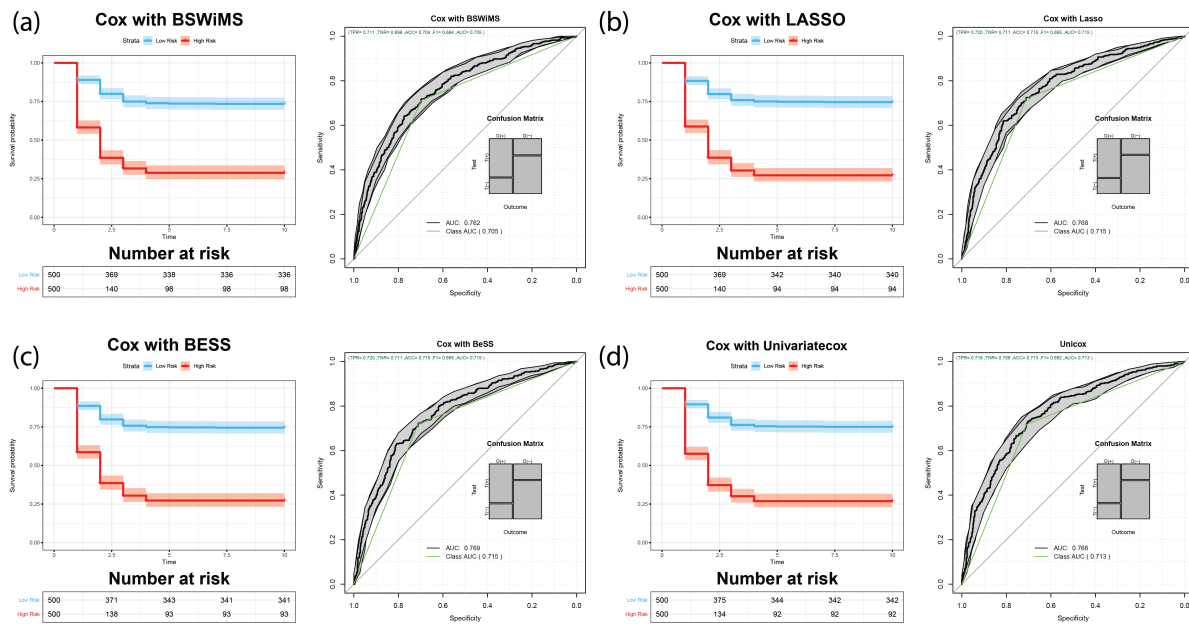


Figure 4.5: KM and ROC Curves for filters methods in the Simulation experiment with 21 features (a) BSWiMS with Cox (b) LASSO with Cox (c) GSPDAS with Cox (d) Univariate Cox Analysis

### 101 features (11 real - 90 random)

In this experiment the same number of real features were used. We used 90 random variables (50% of normal and 50% binomial random distributed variables) to the dataset to test the CoxBenchmarking process. The first Model, again BSWiMS selected 4.70 features on average. Its Jaccard Index is 0.64. Eleven features were used in total, four of them were random numbers. Of the remaining seven, three were selected in every single model. Surgeries, Blocks and Lesions. Defensive Rating, FGP, Offensive Rating and age were selected from 0.65-0.15%. BSWiMS reported ACC = 0.71 (0.68, 0.74), AUC = 0.78 (0.76, 0.81), SEN = 0.72 (0.68, 0.75), SPE = 0.71 (0.67, 0.75), CIRisks = 0.67 (0.65, 0.68) and CIFU = 0.48 (0.45, 0.51). Model II uses LASSO and selected 21.70 features on average with a Jaccard index is 0.47. Seven features were selected in all the 20 iterations, 6 of them are real features. Surgeries, Blocks, Lesions, Defensive Rating, Field Goal Percentage and age. The other feature selected in 100% of the models was random. Offensive Rating was used on 19 out the 20 iterations and BMI was used in 10 models. Lasso finished with ACC = 0.72 (0.69, 0.75), AUC = 0.79 (0.76, 0.82), SEN = 0.72 (0.69, 0.76), SPE = 0.72 (0.67, 0.76), CIRisks = 0.67 (0.65, 0.68) and CIFU = 0.58 (0.55, 0.61). Model III and IV got the same statistical results but the number of features selected and the Jaccard Index is different. The RIDGE method selected a mean of 99.95 and a Jaccard Index of 0.94 and ELASTICNET selected 100.10 characteristics with a Jaccard index of 0.95. Both methods select all the real features in all the models. Their stats were ACC = 0.71(0.68, 0.74), AUC = 0.76 (0.73, 0.79), SEN = 0.71 (0.67, 0.75), SPE = 0.7 (0.66, 0.75), CIRisks = 0.65 (0.64, 0.67) and CIFU = 0.67 (0.64, 0.69).

Method	ACC (95% CI)	AUC (95% CI)	SEN (95% CI)	SPE (95% CI)	C-index Risks (95% CI)	C-Index FU (95% CI)
I	0.71 (0.68,0.74)	0.78 (0.76,0.81)	<b>0.72</b> <b>(0.68,0.75)</b>	0.71 (0.67,0.75)	0.67 (0.65,0.68)	0.48 (0.45,0.51)
II	0.72 (0.69,0.75)	<b>0.79</b> <b>(0.76,0.82)</b>	<b>0.72</b> <b>(0.69,0.76)</b>	<b>0.72</b> <b>(0.67,0.76)</b>	0.67 (0.65,0.68)	0.58 (0.55,0.61)
III	0.71 (0.68,0.74)	0.76 (0.73,0.79)	0.71 (0.67,0.75)	0.7 (0.66,0.75)	0.65 (0.64,0.67)	<b>0.67</b> <b>(0.64,0.69)</b>
IV	0.71 (0.68,0.74)	0.76 (0.73,0.79)	0.71 (0.67,0.75)	0.7 (0.66,0.75)	0.65 (0.64,0.67)	<b>0.67</b> <b>(0.64,0.7)</b>
V	0.7 (0.68,0.73)	0.77 (0.74,0.8)	0.71 (0.67,0.75)	0.7 (0.65,0.74)	0.66 (0.64,0.67)	0.47 (0.44,0.5)
VI	0.71 (0.68,0.74)	0.78 (0.76,0.81)	0.72 (0.68,0.75)	0.71 (0.66,0.75)	0.67 (0.65,0.68)	0.47 (0.45,0.5)
VII	0.71 (0.68,0.74)	<b>0.79</b> <b>(0.76,0.82)</b>	0.71 (0.67,0.75)	<b>0.72</b> <b>(0.67,0.76)</b>	<b>0.67</b> <b>(0.65,0.69)</b>	0.47 (0.44,0.5)

Table 4.6: Classification and survival stats for wrapper methods in the Simulation experiment of 21 features (11 real - 100 random). I = BSWiMS, II = LASSO, III = RIDGE, IV = ELASTICNET, V = GSPDAS (BESS), VI = SPDAS (BESS.SEQUENTIAL), VII = SPDAS.BIC (BESS.SEQUENTIAL.BIC). Best scores for each stat are bolded.

Model V uses GSPDAS and chose a mean of 58.20 features with a Jaccard index of 0.45. This was a special case, just surgeries variable was used in all the models. But 8 variables were used in 19 models. Of them, 5 variables were real: Blocks, Lesions, Defensive Rating, Field Goal Percentage, age. Offensive rating and BMI were used in 18 iterations, AST in 16, minutes in 10 and games in 5. BeSS finished an ACC = 0.7 (0.68,0.73), AUC = 0.77 (0.74, 0.8), SEN = 0.71 (0.67, 0.75), SPE = 0.7 (0.65, 0.74), CIRisks = 0.66 (0.64, 0.67) and CIFU = 0.47 (0.44, 0.5). Model VI uses SPDAS algorithm selecting an average 4.90 with a Jaccard Index of 0.64. Surgeries, Lesions and Blocks were used in every single iteration. Eight extra features were used in the models, the half of them are real: Deffensive rating, FGP, Offensive rating and age. All the other real features were completely ignored. SPDAS finished with ACC = 0.71 (0.68, 0.74), AUC = 0.78 (0.76, 0.81), SEN = 0.72 (0.68, 0.75), SPE = 0.71 (0.66, 0.75), CIRisks = 0.67 (0.65, 0.68) and CIFU = 0.47 (0.45, 0.5). Finally, SPDAS with BIC criterion selected a mean of 6.40 features with a Jaccard Index of 0.57. Just ass the last method, it selected the three same real features in all the models. The main difference is that this method selected a total 18 features in the models. The final model finished with ACC = 0.71 (0.68, 0.74), AUC = 0.79 (0.76, 0.82), SEN = 0.71 (0.67, 0.75), SPE=0.72 (0.67, 0.76), CIRisks = 0.67 (0.65, 0.69) and CIFU = 0.47 (0.44, 0.5).

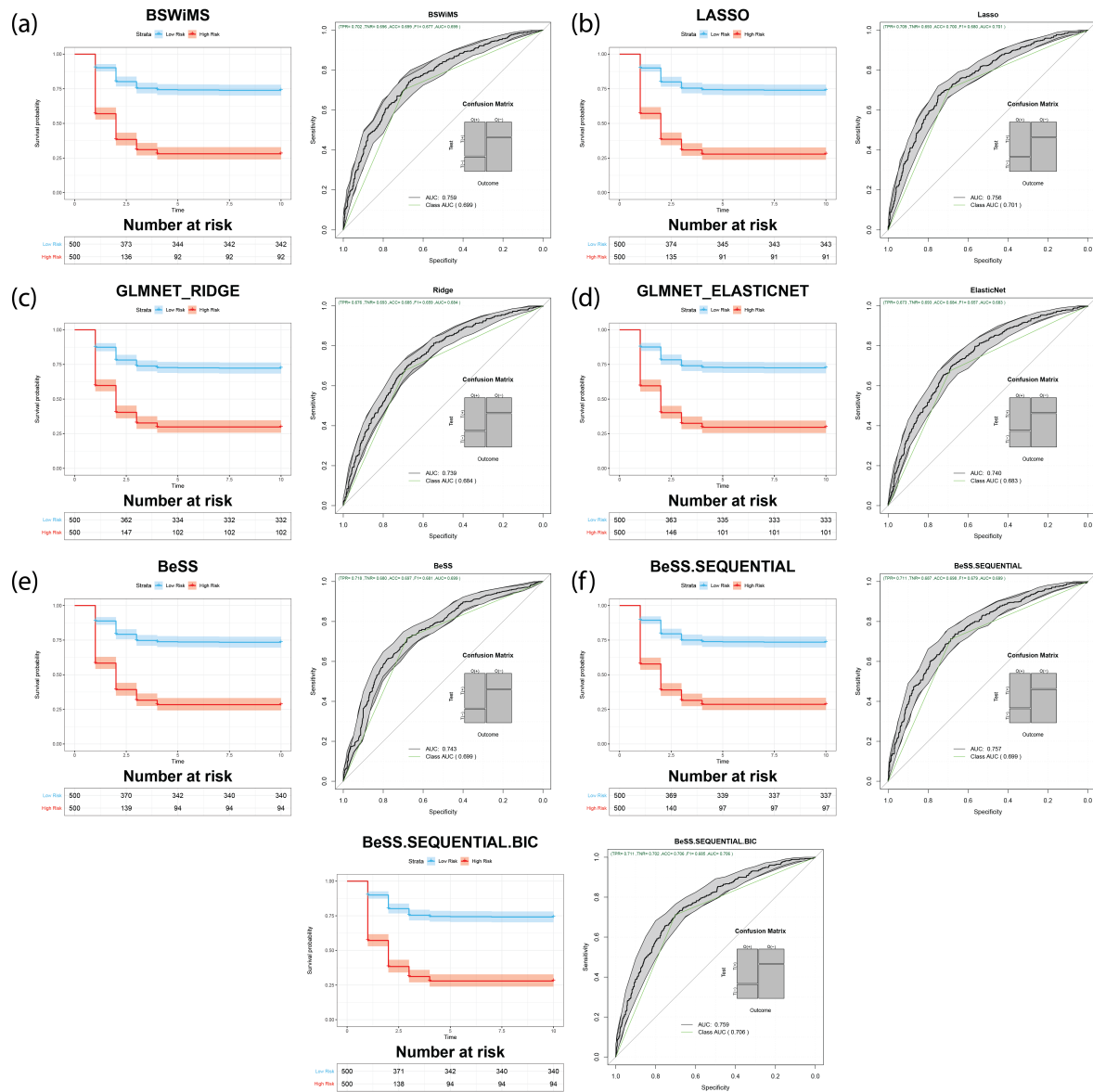


Figure 4.6: KM and ROC Curves for wrappers methods in the Simulation experiment with 101 features (a) BSWiMS (b) LASSO (c) RIDGE (d) ELASTICNET (e) GSPDAS (BeSS) (f) SPDAS (g) SPDAS with BIC



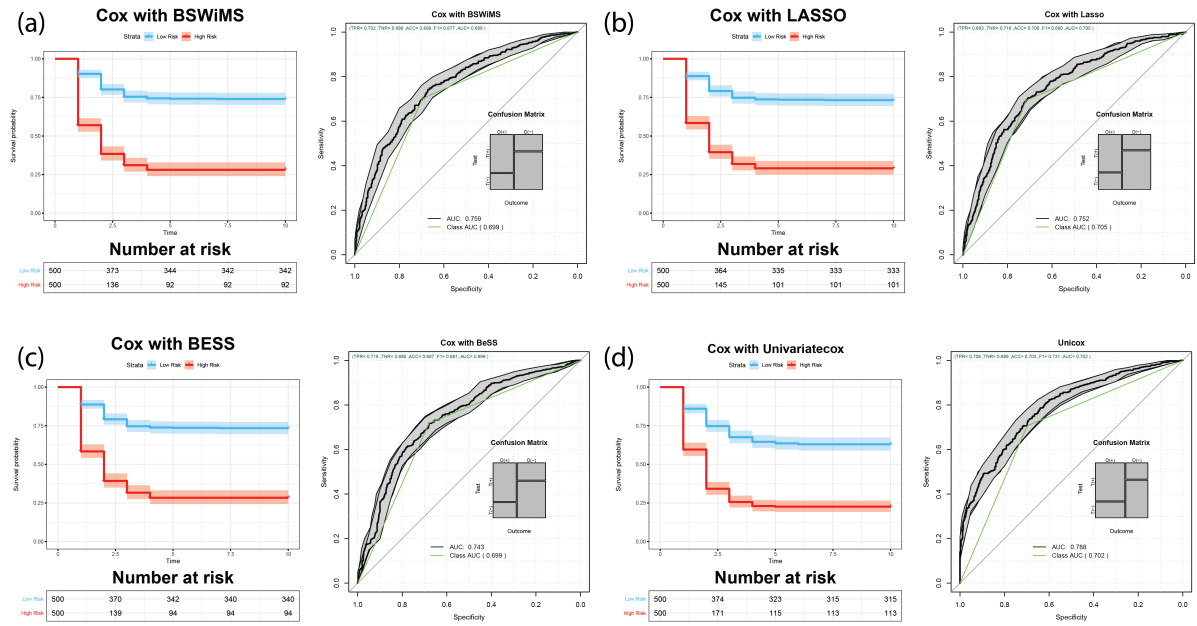


Figure 4.7: KM and ROC Curves for filter methods in the Simulation experiment with 101 features (a) BSWiMS with Cox (b) LASSO with Cox (c) GSPDAS with Cox (d) Univariate Cox Analysis

Method	ACC (95% CI)	SEN (95% CI)	SPE (95% CI)	C-index Risks (95% CI)	C-Index FU (95% CI)
I	0.71 (0.68,0.74)	<b>0.72</b> ( <b>0.68,0.75</b> )	0.71 (0.67,0.75)	<b>0.48</b> ( <b>0.45,0.51</b> )	<b>0.48</b> ( <b>0.45,0.51</b> )
II	<b>0.72</b> ( <b>0.69,0.75</b> )	0.71 (0.67,0.75)	<b>0.73</b> ( <b>0.69,0.77</b> )	0.47 (0.44,0.49)	0.47 (0.44,0.49)
III	0.7 (0.68,0.73)	0.71 (0.67,0.75)	0.7 (0.65,0.74)	0.47 (0.44,0.5)	0.47 (0.44,0.5)
IV	0.71 (0.68,0.74)	<b>0.72</b> ( <b>0.68,0.75</b> )	0.71 (0.66,0.75)	<b>0.48</b> ( <b>0.45,0.5</b> )	<b>0.48</b> ( <b>0.45,0.5</b> )

Table 4.7: Classification and survival stats for filter methods with 11 real features and 100 random variables. I = Cox with BSWiMS, II = Cox with LASSO, III = Cox with BESS IV = Univariate Cox. Best scores for each stat are bolded.

### 1001 features (11 real - 990 random)

Like happened in the past experiment with more than 1000 variables. The only features that were used in all the iterations were real: Lesions and surgeries. But in this case, BSWiMS selected also Blocks and Defensive Rating with 19 and 5 selections respectively. FGP was selected just one time and two random features were also selected. It used a mean of 3.4

features with a Jaccard Index of 0.79. The time that BSWiMS took was 1.12 s on average. It finished with an ACC = 0.71 (0.68, 0.73), AUC = 0.78 (0.75, 0.81), SEN = 0.71 (0.67, 0.75), SPE = 0.70 (0.65, 0.74), CIRisks = 0.67 (0.65, 0.68) and CIFU = 0.48 (0.45, 0.51).

LASSO selected a mean of 31.4 features with a low Jaccard Index of 0.23. Blocks, Defensive Rating, Lesions, and Surgeries were used in every single iteration. Age in 18 models, FGP in sixteen times, Offensive Rating in 12; and finally, BMI in 4 models. Five random features were used in more than half of the iterations. LASSO got an ACC = 0.71 (0.68, 0.73), AUC = 0.78 (0.76, 0.81), SEN = 0.70 (0.66, 0.74), SPE = 0.71 (0.66,0.75), CIRisks = 0.67 (0.65, 0.68) and CIFU = 0.65 (0.63, 0.68).

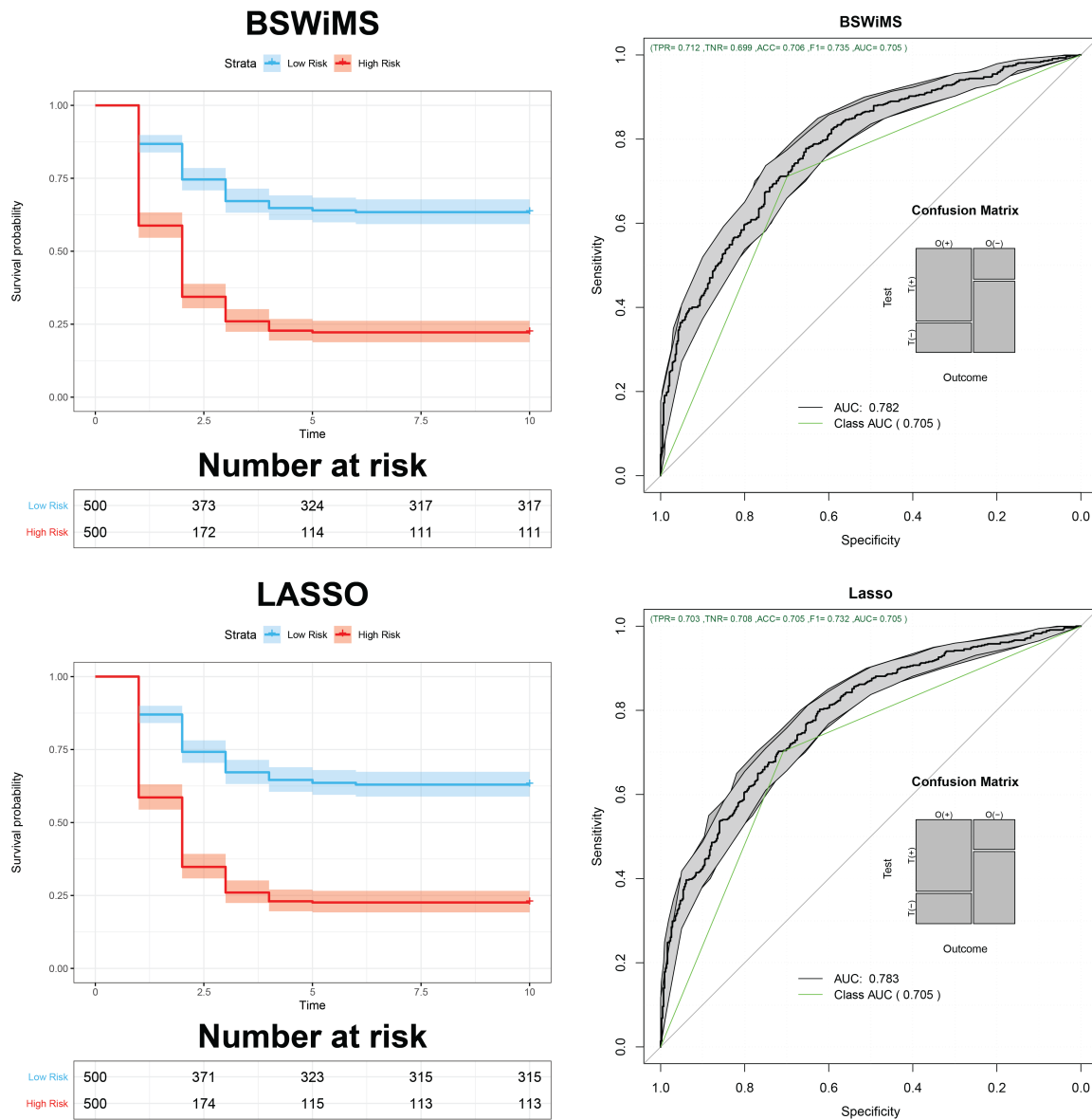


Figure 4.8: KM and ROC Curves for wrappers methods with the 11 real features and 990 random created with (a) BSWiMS (b) LASSO (c) RIDGE (d) ELASTICNET (e) GSPDAS (BeSS) (f) SPDAS (g) SPDAS with BIC

## 4.2 TADPOLE-ADNI

Discovering, characterizing and validating imaging-biomarkers associated with AD requires a well-designed study. The Alzheimer’s Disease Neuroimaging Initiative (ADNI) is a large study aimed to discover and test novel imaging-biomarkers [27]. ADNI has generated hundreds of research papers in this area, but most of them have used supervised classifications or statistical approaches for the characterization of MCI patients that presented with AD conversion. These research papers have been useful in discovering early imaging findings, but most of them have not evaluated effectively the time to AD conversion in their discovery efforts [28]–[30]. In this case, this thesis supported two experiments to discover how imaging data can be used to predict the conversion. Both analyses were developed with the CoxBenchmarking method. The first of the experiments was performed with the first version of CoxBenchmarking. The results with this version report the statistics only of the default methods of each of the packages. Experiment number two uses the final version of CoxBenchmarking and returns to the experimentation with the first data set, which is subsequently compared with clinical information from CSF measures and Cognitive Assessments.

Method		FS (JI)	C-Index Follow Up (95% CI)	LogRank pvalue	AUC (95% CI)	ACC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
LASSO (Coxnet)	W	24.40 (0.31)	<b>0.84</b> <b>(0.82-0.86)</b>	<b>3.33x10<sup>-16</sup></b>	<b>0.73</b> <b>(0.68-0.78)</b>	<b>0.68</b> <b>(0.64-0.72)</b>	<b>0.69</b> <b>(0.62-0.76)</b>	<b>0.68</b> <b>(0.62-0.74)</b>
	F		0.72 (0.69-0.74)	3.73x10 <sup>-13</sup>	0.72 (0.68-0.77)	0.67 (0.62-0.72)	0.66 (0.59-0.73)	0.67 (0.61-0.72)
BSWiMS	W	12.75 (0.34)	0.81 (0.79-0.83)	1.58x10 <sup>-14</sup>	<b>0.73</b> <b>(0.68-0.78)</b>	0.67 (0.63-0.72)	0.69 (0.61-0.75)	0.67 (0.61-0.72)
	F		0.74 (0.71-0.76)	8.10x10 <sup>-15</sup>	0.73 (0.68-0.77)	0.67 (0.63-0.72)	0.67 (0.60-0.74)	0.67 (0.61-0.73)
BeSS	W	52.85 (0.21)	0.63 (0.60-0.66)	1.99x10 <sup>-10</sup>	0.68 (0.63-0.73)	0.62 (0.58-0.67)	0.61 (0.54-0.68)	0.63 (0.57-0.69)
Univariate Cox	F	101.30 (0.67)	0.67 (0.64-0.70)	6.39x10 <sup>-11</sup>	0.67 (0.62-0.72)	0.63 (0.58-0.67)	0.61 (0.54-0.68)	0.64 (0.57-0.69)

Table 4.8: Models predictions statistics. contains c-index of follow-up times predictions, the p-value on log rank test between low-high risk curves, area under the curve, accuracy, sensitivity and specificity with their 95% confidence intervals. W=Wrappers, F=Filters, FS = Feature Size, JI = Jaccard Index. Best scores for each stat are bolded.

### 4.2.1 Survival Models Associated with MCI to AD Conversion with qMRI features

Table 4.8 shows the main results of the RHOCV on the six tested models. We report the major findings per method and all the performance statistics with 95% CI. The BSWiMS strategy selected the smallest models. They contained an average of 13 features with an average Jaccard index of 0.34. The mean volume of the amygdala and entorhinal and the

mean cortical thickness average of bankssts were selected on every iteration. The BSWiMS model had c-index of 0.81 (0.79-0.83) with ACC = 0.67 (0.63, 0.71), SEN = 0.69 (0.61, 0.75), SPE = 0.67 (0.61,0.72), and AUC = 0.73 ( 0.68, 0.78). The Cox Modeling based on BSWiMS reported the following classification performance: ACC = 0.67 (0.63, 0.72), SEN = 0.67 (0.60, 0.74), SPE = 0.67 (0.61, 0.73), and AUC = 0.73 (0.68, 0.77). Hence BSWiMS models were very similar to CoxPH fitted model. Figure 4.9(a) shows the Kaplan-Meier curves of the subjects predicted at risk of conversion vs the subjects predicted as stable for the Cox model created by BSWiMS features.

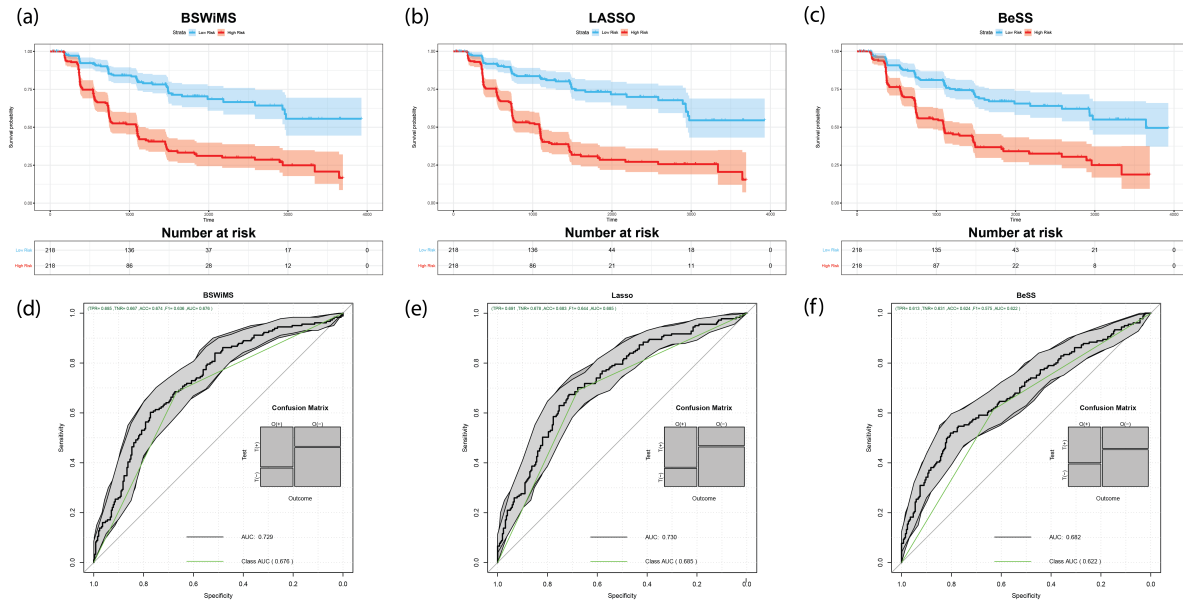


Figure 4.9: Kaplan Meier (KM) and ROC curves for wrappers/embedded section. CoxNet showed the best accuracy on the classification and the best c-index on Risk and Follow-up times. (a) Model 1 BSWiMS KM (b) CoxNet KM (c) BeSS KM (d) BSWiMS ROC (e) CoxNet ROC (f) BeSS ROC

The CoxNet/LASSO method generated models with an average set of 24 features with a Jaccard index of 0.28. The most common features were APOE4, the mean cortical thickness average of Bankssts and the mean volume (cortical parcellation CP) of entorhinal. 75% of the repetitions selected the mean volume (cortical parcellation) of inferior temporal, the absolute difference of cortical thickness average of par s opercularis, the mean volume (WM parcellation) of the amygdala and the mean cortical thickness standard deviation of bankssts. This model reported c-index = 0.84 (0.82-0.86), ACC = 0.68 (0.64, 0.73), SEN = 0.69 (0.62, 0.76), SPE = 0.68 (0.62, 0.74), and AUC = 0.73 (0.68, 0.78). The Cox regression models fitted with LASSO features returned the following performance: ACC = 0.67 (0.62, 0.71), SEN = 0.66(0.59, 0.73), SPE = 0.67 (0.60, 0.72) and AUC = 0.72 (0.68, 0.77). These results indicate that CoxPh performance is lower than L1 fitted model, implying that L1 penalization helped in improving the prediction of which subjects converted. Figure 4.9(b) shows the Kaplan-Meier curves. The BeSS method returned on average models with 53 features with a Jaccard index of 0.21. Three features were selected on every single repetition: APOE4, mean cortical thickness standard deviation of bankssts and mean volume (cortical parcellation) of

entorhinal.

Variable	FT	MT	Event Mean(SD)	No event Mean (SD)	MV HR (95% CI)	UV HR (95% CI)	M1	M2	M3	M4
CP entorhinal	<i>M</i>	<i>V</i>	1783.72 (426.17)	1535.31 (423.91)	0.63*** (0.50,0.80)	0.47**** (0.39,0.57)	1	1	1	1
WMP amygdala	<i>M</i>	<i>V</i>	1247.77 (280.46)	1053.28 (276.06)	0.88* (0.69,1.13)	0.50**** (0.41,0.60)	2	5	12	2
CP inferior temporal	<i>M</i>	<i>V</i>	9681.64 (1617.63)	8896.89 (1803.11)	0.79 $\alpha$ (0.62,1.00)	0.49**** (0.40,0.61)	17	4	20	3
AVG Bankssts	<i>M</i>	<i>CT</i>	2.39 (0.21)	2.27 (0.23)	0.76* (0.60,0.97)	0.54**** (0.45,0.67)	3	2	63	6
APOE4	<i>P</i>	<i>G</i>	NA	NA	1.74**** (1.40,2.17)	1.83**** (1.50,2.24)	5	3	2	8
SD Bankssts	<i>M</i>	<i>CT</i>	0.51 (0.08)	0.54 (0.08)	1.60** (1.20,2.12)	1.74**** (1.35,2.24)	4	6	3	25
AVG pars opercularis	<i>A</i>	<i>CT</i>	2.38 (0.18)	2.29 (0.20)	1.46** (1.09,1.94)	1.76*** (1.32,2.36)	39	7	17	32
AVG inferior parietal	<i>A</i>	<i>CT</i>	2.25 (0.19)	2.15 (0.21)	1.33* (1.04,1.69)	1.5** (1.17,1.92)	42	9	13	47
AVG middle temporal	<i>A</i>	<i>CT</i>	2.70 (0.21)	2.56 (0.23)	1.38* (1.07,1.78)	1.52** (1.17,1.97)	21	11	14	50
SD Rostral middle frontal	<i>M</i>	<i>CT</i>	0.62 (0.05)	0.61 (0.048)	0.63*** (0.50,0.81)	0.80* (0.64,0.98)	34	13	5	76

Table 4.9: Characteristics and ranking of ten features selected in almost the half of the iterations. The ranking was ordered based on the number of times selected and then ordered depending on the p-value of univariate cox analysis. [FT = feature type; M=mean, P= polymorphism, A=absolute difference], [MT = measure type; V=volume (mm3), G = gene, M = cortical thickness (mm)], [M1 = BSWiMS, M2 = COXNET/LASSO, M3 = BeSS, M4 = Univariate Cox] P. Value significance:  $\alpha < 0.1$ , \* < 0.05, \*\* < 0.01, \*\*\* < 0.001, \*\*\*\* <  $10^{-04}$

75% of the time the following 3 features were selected: mean cortical thickness standard deviation of temporal pole, mean cortical thickness standard deviation of the rostral middle frontal and mean surface area of cuneus. BeSS models reported c-index = 0.63 (0.60, 0.66), ACC = 0.63 (0.58, 0.67), SEN = 0.61 (0.54, 0.68), SPE = 0.63 (0.57, 0.69), and AUC = 0.68 (0.63,0.73). Finally, the models created by univariate Cox filter were the largest. The average size of the models included 103 elements with a Jaccard index of 0.65. 54 features were selected in all the iterations. Among the selected features were APOE4, the mean cortical thickness average of Parahippocampal, the cortical thickness average and the volume (cortical parcellation) of pars opercularis. Classification performance of univariate filter were: ACC =

0.63 (0.58, 0.67), SEN = 0.61 (0.54,0.68), SPE = 0.64 (0.57,0.69) and AUC = 0.67 (0.62,0.72). Hence the Cox models based on simple univariate filter had the least robust performance. Figure 4.9 and Figure 4.10 show the complete Kaplan-Meier curves and the ROC plots based on the median estimations for ML-methods and filter-based-methods.

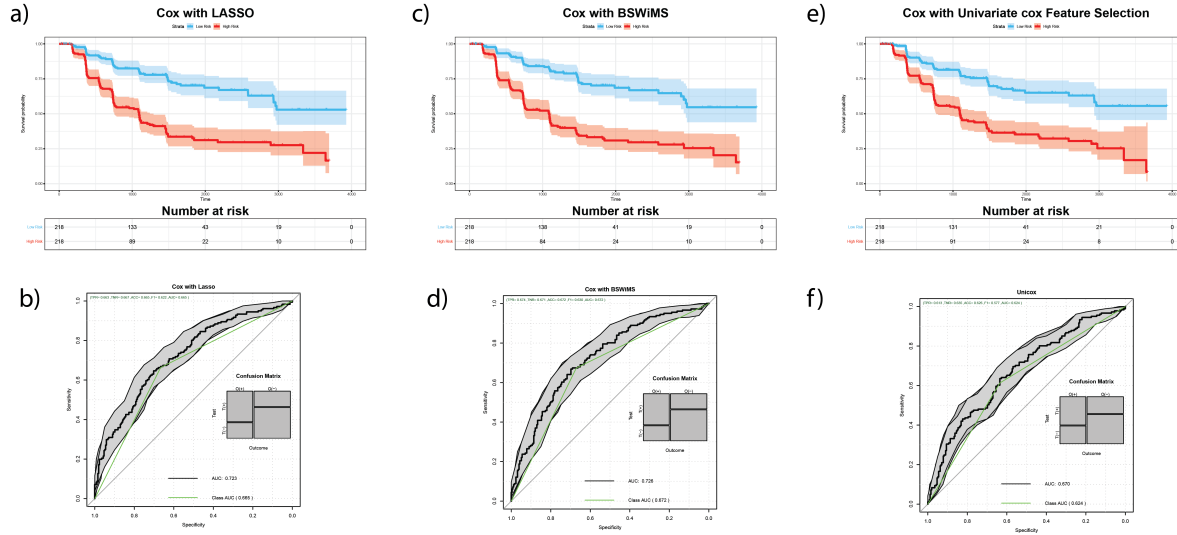


Figure 4.10: Kaplan Meier (KM) and ROC curves for filters section. Cox Model build with BSWiMS features showed the best accuracy on the classification and the best c-index on Risk and Follow-up times. (a) Model 4 Cox with BSWiMS KM (b) Cox with BSWiMS ROC (c) Model 5 Cox with CoxNet KM (d) Cox with CoxNet ROC (e) Model 6 Cox with Univariate Cox KM (f) Cox with Univariate Cox ROC

We performed a detailed analysis of the set of selected features across ML methods. The analysis of the RHOCV reported that ten features were common on 50% of the sets. To evaluate the importance of these ten features as a risk factor for MCI to AD conversion, we refit the Cox model using these ten features. We then reported the hazard ratios (HR) and their corresponding 95% CI: The mean volume (CP) of entorhinal HR = 0.63 (0.50, 0.80), mean cortical thickness SD of Bankssts HR = 1.60 (1.20,2.12), APOE4 HR = 1.74 (1.40,2.17), mean volume (WMP) of amygdala HR = 0.88 (0.69,1.13), mean cortical thickness AVG of Bankssts HR = 0.76 (0.60,0.97), mean volume (CP) of inferior temporal HR = 0.79 (0.62,1.00), absolute difference cortical thickness AVG of middle temporal HR = 1.38 (1.07, 1.78), absolute difference of cortical thickness AVG of pars opercularis HR = 1.46(1.09, 1.94), absolute difference cortical thickness average of inferior parietal HR = 1.33 (1.04, 1.70), mean cortical thickness standard deviation of Rostral middle frontal HR = 0.64(0.50, 0.81). A heatmap representation with the ten features correlation with the outcome can be found in Figure 4.11. Table 4.9 provides more details of the ten characteristics. The last two columns of table III shows the rank of the features of the four ML approaches. The MV HR and the UV HR correspond to the Hazard ratios of the feature inside a Multivariate model and the HR computed by the univariate approach respectively. It is clear that feature ranking and importance depended on the ML method.

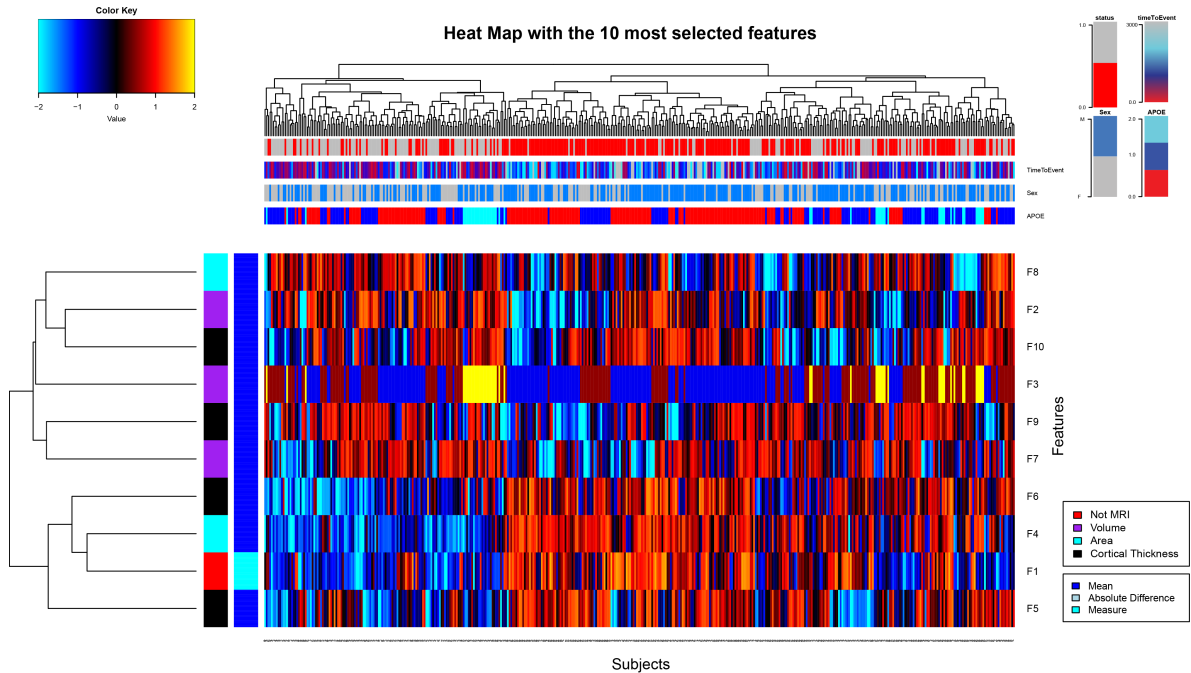


Figure 4.11: A heat map representation of the features associated with MCI to AD conversion. The figure shows the ten features selected by all the 4 methods in at least in the half of the iterations (horizontal axis) and subjects on the vertical axis. (F1) Mean volume (CP) of entorhinal, (F2) mean cortical thickness SD of Bankssts, (F3) APOE4, (F4) mean volume (WMP) of amygdala, (F5) mean cortical thickness AVG of Bankssts, (F6) mean volume (CP) of inferior temporal, (F7) absolute difference cortical thickness AVG of middle temporal, (F8) absolute difference of cortical thickness AVG of pars opercularis, (F9) absolute difference cortical thickness AVG of inferior parietal, (F10) mean cortical thickness SD of Rostral middle frontal

#### 4.2.2 Prediction of MCI to AD Risk of Conversion Survival Models: qMRI vs CSF Measures and Cognitive Assessments

Table 4.11 shows main classification (ACC) and survival stats (c-index FT) for all the models in Experiment I, II, III and VI. Some configurations were tested with all the 9 ML/SL different strategies. In this section, we report the main findings in each group and the strategies explored within the group. Considering just the CSF measures group (Experiment I), BSWiMS selected a mean of 1.75 measures and a Jaccard index of 0.80. LASSO selected the smallest mean of CSF measures with 2.95 and a JI of 0.97. selected 2.95 features and 0.96 JI. GPDAS chose 2 features on average with a Jaccard index of 0.77. Its version with sequential algorithm and BIC adjusted selected 1.50 measures with 1 JI. RIDGE, SPDAS, and Unicox selected 3 features i.e. all of them, thus a Jaccard Index of 1. Since all the models almost selected all the measures 100% of the time, the best performance was reported by RIDGE with 0.76 Follow Up c-index and its 95% confidence interval (CI) of (0.74,0.79). Besides, the best c-index on Risks was found by GPDAS as a filter with 0.75(0.72,0.78), Figure 4.12(a) shows the ROC and Kaplan Meier curves for GPDAS as a filter model.

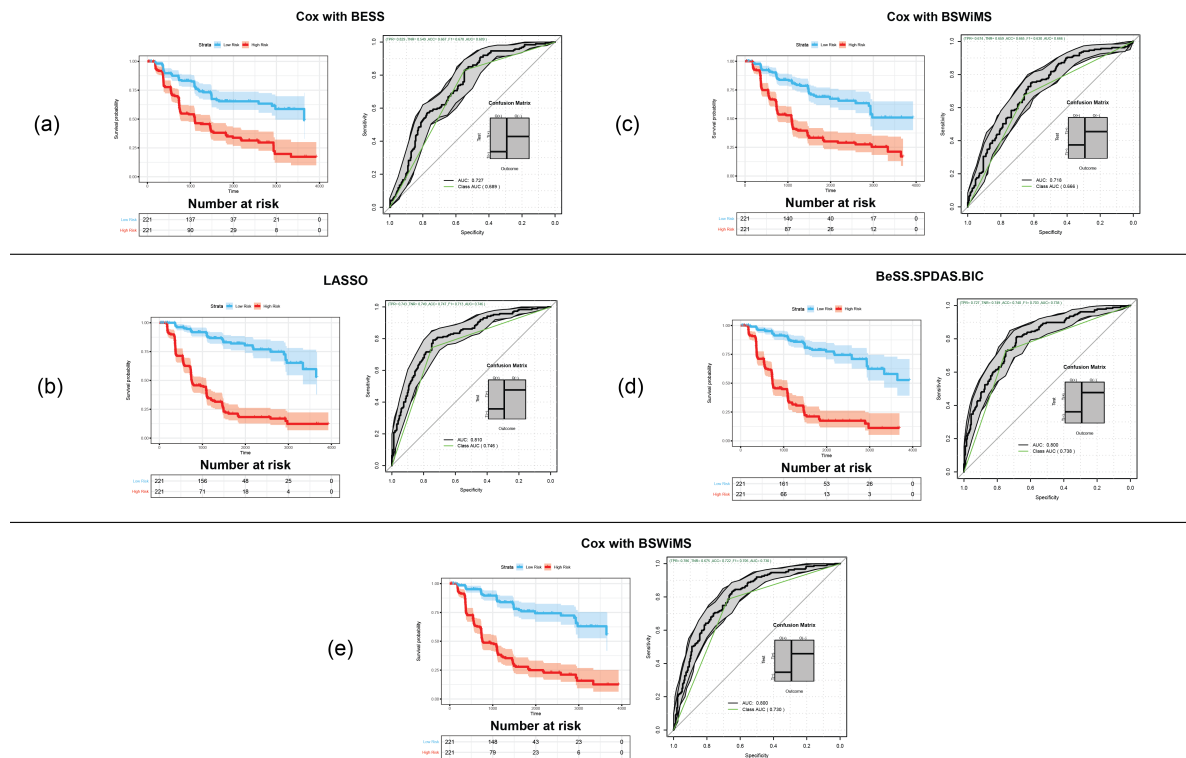


Figure 4.12: KM and ROC Curves for (a) Experiment I model GPDAS as a filter c-index Risks = 0.75 (0.72, 0.78), ACC = 0.67 (0.62, 0.71) (b) Experiment II model LASSO c-index Risks = 0.65 (0.62, 0.67), ACC = 0.76 (0.72, 0.8) (c) Experiment III model BSWiMS as a filter c-index Risks = 0.74 (0.72, 0.77), ACC = 0.67 (0.62, 0.71) (d) Experiment IV model SPDAS.BIC c-index Risks = 0.67 (0.64, 0.70), ACC = 0.74 (0.7, 0.78) (e) Experiment V model BSWiMS as a filter c-index Risks = 0.68 (0.65, 0.71), ACC = 0.72 (0.68, 0.76)

The worst p-value found in the LogRank test was 3.34E-06 by SPDAS.BIC. With Cog-assessments features (Experiment II) BSWiMS selected a mean of 4.30 and JI of 0.80. Just two measurements were always selected: 13-tasks ADAS version and RAVLT immediate score. FAQ and CDRSB were selected in more than 80% of the iterations; MMSE was never selected. LASSO selected a mean of 4.30 Cog-assessments scores and 0.90 of JI. ADAS13, FAQ, RAVLT immediate and CDRSB were selected in 100% of the iterations, ADAS11 was not selected. RIDGE selected all the forms all the time. GPDAS has an average of 5.45 features and JI of 0.78. It selected the same features as LASSO in 100% of the cases. SPDAS has an average of 4.35 and 0.85 JI. It selected ADAS13 and FAQ in 100% of the iterations and RAVLT immediate and CDRSB in the 95% of them. The BIC version selected 4 features on average and 0.76 of JI. Just ADAS13 was selected on all the occasions and FAQ, RAVLT immediate and CDRSB were selected in more than 85%. Finally, the Univariate cox analysis selected all features, but RAVLT forgetting score, were considered 100% of the times, the not selected one was not chosen just in one of the iterations. mean of 4 features with a JI of 0.76. Among this group, the best accuracy and area under the curve for this outcome were found. LASSO reported the best c-index FT. The reported classifications stats are ACC = 0.76 (0.72, 0.8), AUC = 0.81 (0.76, 0.85), SEN = 0.76 (0.69, 0.82), SPE = 0.76 (0.7, 0.81) and survival stats c-index FT = 0.69 (0.66, 0.72) and c-index Risks 0.65 (0.62, 0.67). Figure 4.12(b) shows



the KM and ROC Curves for the LASSO model. All the p-values on the LogRank Test were 0.

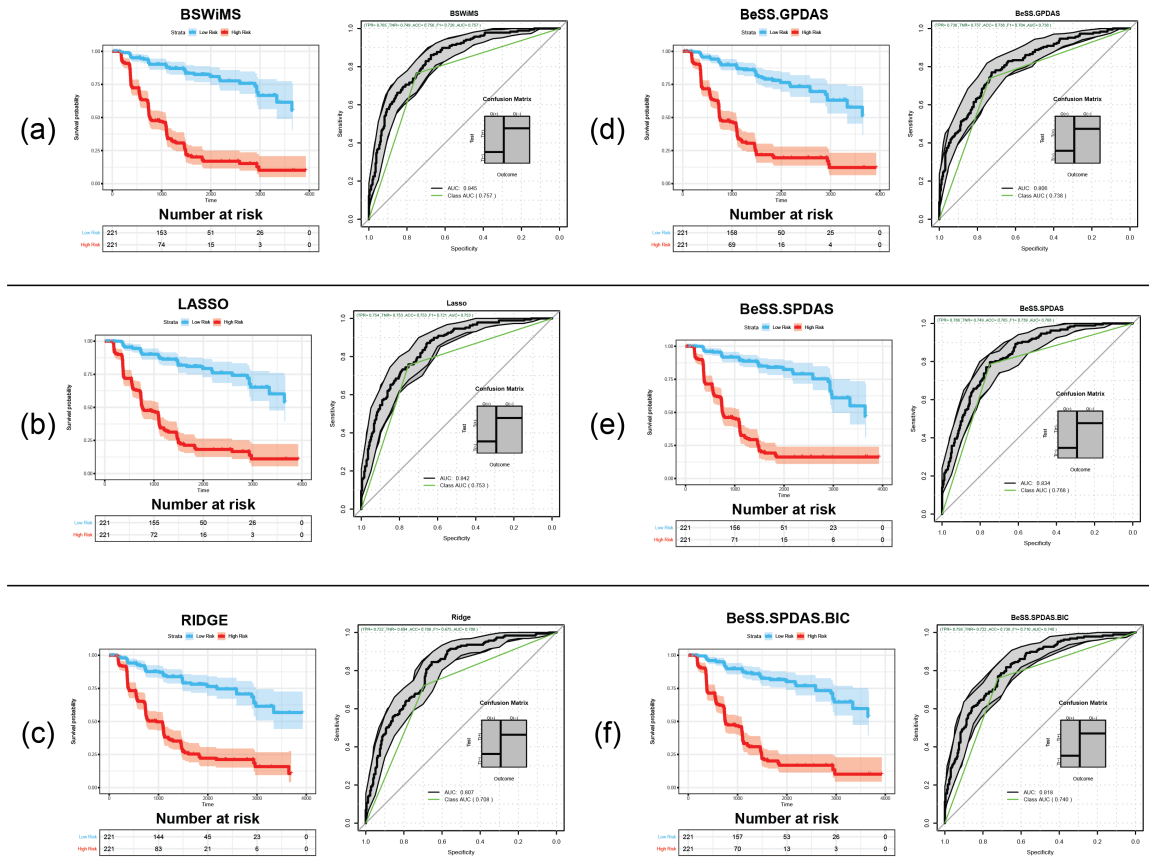


Figure 4.13: KM and ROC Curves for all the models of Experiment IV (CSF Measures + Cog-assessments + Radiomics). (a) BSWiMS Model [ACC = 0.76 (0.71, 0.8) c-index FT = 0.69 (0.66, 0.72) c-index Risks = 0.65 (0.63, 0.68)] (b) LASSO Model [ACC = 0.75 (0.71, 0.79) c-index FT = 0.77 (0.75, 0.79) c-index Risks = 0.66 (0.63, 0.69)] (c) RIDGE Model [ACC = 0.71 (0.66, 0.75) c-index FT = 0.91 (0.89, 0.92) c-index Risks = 0.63 (0.61, 0.66)]. (d) GPDAS Model [ACC = 0.74 (0.69, 0.78) c-index FT = 0.59 (0.55, 0.62) c-index Risks = 0.65 (0.62, 0.68)]. (e) SP DAS Model [ACC = 0.76 (0.72, 0.8) c-index FT = 0.63 (0.6, 0.67) c-index Risks = 0.66 (0.63, 0.68)]. (f) SP DAS.BIC Model [ACC = 0.74 (0.69, 0.78) c-index FT = 0.61 (0.58, 0.64) c-index Risks = 0.66 (0.63, 0.68)]

Using just Radiomics information (qMRI features + APOE4 + SEX) (Experiment III), BSWiMS selected an average of 11.30 features and 0.29 JI. The mean volume of amygdala was selected in all the iterations. Mean volume of entorhinal and mean cortical thickness of averages and standard deviation of the Bankssts, APOE4 were selected in more of the 75% of the iterations. Coxnet/LASSO chose 22.75 features on average and JI of 0.30. APOE4 and mean volume cortical parcellation (CP) of entorhinal were always selected. In more than 75% of the times, LASSO selected: Mean cortical thickness average of Bankssts, mean volume (WM Parcellation) of amygdala, mean cortical thickness average of Pars Opercularis, mean cortical thickness standard deviation of Bankssts and mean volume (Cortical Parcellation) of

Inferior Temporal. RIDGE selected an average of 292.95 features with 0.89 of the Jaccard index. Both methods chose 149 features in all the repetitions. GPDAS selected 52.60 features and its JI was 0.21. Despite the big number of selections, just one feature was used all the time, APOE4. Mean volume cortical parcellation (CP) of entorhinal, mean cortical thickness standard deviation of Bankssts, mean surface area of cuneus, mean volume (CP) of Rostral Anterior Cingulate, mean cortical thickness standard deviation of Temporal Pole and absolute difference of volume (WM Parcellation) of Amygdala were selected in at least the 75% of the repetitions. SPDAS algorithm selected the smallest number of features with 4.85 and Jaccard Index of 0.30. Mean volume (CP) of entorhinal was selected all the times and APOE4 at least 75% of the iterations. SPDAS.BIC selected 19.70 features on average with 0.22 JI. APOE4 and once again mean volume (CP) of entorhinal were selected in the 100% of the repetitions. Finally, in this experiment, Univariate cox analysis selected a mean of 98.80 features with a Jaccard index of 0.63. 45 features were selected all the time. Among all the models, RIDGE had the best performance on the c-index of follow up times with 0.92 (0.91, 0.93) and its ACC = 0.68 (0.63, 0.72). On the other hand, the best performance on c-index Risks was found with BSWiMS as a filter with 0.74 (0.72, 0.77), KM and ROC Curves for this model are shown on Figure 4.12(c). The smallest p-value found on the LogRank Test was 4.08E-10 in the Unicox model.

Experiment IV uses information about Radiomics and Cog-assessments scores. BSWiMS selected 10.70 features JI = 0.5174. It selected all the Cog-assessments except for MMSE and RAVLT forgetting in all the iterations and mean volume (WM Parcellation) of Amygdala in 75% of the times. LASSO selected 27.25 features with a JI of 0.36. It chose ADAS13, CDRSBm FAQ, RAVLT immediate, APOE4, absolute difference of cortical thickness average of Superior frontal, and mean volume (CP) of inferior temporal. Absolute differences in cortical Thickness SD of Lingual and volume (CP) of Pars Triangularis, and mean volume (CP) of Entorhinal were selected more than 15 times. RIDGE selected 290.60 features with 0.86 JI. It selected 122 features including all the forms, all the time. GPDAS selected 52.70 features on average with JI=0.23. ADAS13, RAVLT immediate were selected 100% of the cases. CDRSB, FAQ, APOE, mean cortical Thickness SD of Bankssts and TemporalPole, absolute differences of volume (CP) of Pars Triangularis and Surface Area of Transverse Temporal, mean volume (CP) of Caudal Middle Frontal and mean surface area of Cuneus were selected more than 75% of the times. SPDAS chose 5.45 features with 0.46 JI. ADAS13 was selected always and mean volume (CP) of Inferior Temporal, FAQ, RAVLT immediate were selected more than 15 times. SPDAS.BIC selected 25.85 features with a JI of 0.27. ADAS13, FAQ, RAVLT immediate were selected 100% of the time. Mean volume (CP) of Inferior Temporal, absolute differences of volume (CP) of Pars Triangularis and surface area of Transverse Temporal, and CDRSB were selected more than 75% of the iterations. Finally, Unicox selected 105.50 05 features with 0.66 of JI. It selected 51 features all the time, including all the forms. The model with the best performance on the c-index Follow-up Times was RIDGE with 0.92 (0.91, 0.93) and the best c-index Risks was found with SPDAS.BIC 0.67 (0.64, 0.7). SPDAS.BIC KM and ROC curves are shown on Figure4.12(d). All the p-values on LogRank Test were zero.

Experiment V uses information about Radiomics and CSF measures. BSWiMS selected an average of 5.40 features with 0.35 of JI. It selected  $A\beta_{1-42}$  all the times and mean volume of entorhinal in the 75%. Coxnet/LASSO used 24.95 features with a JI of 0.29. Once again,

the same features as BSWiMS were selected all the times, APOE4, the absolute difference of cortical thickness average of pars opercularis, mean volume (CP) of inferior temporal and mean cortical thickness average of bankssts. RIDGE and ELASTICNET used 293.60 features with a Jaccard index of 0.88. They selected the three CSF Measures, APOE4 and 142 qMRI features. GPDAS selected 51.65 features with 0.18 of JI.  $A\beta_{1-42}$  and mean volume of entorhinal all the iterations and APOE4 and other 3 qMRI features in at least 75% of the cases. SPDAS algorithm just used 4.05 features with 0.43 of JI. The same two features were selected 100% of the time. SPDAS.BIC has a mean of 18.40 features and 0.23 JI. Same both features selected at 100% but the volume (CP) of rostral anterior cingulate join them in 75% of the cases. Univariate cox selected a mean of 107.35 features with JI of 0.68, 52 features were selected all the times, the three CSF Measures were included. Experiment V got the best performance among all the experiments on the RIDGE method with a c-index Follow-up times of 0.93 (0.91, 0.94) and its best performance on c-index Risks with 0.68 (0.65, 0.71) in the model build that uses BSWiMS as a filter, whose KM and ROC curves are part of the Figure4.12(e). The biggest p-value for the log-rank test was found on the GPDAS model with  $1.41E-12$ .

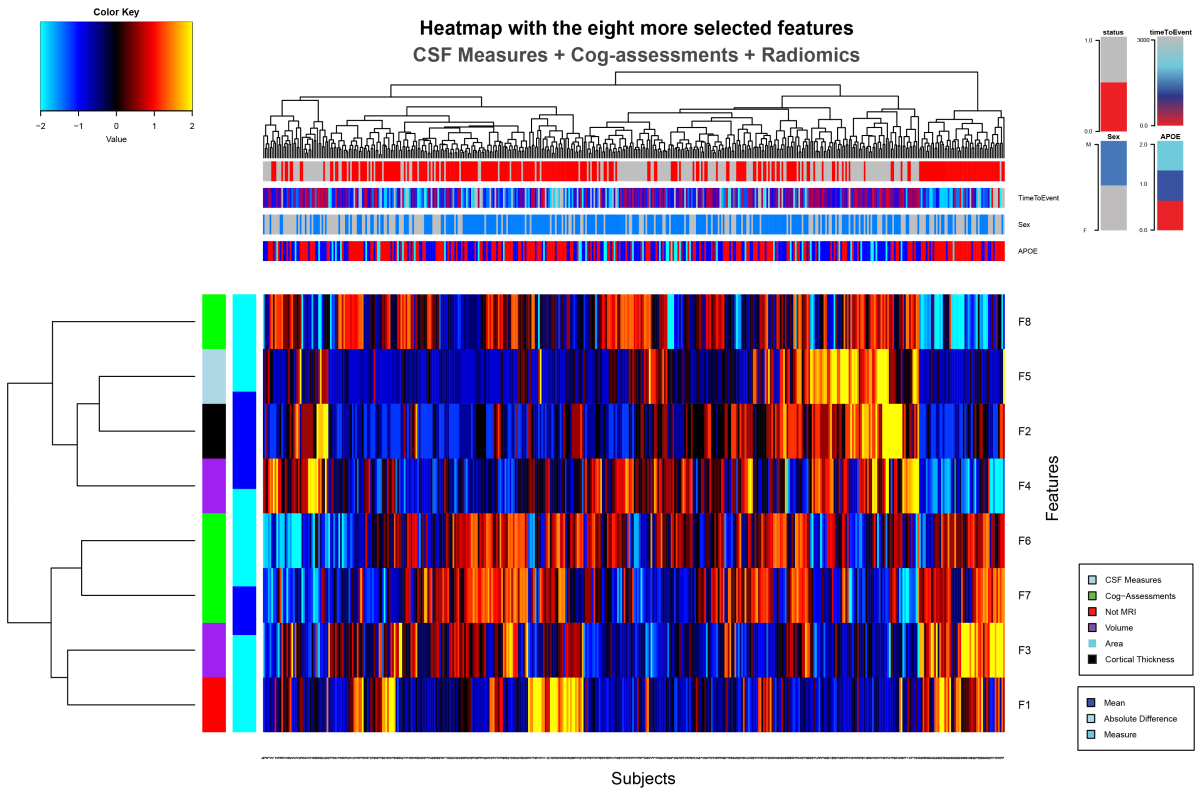


Figure 4.14: A heat map representation of the features associated with MCI to AD conversion. The figure shows the 8 features selected by all the 6 methods in at least in the half of the iterations (horizontal axis) and subjects on the vertical axis. (F1)  $A\beta_{1-42}$ , (F2) CDSRB, (F3) RAVLT immediate, (F4) ADAS13, (F5) FAQ, (F6) Mean volume CP Inferior Temporal, (F7) Mean volume CP Entorhinal, (F8) Mean cortical thickness SD Bankssts.

Finally, experiment number VI used the combination of all the feature groups (CSF

Measures + Cog-assessments + Radiomics). Summary of c-index Follow-up Times and ACC are shown in Table 4.11. Figure 4.13 shows KM and ROC curves for all the Wrapper models in this experiment. BSWiMS selected a mean of 11.55 features with a Jaccard Index of 0.51. Four features were used in every single iteration. 3 of them are Cog-assessments scores: ADAS13, CDRSB, RAVLT immediate, and the remaining feature is the CSF Measure  $A\beta_{1-42}$ . FAQ, ADAS11 and mean volume (WM Parcellation) of Amygdala were selected in the 75% of the times.

V	FT	MT	Event Mean (SD)	No event Mean (SD)	MV HR (95% CI)	UV HR (95% CI)	M1	M2	M3	M4	M5	M6	M7
1	C	S	1.85 (0.92)	1.25 (0.69)	1.40*** (1.20,1.60)	1.8**** (1.6,2)	3	4	4	2	3	2	4
2	C	S	28.71 (8.04)	36.22 (10.39)	0.61*** (0.49,0.77)	0.45**** (0.37,0.54)	2	2	2	1	5	1	2
3	C	S	20.68 (6.25)	14.44 (5.66)	1.60*** (1.30,1.90)	2.70**** (2.30,3.20)	1	1	1	3	2	6	1
4	C	S	5.01 (4.65)	2.21 (3.3)	1.20* (1.00,1.40)	1.60**** (1.40,1.80)	5	3	3	7	6	5	3
5	M	V	1535.31 (423.91)	1783.72 (426.17)	0.79* (0.63,0.99)	0.47**** (0.39,0.57)	6	5	5	8	7	8	5
6	P	P	766.81 (715.52)	598.21 (666.43)	0.54*** (0.43,0.68)	0.44**** (0.35,0.55)	4	6	6	4	1	3	6
7	M	V	8896.89 (1803.1)	9681.64 (1617.63)	0.63*** (0.50,0.79)	0.49**** (0.40,0.61)	8	7	7	6	4	4	7
8	M	CT	0.54 (0.08)	0.51 (0.08)	1.40* (1.10,1.90)	1.70**** (1.40,2.2)	7	8	8	5	8	7	8

Table 4.10: Characteristics and ranking of eight features selected in almost the half of the iterations. The ranking was ordered based on the number of times selected, then ordered depending on the p-value of univariate cox analysis and finally, the concordance index of the univariate model. [FT = feature type; M=mean, C=Cog. Assessment, P=CSF Measure], [mt = measure type; v=volume (mm<sup>3</sup>), p = protein, ct = cortical thickness (mm)], [M1 = BSWiMS, M2 = LASSO, M3 = RIDGE, M4 = GPDAS, M5=SPDAS, M6=SPDAS.BIC, M7=Univariate Cox] P. Value significance: < 0.1, \* < 0.05, \*\* < 0.01, \*\*\* < 0.001, \*\*\*\* < 10<sup>-04</sup>

CoxNet/LASSO selected 29.90 features on average with a JI of 0.35. In this model, the features used in the 100% of the cases are more group variant than in the other models:  $A\beta_{1-42}$ , ADAS13, CDRSB, FAQ, RAVLT immediate, mean volume (CP) of Entorhinal and volume (CP) of Inferior Temporal. Four qMRI features were selected in the 75% of the cases (absolute differences of cortical thickness average of superior frontal, cortical thickness standard deviation of left lingual, cortical thickness average of Pars opercularis, and volume (CP) of pars triangularis. RIDGE selected 294.45 features on average with 0.96 JI. 143 features were selected for all the iterations, all the CSF Measures and Cog-assessments were included

in this group. GPDAS selected 52.60 features with a Jaccard Index of 0.25. Among the features used in the 20 iterations are  $A\beta_{1-42}$ , CDRSB, RAVLT immediate and absolute difference of volume (CP) of Pars Triangularis.

		Conversion Accuracy				Concordance Conversion Time			
M	MT	I	II	III	VI	I	II	III	VI
<b>1</b>	W	0.67	0.74	0.66	<b>0.76**</b>	0.75	0.68	0.81	0.69
	F	0.67	0.76	0.67	<b>0.76**</b>	0.75	0.66	0.74	0.63
<b>2</b>	W	0.67	0.76	0.68	0.75	0.74	<b>0.69</b>	0.84	0.77
	F	0.67	0.76	<b>0.69</b>	0.76	0.73	0.67	0.72	0.63
<b>3</b>	W	0.66	0.75	0.68	0.71	<b>0.76</b>	<b>0.69</b>	<b>0.92</b>	<b>0.91</b>
<b>4</b>	W	0.67	0.76	0.64	0.74	0.75	0.66	0.63	0.59
	F	0.67	0.72	0.64	0.74	0.75	0.66	0.63	0.59
<b>5</b>	W	0.67	0.76	0.68	0.76	0.73	0.66	0.74	0.63
<b>6</b>	W	<b>0.69</b>	<b>0.76*</b>	0.67	0.74	0.72	0.67	0.69	0.61
<b>7</b>	W	0.67	0.76	0.63	0.70	0.73	0.66	0.67	0.61

Table 4.11: Main classification (Accuracy ACC) and survival stats (c-index FT) for all the models on Experiments I, II, III, and VI. The complete stats for all the experiments are shown in the Table 4 at the appendix section. Bold number on each column indicates the best stat on that specific experiment. Tiebreakers were performed by the AUC value and 95%CI, \*AUC =0.81(0.77,0.85), \*\*AUC =0.84(0.81,0.88). M = Models (1=BSWiMS, 2=LASSO, 3=RIDGE, 4=GPDAS, 5=SPDAS, 6=SPDAS.BIC, 6=Univariate Cox analysis). MT = Model Type (W = Wrapper, F=Filter). I = CSF Measures. II = Cog-assessments. III = Radiomics. VI = CSF Measures + Cog-assessments + Radiomics.

Then, in at least 15 of the iterations ADAS13, FAQ, Absolute differences of the surface area of Transverse Temporal, cortical thickness average of Superior Temporal, and cortical thickness standard deviation of Left Lingual, Mean of cortical thickness standard deviation of Bankssts, and surface area of Cuneus. SPDAS chose 6.45 features on average with a Jaccard index of 0.50. Any feature was selected every single iteration, but  $A\beta_{1-42}$ , ADAS13, CDRSB, RAVLT immediate and mean volume (CP) of Inferior Temporal were selected in at least 75% of the times. SPDAS.BIC selected 21.95 features with a Jaccard Index of 0.28.  $A\beta_{1-42}$ , CDRSB, RAVLT immediate were selected in all the cases. ADAS13, FAQ, mean volume (CP) of inferior temporal, absolute difference of volume (CP) of Pars Triangularis were selected more than 15 times. Finally, Unicox analysis selected a mean of 116.50 features with a JI of 0.69. 61 features were selected every time, all the CSF Measures and Cog-assessments scores are part of this list. The best performance of the Follow-up Times c-index was found on the RIDGE model with 0.91(0.89,0.92), but Coxnet/LASSO found a better performance

over c-index Risks. LASSO reported an ACC = 0.75(0.71,0.79), AUC = 0.84(0.81,0.88), SEN = 0.75(0.69,0.81), SPE = 0.75(0.70,0.80), c-index Follow-up Times = 0.77(0.75,0.79) and c-index Risks = 0.66(0.63,0.69). All the p-values on the Log-rank test were zero.

We performed a detailed analysis of the set of selected features across ML methods. The analysis of the RHOCV reported that eight features were common on 50% of the sets. To evaluate the importance of these features as a risk factor for MCI to AD conversion, we refit the Cox model using them. We then reported the hazard ratios (HR) and their corresponding 95% CI:  $A\beta_{1-42}$ : HR = 0.9989 (0.9984, 0.9993), CDRSB: HR = 1.45 (1.21,1.73), RAVLT immediate: HR = 0.95 (0.93, 0.97), ADAS13: HR = 1.07 (1.04, 1.10), FAQ: HR = 1.04 (1.00, 1.08), mean volume (CP) of Inferior Temporal: HR = 0.63 (0.50, 0.79), mean volume (CP) of Entorhinal: HR= 0.79 (0.63, 0.99), mean cortical thickness SD of Bankssts: HR= 1.42 (1.08, 1.86). A heatmap representation with the ten features correlation with the outcome can be found in Figure 4.14. Table 4.10 provides more details of the ten characteristics. The last two columns of Table 4.10 show the rank of the features of the 6 ML approaches. The MV HR and the UV HR correspond to the Hazard ratios of the feature inside a Multivariate model and the HR computed by the univariate approach respectively. It is clear that feature ranking and importance depended on the ML method.

### 4.3 Osteoarthritis Initiative: OAI

Wrapper Methods							
	BSWiMS	LASSO	Ridge	Elasticnet	GSPDAS	SPDAS	SPDAS.BIC
<i>ACC</i>	0.81 (0.79, 0.84)	0.8 (0.77, 0.82)	<b>0.82</b> <b>(0.8, 0.85)</b>	<b>0.82</b> <b>(0.8, 0.85)</b>	0.78 (0.75, 0.8)	0.78 (0.75, 0.8)	0.78 (0.75, 0.81)
<i>CIFT</i>	0.73 (0.7, 0.76)	0.69 (0.65, 0.72)	<b>0.79</b> <b>(0.76, 0.82)</b>	<b>0.79</b> <b>(0.76, 0.82)</b>	0.66 (0.63, 0.69)	0.64 (0.6, 0.67)	0.63 (0.6, 0.66)
Filter Methods							
	Cox.BSWiMS	Cox.LASSO	Cox.GSPDAS	UniCox			
<i>ACC</i>	0.78 (0.75, 0.8)	<b>0.8</b> <b>(0.77, 0.82)</b>	0.78 (0.75, 0.8)	0.76 (0.73, 0.78)			
<i>CIFT</i>	<b>0.85</b> <b>(0.82, 0.88)</b>	0.82 (0.79, 0.86)	0.66 (0.63, 0.7)	0.72 (0.69, 0.76)			

Table 4.12: Main classification (Accuracy ACC) and survival stats (c-index FT) for all the models with Wrapper and Filter Methods. The complete stats for all the models are shown in the Table 4.13 for wrapper methods and Table for filter Methods. Best scores for each stat are bolded.

Table 4.13 shows some stats for wrapper and filter methods methods in OAI and Table 4.12 summarizes the the classification and survival stats of the both strategies. BSWiMS selected a mean of 76.30 features with a Jaccard Index of 0.59. It took 73.67 seconds to build each model. Thirteen features were selected in every model and 64 were selected in more than half of the iterations. The thirteen features are the following: Mean and absolute difference and raw measure for Kellgren and Lawrence grades, calc of Function, Sports, and Recreational

Activities Score of KOOS (KOOS Sports), grades for osteophytes femur lateral compartment), mean of sclerosis grades for tibia medial compartment, KOOS Quality life and symptoms score, osteophytes grades for femur medial compartment and sclerosis grade for tibia lateral compartment.

Wrappers						
	<i>ACC</i>	<i>AUC</i>	<i>SEN</i>	<i>SPE</i>	<i>CIRisks</i>	<i>CIFU</i>
<i>BSWiMS</i>	0.81 (0.79, 0.84)	0.84 (0.82, 0.87)	0.82 (0.78, 0.86)	0.81 (0.78, 0.84)	0.52 (0.5, 0.54)	0.73 (0.7, 0.76)
<i>LASSO</i>	0.8 (0.77, 0.82)	0.83 (0.81, 0.86)	0.85 (0.81, 0.89)	0.77 (0.74, 0.8)	0.52 (0.5, 0.55)	0.69 (0.65, 0.72)
<i>Ridge</i>	<b>0.82</b> <b>(0.8, 0.85)</b>	0.85 (0.82, 0.87)	0.84 (0.79, 0.88)	<b>0.82</b> <b>(0.78, 0.84)</b>	<b>0.53</b> <b>(0.5, 0.55)</b>	<b>0.79</b> <b>(0.76, 0.82)</b>
<i>Elasticnet</i>	<b>0.82</b> <b>(0.8, 0.85)</b>	<b>0.85</b> <b>(0.82, 0.87)</b>	0.84 (0.79, 0.88)	0.81 (0.78, 0.84)	<b>0.53</b> <b>(0.5, 0.55)</b>	<b>0.79</b> <b>(0.76, 0.82)</b>
<i>GSPDAS</i>	0.78 (0.75, 0.8)	0.8 (0.77, 0.83)	0.87 (0.83, 0.91)	0.74 (0.7, 0.77)	0.52 (0.5, 0.54)	0.66 (0.63, 0.69)
<i>SPDAS</i>	0.78 (0.75, 0.8)	0.81 (0.79, 0.84)	<b>0.88</b> <b>(0.83, 0.91)</b>	0.73 (0.7, 0.77)	0.52 (0.5, 0.54)	0.64 (0.6, 0.67)
<i>SPDAS.BIC</i>	0.78 (0.75, 0.81)	0.82 (0.79, 0.84)	0.87 (0.83, 0.91)	0.74 (0.71, 0.77)	0.52 (0.5, 0.55)	0.63 (0.6, 0.66)
Filters						
	<i>ACC</i>	<i>AUC</i>	<i>SPE</i>	<i>SEN</i>	<i>CIRisks</i>	<i>CIFU</i>
<i>Cox.BSWiMS</i>	0.78 (0.75, 0.8)	0.81 (0.78, 0.84)	0.86 (0.81, 0.89)	<b>0.74</b> <b>(0.71, 0.78)</b>	<b>0.85</b> <b>(0.82, 0.88)</b>	<b>0.85</b> <b>(0.82, 0.88)</b>
<i>Cox.LASSO</i>	<b>0.8</b> <b>(0.77, 0.82)</b>	<b>0.83</b> <b>(0.8, 0.85)</b>	0.87 (0.82, 0.9)	0.77 (0.73, 0.8)	0.82 (0.79, 0.86)	0.82 (0.79, 0.86)
<i>Cox.GSPDAS</i>	0.78 (0.75, 0.8)	0.8 (0.77, 0.83)	<b>0.87</b> <b>(0.83, 0.91)</b>	0.74 (0.7, 0.77)	0.66 (0.63, 0.7)	0.66 (0.63, 0.7)
<i>UniCox</i>	0.76 (0.73, 0.78)	0.77 (0.74, 0.8)	0.86 (0.82, 0.9)	0.71 (0.68, 0.75)	0.72 (0.69, 0.76)	0.72 (0.69, 0.76)

Table 4.13: Classification and survival stats for all the models with Wrapper and Filter Methods in the OAI analysis. Best scores for each stat are bolded.

The Second method was created with LASSO strategy, it uses 31.45 features on average with a low Jaccard index of 0.33. LASSO took a mean of 48.55s to build each model. In this case, the number of features selected in all the iterations was just 4: mean and difference of Kellgren and Lawrence grades, KOOS Sports, osteophytes grades for femur lateral compartment. 20 features were selected in more than half of the times. The third model, developed with RIDGE strategy selected a mean of 214 features with a Jaccard Index of 0.84. The mean time used to build each model was 3.45 seconds. Besides its big number of selected features, just 105 features were selected in all the iterations and 120 more were chosen in more than half of the repetitions. Fourth model uses ELASTICNET that is almost the same than the third one. It selected the same amount of features in more than the half of the times but it

chose eleven more features in all the iterations. ELASTICNET took 3.56 seconds on average to build the model of 216.05 features with a Jaccard index of 0.86.

The fifth model uses GSPDAS algorithm by BeSS. It selected 39.45 features on average for each model and finishes with a Jaccard Index of 0.22. It only took a mean 4.95 seconds to build each model. It just used one feature in all the models, the mean of the grades of Kellgren and Lawrence (KL). Two other features were used in more than the half of iterations KOOS Sports and the difference between Kellgren and Lawrence scores in the both knees. Sixth model also uses BeSS but this time with SPDAS algorithm.

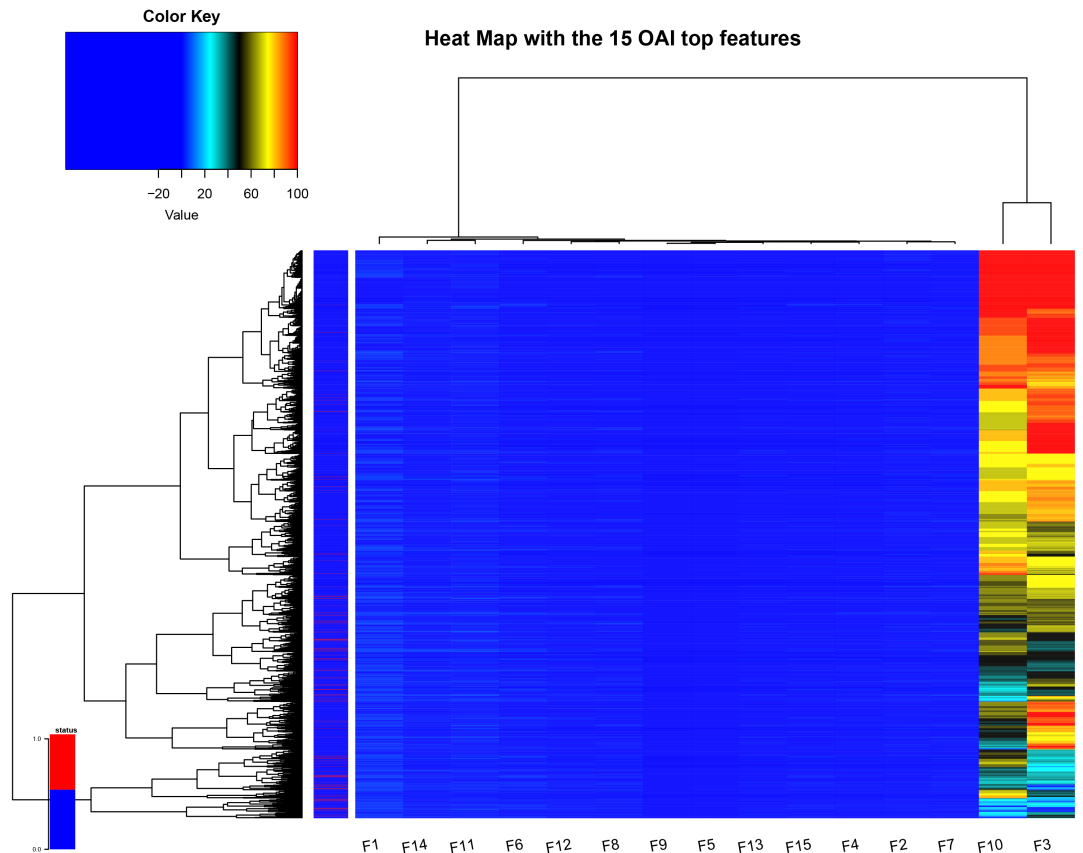


Figure 4.15: A heat map representation of the features associated with TKR outcome on OAI patients. The figure shows the 15 features selected by all the 8 methods (Seven wrappers and Unicox) in at least in the half of the iterations (horizontal axis) and subjects on the vertical axis. (F1) Mean KL, (F2) Absolute difference KL, (F3) KOOS Sports, (F4) absolute difference osteophytes grades of femur lateral compartment, (F5) absolute difference lateral tibial plateau margin, (F6) Mean osteophytes grades of femur lateral compartment, (F7) absolute difference medial minimum JSW, (F8) Mean sclerosis grades tibia medial compartment, (F9) mean x coordinate of minimum JSW, (F10) KOOS Quality of life, (F11) absolute difference of FTA, (F12) raw feature osteophytes grades of femur medial compartment, (F13) absolute difference sclerosis grades of femur medial compartment, (F14) Raw difference between position  $x=150$  and  $x=850$ , (F15) Raw osteophytes grades femur lateral compartment.



The main difference of the both results is the time that takes to build each model in the repetitions which took 124.62s. 4.75 features were selected on average in each model and the Jaccard Index of 0.43. It uses the same features than GSPDAS algorithm. SPDAS with BIC criterion selected a mean of 9.35 features with 0.32 of Jaccard Index. In this case, the same one feature was selected in all the iterations but 5 other features were used in the half of the iterations: the difference between Kellgren and Lawrence grades, KOOS Sports, osteophytes grades for femur lateral compartment, absolute difference lateral tibial plateau margin and mean x coordinate of minimum JSW. Eight model uses Unicox analysis to study the case. In this case, it uses 72.35 features with a Jaccard Index of 0.48. Just Kellgren and Lawrence grades were used in all the repetitions. After all the iterations with all the models we chose the group of features that were selected in more than the half of the iterations. We build a Cox Model for those features to build a final Cox Model that studies the OAI TKR outcome. In total, 160 repetitions were made. With their HR and 95%CI were selected in more than 80 repetitions: Mean and absolute difference of Kellgren and Lawrence grades HR = 1.41 (1.2, 1.67), KOOS Sports HR = 1.27 (1.05, 1.53), absolute difference HR = 0.98 (0.97, 0.99) and mean HR = 1 (0.4, 2.48) of osteophytes grades for femur lateral compartment HR = 0.98 (0.8, 1.2), absolute difference lateral tibial plateau margin HR = 1.05 (0.87, 1.28), difference between medial JSW between kness HR = 0.97 (0.8, 1.18), mean sclerosis grades of tibia medial compartment HR = 1.14 (0.97,1.35) , mean x coordinate of minimum JSW HR = 3.22 (0.38, 27.38), KOOS Quality of Life score HR = 1 (0.99, 1.01), absolute difference femoral tibial angle (FTA) reading HR = 1.24 (1.15, 1.35), osteophytes grades of femur medial compartment raw measure HR = 1.14 (0.95, 1.36), absolute difference sclerosis grades femur medial compartment HR = 1.14 (0.95, 1.36), Difference between x position 150 and 850 for JSW in the right knee HR = 1.04 (0.94, 1.15), osteophytes grades femur lateral compartment raw measure in the left knee HR = 0.92 (0.69, 1.23). Figure 4.15 shows the relationship heatmap of those characteristics with the outcome.

## 4.4 Prognostic Wisconsin Breast Cancer Database

In this case, we found some results similar to the ones in the literature. Table 4.14 shows the Wrappers method's results belong to its 95% confidence intervals (CI) for the Prognostic Wisconsin Breast Cancer Database. Table 4.15 shows the results on filters. In the next paragraphs, we will describe the models built, their stats and their 95% confidence intervals between parenthesis (i.e xx(yy, zz)). The first of the models were analyzed with BSWiMS strategy. It selected a mean of 4.25 features by iteration with a Jaccard index of 0.28. BSWiMS selected seven features in more than half of the times, perimeter and radius of the third cell nucleus, perimeter, and area of the first cell nucleus and the tumor size. Being the perimeter of the third cell the most selected feature in 90% of the iterations. BSWiMS got ACC =0.64(0.57,0.7), SEN =0.6(0.44,0.74), SPE=0.65(0.57,0.72) and C-Index FU 0.81(0.78,0.83). The cox model created with the same features (The filter method with BSWiMS) uses the same number of features but it got ACC =0.63(0.56,0.7), SEN =0.64(0.49,0.77), SPE=0.63(0.55,0.71) and C-index on Risk and Follow-up Times of 0.74(0.7,0.78).

Method	ACC (95% CI)	SEN (95% CI)	SPE (95% CI)	C-index Risks (95% CI)	C-Index FU (95% CI)
I	0.65 (0.58,0.72)	0.66 (0.51,0.79)	0.65 (0.57,0.72)	<b>0.55</b> <b>(0.5,0.6)</b>	0.80 (0.77,0.83)
II	0.69 (0.62,0.76)	0.74 (0.6,0.86)	0.68 (0.59,0.75)	<b>0.55</b> <b>(0.5,0.6)</b>	0.81 (0.78,0.84)
III	<b>0.69</b> <b>(0.62,0.76)</b>	0.70 (0.55,0.83)	0.69 (0.61,0.76)	<b>0.55</b> <b>(0.5,0.6)</b>	<b>0.83</b> <b>(0.8,0.86)</b>
IV	0.68 (0.61,0.74)	0.70 (0.55,0.83)	0.67 (0.59,0.74)	<b>0.55</b> <b>(0.51,0.6)</b>	<b>0.83</b> <b>(0.81,0.86)</b>
V	0.66 (0.59,0.72)	0.70 (0.55,0.83)	0.64 (0.56,0.72)	0.46 (0.41,0.5)	0.69 (0.65,0.73)
VI	0.71 (0.64,0.77)	<b>0.77</b> <b>(0.62,0.88)</b>	<b>0.70</b> <b>(0.62,0.77)</b>	0.51 (0.46,0.56)	0.76 (0.73,0.79)
VII	0.71 (0.64,0.77)	<b>0.77</b> <b>(0.62,0.88)</b>	<b>0.70</b> <b>(0.62,0.77)</b>	0.51 (0.46,0.56)	0.76 (0.73,0.79)

Table 4.14: Classification and survival stats for filter methods that analyzed the Prognostic Wisconsin Breast Cancer Database. I = BSWiMS, II = LASSO, III = RIDGE, IV = ELASTICNET, V = GSPDAS (BESS), VI = SPDAS (BESS.SEQUENTIAL), VII = SPDAS.BIC (BESS.SEQUENTIAL.BIC). Best scores for each stat are bolded.

Method	ACC (95% CI)	SEN (95% CI)	SPE (95% CI)	C-index Risks (95% CI)	C-Index FU (95% CI)
I	0.63 (0.56,0.7)	0.64 (0.49,0.77)	0.63 (0.55,0.71)	<b>0.74</b> <b>(0.7,0.78)</b>	<b>0.74</b> <b>(0.7,0.78)</b>
II	0.66 (0.59,0.73)	0.66 (0.51,0.79)	0.66 (0.58,0.74)	0.73 (0.69,0.77)	0.73 (0.69,0.77)
III	<b>0.67</b> <b>(0.6,0.74)</b>	<b>0.68</b> <b>(0.53,0.81)</b>	<b>0.67</b> <b>(0.59,0.74)</b>	0.68 (0.64,0.73)	0.68 (0.64,0.73)
IV	0.63 (0.55,0.69)	0.66 (0.51,0.79)	0.62 (0.53,0.69)	0.72 (0.68,0.76)	0.72 (0.68,0.76)

Table 4.15: Classification and survival stats for filter methods that analyzed the Prognostic Wisconsin Breast Cancer Database. I = Cox with BSWiMS, II = Cox with LASSO, III = Cox with BESS IV = Univariate Cox. Best scores for each stat are bolded.

The second wrapper method used was LASSO and it selected a mean of 6.85 features in each iteration with a Jaccard index of 0.27. Although the Jaccard number is smaller, the number of features selected in more than half of the occasions remains. However, the selected characteristics are different. 75% of the time the lymph node status which is the number of positive axillary lymph nodes observed at the time of surgery is selected. Followed by

the perimeter of the third cell that was selected 70% of the time. Subsequently, the size of the tumor and the symmetry of the first cell chosen in 65%. Finally, the perimeter of the second cell by 60% and the fractal dimension of the first cell by 55%. LASSO find an ACC =0.69(0.62,0.76), SEN =0.74(0.6,0.86), SPE=0.68(0.59,0.75) and C-Index Follow-Up=0.81(0.78,0.84).

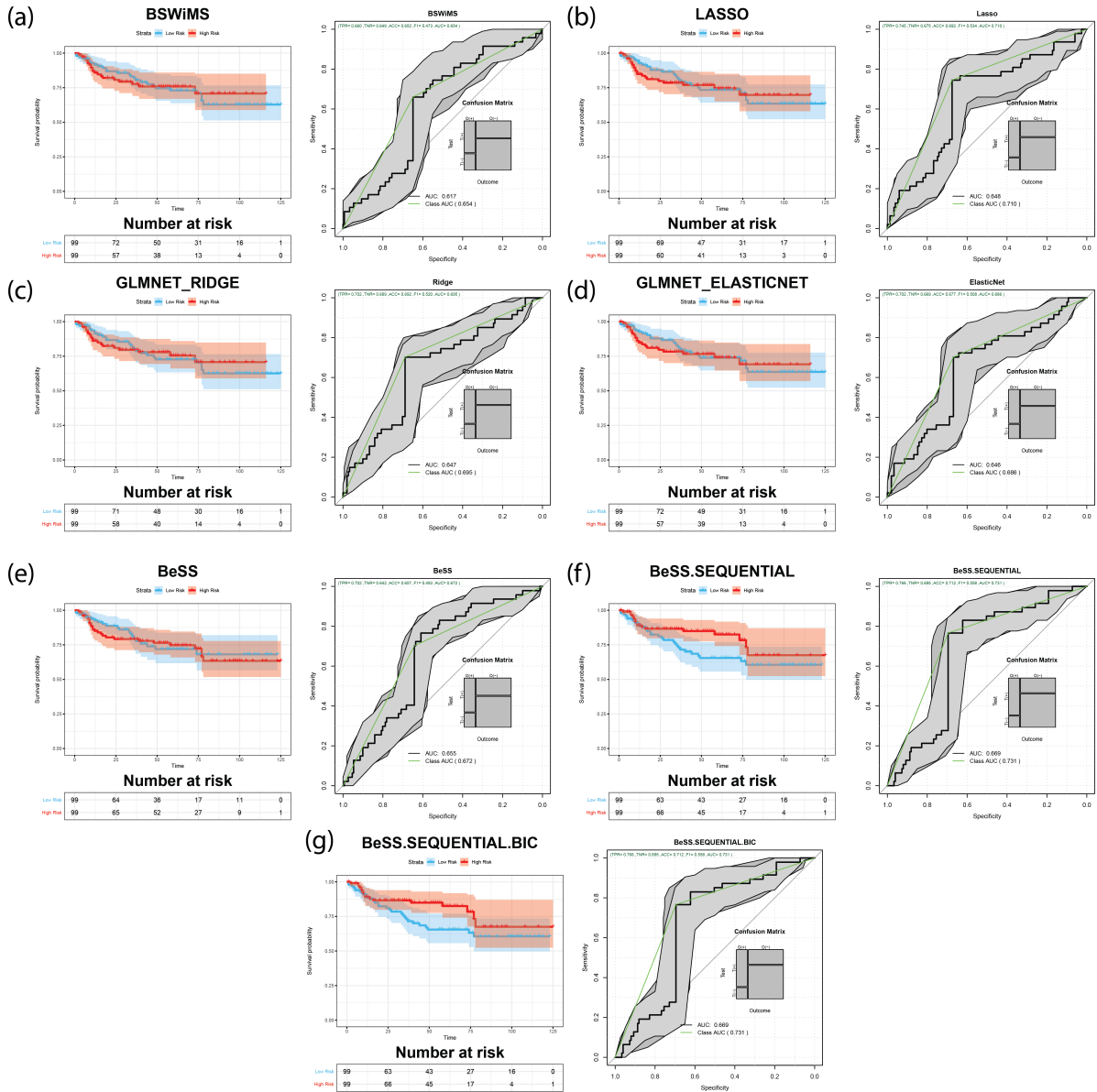


Figure 4.16: KM and ROC curves of the wrapper models built with Prognostic Wisconsin Breast Cancer Database. (a) BSWiMS (b) LASSO (c) RIDGE (d) ELASTICNET (e) GSP-DAS (BeSS) (f) SPDAS (g) SPDAS with BIC

The third model constructed uses the RIDGE method. It selected the biggest number group of features with 31.55. It's Jaccard index is very big with 0.97. 27 of the 32 features

were selected every time and just the other five range from 0.95 to 0.85 percent of the iterations. Ridge got an ACC = 0.69(0.62,0.76), SEN = 0.70(0.55,0.83), SPE = 0.69(0.61,0.76), and a C-Index FU = 0.83(0.8,0.86). Almost with the same stats the fourth method, ELASTIC-NET, used 31.40 features by mean in each model and got a Jaccard index of 0.96. In this case, just 22 characteristics were selected every time but the other ones were select from 0.95% to 0.90%. ELASTICNET found an ACC = 0.68 (0.61, 0.74), SEN = 0.70 (0.55, 0.83), SPE = 0.67 (0.59, 0.74), C-Index FU = 0.83 (0.81, 0.86).

The fifth model was developed with BeSS. The default version of BeSS which uses de GSPDAS algorithm selected a mean of 25.50 features with a Jaccard Index of 0.68. It selected 4 features all the times, Lymph Node status, Fractal dimension of the first cell and compactness of the third and second cells. All the other features were selected at least in half of the iterations. GSPDAS found an ACC = 0.66 (0.59, 0.72), SEN = 0.70 (0.55, 0.83), SPE = 0.64 (0.56, 0.72) and a C-Index FU = 0.69 (0.65, 0.73). Its C-Index Risks was the only one which was lower than 0.50 with 0.46 (0.41, 0.5). The sixth model uses BeSS once again, but this time with the SPDAS algorithm. It found the least mean number of features with 1.95 on each model but its Jaccard index is not the lowest (0.23). None of the features were selected in at least half of the iterations, the most selected feature was Lymph Node Status in 40% of the models. Model VI got an ACC = 0.71 (0.64, 0.77), SEN = 0.77 (0.62, 0.88), SPE = 0.70 (0.62, 0.77) and C-Index FU = 0.76 (0.73, 0.79).

SPDAS with BIC also known as Model VII selected just 2.40 features on average. Its Jaccard index is the lowest of the models with 0.20. It also did not select any of the features in all the models, and the same feature was selected in 8 of the 20 models. The stats of this model were the same as the last method, the differences between the stats of these models were manifested from the 5th decimal place. Figure 4.16 shows the KM and ROC curves of all the wrappers methods (7 models).

## 4.5 Prognostic San Jose Hospital Breast Cancer Database

The last section of this chapter reports the results of the Prognostic San Jose Breast Cancer Dataset analysis. As took place in other of the experiments not all the methods could finish the task. In this case and with the configuration of the data detailed in Chapter 3, just BSWiMS could fit the model and get some results. This method chose a large number of features with an average of 17.6 for each model. BSWiMS holds its Jaccard index in 0.10 taking about 4.90 seconds on average to build the models. BSWiMS finished with an ACC = 0.67 (0.55,0.78), AUC = 0.66 (0.53, 0.78), SEN = 0.76 (0.50, 0.93), SPE = 0.64 ( 0.50, 0.77), CI Risks = 0.52 (0.43, 0.60) and CI Follow-Up Times of 0.72 (0.64, 0.80). A total of 160 features were used. Of these, 82 features were used only in one repetition. In contrast, only three characteristics were used in more than 10 iterations (half of the models produced). Absolute differences between mediolateral oblique view (MLO) and Craniocaudal view (CC) of z range of level HH 1 of Wavelets Transforms (F1, 18 times), the same measure but this time for LH 1 level (F2, 13 times), and dynamic range of HH 1 Level (F4, 10 times), were the most selected features. We took them to analyze them separately from the other.

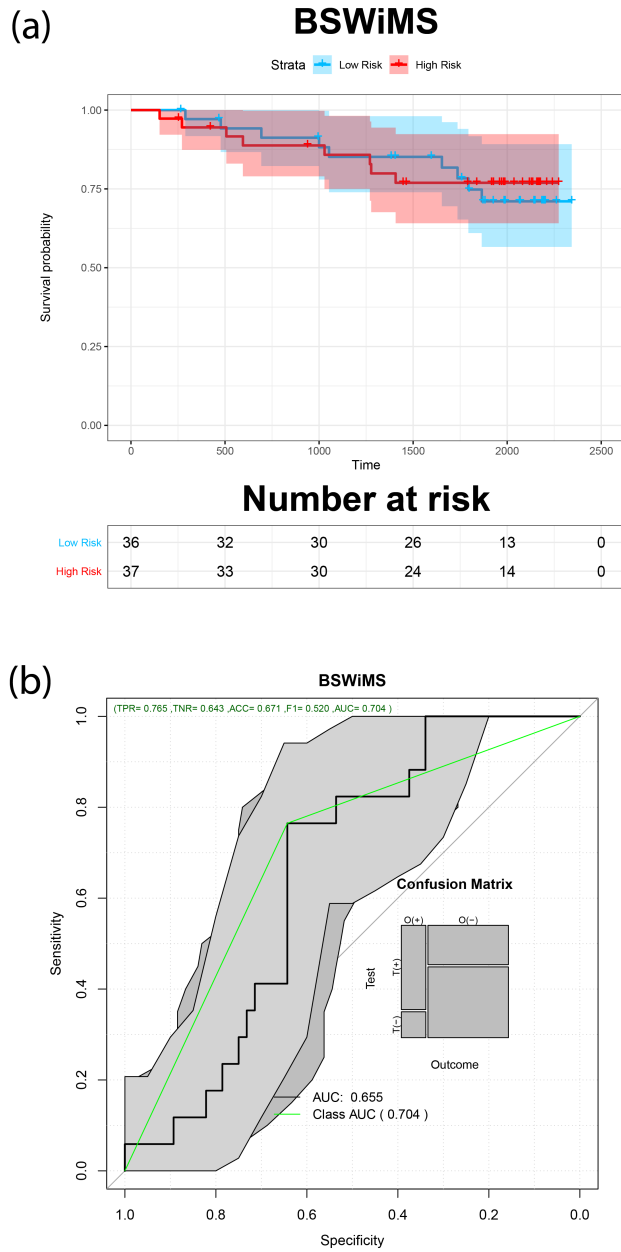


Figure 4.17: (a) KM and (b) ROC curves for the only model which got results on the San Jose Survival analysis, BSWiMS.

We build a Multivariate Cox Model with those features and the Concordance Index of the resultant model is 0.743 and its p-value on the log-rank test is  $p = 6e^{-05}$ . The HR for those features were also calculated. F1 has a HR = 1.056 (0.6533, 1.708), F2 has HR = 2.001 (1.0393, 3.854) and F3 has 1.004 (0.9994, 1.009). The other methods could not get any stats because they can fit a model with this data. In the discussion section, we will detail and explain why we think that this happened. Figure 4.17 shows the KM and ROC Curves of the only resultant model BSWiMS.



# Chapter 5

## Discussions and Conclusions

This chapter concludes this thesis trying to discuss the results found from the analysis of different data through the ML-Survival models comparison tool: Cox Benchmarking. Regarding that each experiment has different objectives, they will have its own subsection where their discussion will take place. Each subsection will contain every field that we consider as a field in which Cox Benchmarking can help make a decision. Besides, the experiment's results will also be analyzed and compared to those in previous works or to the absolute truth in the case of data simulation experiments. This comparison takes place to ensure that our benchmarking process is using great methods that can find solutions like the ones that are scientifically published. In all the cases, we find that our strategy and our methods got a statistically comparable behavior to those in the scientific literature but in our case, we have the advantage of allowing comparison of some methods in the same analysis. Generalized discussions and conclusions regarding the main objective of the thesis are found in the last section of this Chapter.

On the other hand, within the comparison of methods, you will be able to notice the main differences between the methods, and the great advantage of the special behavior of each of them in the different problems to be treated. The use of different types of data allowed the assessment and comparison of the performance of the different methods to be fair. As we can see in the following sections, each method has a different performance depending on the type of data, its size, amount of features or the particular distribution that each situation has. In the end, although we cannot select one of the methods as better than the previous ones, we can suggest that the three default methods work correctly in the survival analysis and that it depends on the user the correct selection of the tool that is the most indicated. In the conclusions section, we can find more details about our judgment on the methods used. Finally, this section is about creating new research. With the questions and limitations found in this thesis, we propose possible problems to be solved in the new research on Future work section. Issues that could be taken as a starting point for more scientific work and that could even be indexed publications of the same authors of the thesis. Future work and possible research left will be detailed as part of the discussion and limitations of each section.

## 5.1 Discussion

As was observed in each of the results, the CoxBenchmarking method was coupled to the different situations, whether clinical or simulated. The analysis of these data sets with this tool only required the preparation of the data. This, in summary, requires the existence of two variables; the first is the true value of the event (status) and the second is the time until the same event (time-to-event). The ability of CoxBenchmarking to use censored information (because of using the Cox model) allows the amount of data we can use to be greater. The development of this tool focused on facilitating the survival analysis of any kind of information, but mainly for the help in CAD systems to investigate the prognosis of chronic degenerative diseases. All experiments used almost the same strategy, and always used the same methods to analyze and compare the performance of the tool. The results were already reported and discussed in the previous sections, so this segment will be dedicated to summarize the results and find common characteristics of the behavior of our tool. Next, we are going to discuss the results of this thesis considering that the main objective is the evaluation of Machine learning methods that build survival models based on the Cox Model and how they can influence the decisions on the prognosis of chronic degenerative diseases based on the results of the clinical cases studied for this thesis.

**ML Method Validation** RHOCV was decided to be the evaluation method for the machine learning strategies since this evaluation technique allowed measuring the effect of the training set on feature selection, and, at the same time, permitted a training-set unbiased evaluation of the test performance. The strategy chosen for almost all the experiments was an evaluation that created 20 random splits of the dataset into training and test set. For each such split, the train fraction was 0.7 and the remaining 0.3 used for testing. The reported results in all the experiments indicated that ML methods selected models with very different internal features but with very similar statistics between them. Model sizes varied considering the experiment and from method to method. There were very complex models and with very high features (ELASTICNET and RIDGE more than 200 features) and very simple models such as SPDAS with BIC criterion that selected in the least case just 3 features on average.

The ability of RHOCV to provide a more adequate evaluation of the methods, allowed us to know the different behaviors of the methods. Provoking that in each iteration, the feature selection process resulted in something different. Which allows a more adequate approximation of its performance. Regardless of the type of data, the stat values were always similar between the models. With the exception in some statistics such as the Concordance index of Follow-Up Times where in most cases, RIDGE and ELASTICNET were superior to the other methods. CoxBenchmarking provides confidence intervals and statistics that demonstrate the effect size of the results. Both values exhibit a reliable tool for comparing the algorithms. The CoxBenchmarking plot tool is based on these measurements to be able to graphically summarize the statistics and help us in ordering the methods by their performance.

Regarding this classification performance, it is important to note that proportional hazard models were not designed for classifications tasks. To address this issue, in all the experiments and models we assumed that subjects predicted to have an increased risk of conversion ( $\text{Risk} > 1$ ) should correspond to the event occurrence, while subjects at low-risk prediction ( $\text{Risk} \leq 1$ ) should correspond to censored patients. Consider this strategy let us evaluate the



accuracy, sensitivity, specificity, and AUC of the Cox regression models and compare them to other scientific works that investigate risk classification in prognosis context. The reported AUC performance was used in some of our experiments to be compared to other methods. Regarding the reproducibility of feature selection, the Jaccard analysis indicated that the internal structure of the Cox models depended on the training set and the set of features used on each experiment. Depending on the case the Jaccard Index range from values near to 0.1 to big values near to 1. Meaning that the overlapping of the features between models does not affect at all the statistics that resulted in the final model, but the Jaccard Index statistic permitted us to comprehend how the methods operated in the feature selection process.

**Benchmarking** As mentioned earlier, the comparison process is the most important objective of the thesis. We are convinced that this process has been successfully completed with the implementation of this method. We use some of the available algorithms to build models that employ proportional hazard models. However, the method is open for implementation with new techniques and comparison of results with other works. We were able to realize that the methods worked similarly in almost all cases. Depending on the nature of the data and the informational characteristics, the models found very similar statistics and common characteristics among them. Even so, in cases such as in the simulation experiment and BRCA San Jose, where the number of characteristics is very high and exceeded or matched the number of patients; ELASTICNET, RIDGE, GSPDAS, SPDAS, and SPDAS.BIC could not fit a solution. We can justify this behavior because of the type of algorithm with which they work and by the amount of noise that may exist in these experiments. Therefore, we suggest that LASSO and especially BSWiMS are methods are more sensitive to fit a solution in data sets in which the noise is high and the number of features exceeds the number of subjects. This work also aimed to improve the knowledge of the role of these features in the event; hence we reported the list of the top biomarkers along with their standardized HR associated with the event, may help physicians predict how far a specific patient in their prognosis process is. Considering that there are distinct models regarding its complexity and that despite they found similar statistics, we believe that the methods base their importance and their differences in the designation of coefficients for these selected variables. Each of the methods had a highly respectable performance in different experiments. However, it was noted that LASSO and BSWiMS were the most stable models and that they managed to be among the best statistics in almost all experiments. On the other hand, ELASTICNET and RIDGE have a great ability to predict the concordance of time where they had much superiority compared to the other methods in two experiments. Finally, BeSS and its three algorithms have more sensitivity in cases where the amount of data is not large and the amount of noise is minimal. The filter analyzes helped to recognize that the use of the methods does improve the quality and, above all, better approximate the value of the coefficients of the variables in some cases.

## 5.2 Simulation data

This experiment is a fundamental piece in the development of the thesis. Although its clinical importance is null, and its implementation does not require as much knowledge or development time as the development of the tool per se, the results of the tool were the starting point

for the use of the tool in the following data. Because we are the ones who created the simulation variables, we are sure of the expected result when analyzing the data from the simulation. It is well known that although in a simulation scenario the influence of randomness is included (so that the generation of data is more real), this does not change the factors that we decide to influence an expected outcome; For this reason, each of the methods is expected to select these characteristics and their influence to be measured in the same way in which they were used to generate the data.

**Random features** One of the main concerns in the simulation process is the random variables. In this case, we summarized the selection of random variables in the Table 5.1. The table details the number of random variables selected in more than the half of the repetitions for each experiment. In the other hand, Table 5.2 details the false discovery rate (true variables / random variables, both in more of the half of repetitions) and the mean of variables used in each model for the experiments.

Problem	4 features		11 features	
	Real	Random	Real	Random
$N \times N$	4	0	8	1
$N \times (100 - N)$	4	9	7	2
$N \times (1000 - N)$	4	10	7	5

Table 5.1: Random variables selected in more than the half of the iterations for all the experiments

Method	4 features			8 features		
	8	100	1000	21	101	1001
<b>BSWiMS</b>	2.9(0)	2.6(0)	2.25	4.5(0)	4.7(0)	3.4
<b>LASSO</b>	6.6(0.5)	9.9(0.66)	25.5	12.4(>1)	21.7(>1)	31.4
<b>RIDGE*</b>	7.9	96.9	-	20.95	99.95	-
<b>ELASTICNET*</b>	7.9	97.1	-	20.95	100.1	-
<b>GSPDAS</b>	5.3(0.25)	56.3(>1)	-	11.3(>1)	58.2(>1)	-
<b>SPDAS</b>	3.4(0)	3.25(0)	-	6.4(0)	4.9(0)	-
<b>SPDAS.BIC</b>	3.05(0)	4.4(0)	-	6.2(0)	6.4(0)	-

Table 5.2: Mean variables selected by each method in the models. False discovery rate is shown inside the parenthesis. FDR is calculated with the ratio of random variables selected in more than the half of the models.

**Feature selection** Considering the above, the methods must generate models that look like the ground truth used to generate the simulation. That is, they need to select the features that influence the event, especially those that have a greater effect size within the survival model that generated the probability of the event. In this case, regarding Table 3.2 where we detailed

the characteristics that will be related to the event in the simulation. It is considered that the characteristics with an effect size greater than one (that are also the two characteristics with binomial distribution) should be the most selected features by all the methods. Depending on the case, whether 4 or 10 variables are used, these two characteristics should be in the models. The other effect sizes try to reduce the magnitude of the measurements so that the weight of their values in the survival equation is equivalent. However, in those variables, we wish to at least find a Hazard ratio similar to those used for the simulation. Regarding this selection, we will ignore the selection of RIDGE and ELASTICNET functions for discussion, because these functions have always selected almost all the characteristics and their coefficients are what allows them to be similar to the other methods.

The first step for being able to know if the selection of these characteristics is possible and to be able to take into account if the simulation was developed properly, is the generation of a Cox model that uses all the characteristics and thus know the possible effect size that will be generated with this model. The size of the effect should be similar in magnitude and sense. The table 5.3 summarizes the Cox model coefficients generated from these characteristics with the simulated data with the 11 real characteristics. As well as the Hazard Ratio of each variable. The c-index of the Cox model is 0.754 with a log-rank p-value of  $2e-16$ . These results ensure that the data simulation was correct, that the expected variables should be selected, but that the vast majority of variables will be neglected, given the size of their effect. Depending on the number of random characteristics added to the model and the number of real variables, the number of times a real variable is selected in the model is defined. Taking into account that the probability that the random variables follow a binomial distribution is 0.5. Many of the random variables selected may have similarities to the data and therefore be selected. However, the fact that these variables were not used for the generation of simulation data, makes their influence on the outcome remain lower than the actual variables in the final model. Although random variables are selected, we must take into account the selection of real variables. Next, we will analyze the results depending on the number of features.

**4 features** In the three experiments, the Lesions and Surgeries feature were selected in almost all the models as was expected. The other two variables were selected by all strategies, but the number of times selected changes depending on the algorithm. Of the methods that select features, GSPDAS is the only one that uses all real variables in all 20 iterations. However, it also makes use of all random variables although almost always in less than half of the iterations. Instead, SPDAS and SPDAS with BIC select L, S, and ORtg at all times but SPDAS uses more random features than the selection with BIC. On the other hand, BSWiMS is the model that selected the least amount of variables regardless of how many random variables were used. However, he only selected L and S at all times. Offensive Rating was only selected in 11 iterations. LASSO selected the 4 variables when the amount of random data was not high, however, when the noise is higher, the Defensive Rating was selected only in two models. LASSO had an intermediate amount of randomly selected values. This number is between the amount of random data selected by GSPDAS and SPDAS. Finally, Cox's univariate analysis had very similar behaviors the amount of random data aggregated. The same variables as in SPDAS were selected but with almost no random values on both occasions. However, it selected more random values than BSWiMS. In general terms, BSWiMS was the one that selected the least random data and the Univariate analysis that most closely

approached the real model in terms of variables. This behavior is quite expected since the simulation used a Cox model to generate the survival of each subject.

**11 features** In the case of 11 characteristics, there are two behaviors (caused by the number of random variables added to the data set). When ten random variables are added, the real variables are selected in more times than the case with 90 random characteristics. This can be explained because in the first case the aggregate noise is very short compared to the number of features that explain the result. However, the behavior of each method is different and a bit similar to the case of 4 features. BSWiMS once again did not select random features, just in an iteration with the 90 variables case. It selected L, S in all iterations and BLK in more than half of the times. However, the other real features were selected but on very few times (always less than 6 iterations).

<b>Feature</b>	<b>Effect Size</b>	<b>Hazard Ratio</b>
<i>bmi</i>	0.0492158	1.0504470
<i>age</i>	0.0967256	1.1015581
<i>games</i>	-0.0004847	0.9995154
<i>minutes</i>	-0.0033824	0.9966233
<i>AST</i>	-0.0222014	0.9780433
<i>FGP</i>	-0.0495677	0.9516407
<i>BLK</i>	-0.4319036	0.6492720
<i>ORtg</i>	-0.0457648	0.9552666
<i>DRtg.</i>	0.0514445	1.0527907
<i>L</i>	0.8479576	2.3348732
<i>S</i>	1.3369331	3.8073488

Table 5.3: Cox Model summary using the eleven real features with the simulated data. First column shows the features related with the outcome, second column the effect size of each feature and the third the Hazard Ratio of each feature  $\exp(\beta)$

LASSO, in this case, had a better selection of the real characteristics. More than 7 real characteristics were used in all models of both cases. However, many random features were selected in the models although in less than half of the iterations. Games and minutes were used in 5 models. It selected L, S in all iterations and BLK in more than half of the time. However, the other real characteristics were selected but on very few occasions (always less than 6 iterations). LASSO, in this case, had a better selection of the real characteristics. More than 7 real characteristics were used in all models of both cases. However, many random features were selected in the models, although in less than half of the iterations. Games and minutes were used in 5 models. GSPDAS made use of all the variables and equal 7 real variables in all the models. On the other hand, SPDAS selected only 4 and its criteria with BIC in 3 models. It selected L, S in all iterations and BLK in more than half of the time. However, the other real characteristics were selected but on very few occasions (always less than 6 iterations). LASSO, in this case, had a better selection of the real characteristics. More than 7 real characteristics were used in all models of both cases. However, many random features were selected in the models, although in less than half of the iterations. Games and minutes

were used in 5 models. GSPDAS made use of all the variables and equal 7 real variables in all the models. On the other hand, SPDAS selected only 4 and its criteria with BIC in 3 models. GSPDAS selected much more random features compared to the other algorithms member of BeSS package. Univariate Cox selected just 4 features with big coefficients but it ignored the random features.

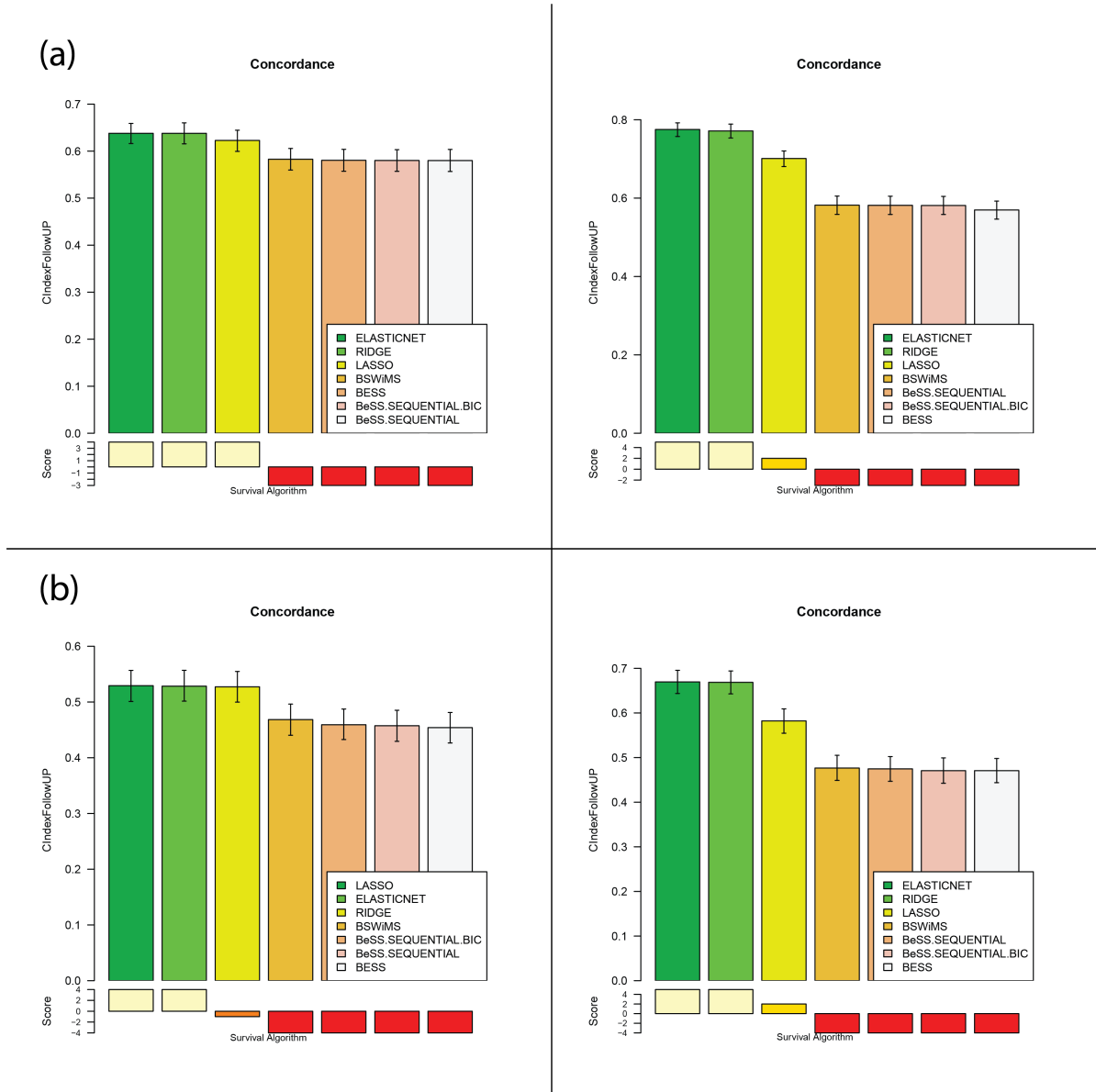


Figure 5.1: Bar plot for Concordance Index Follow-up Times. (a) 4 real and 4 random features (b) 4 real and 96 random features (c) 11 real and 10 random features (d) 11 real and 90 random features.

**ML Method Validation** In order to be sure that the RHOCV strategy uses all simulated data (players) as training and test data, we use a training fraction of 0.7 with 20 iterations. That is, for each method, 20 models are constructed from which its statistics are calculated and the

average of each one is used for the final prediction. This strategy is used to allow a fair and unbiased evaluation of test set performance. Since the simulated data have absolute truth, the complexity of the models depends entirely on each method. Several behaviors were found, having models with very high complexity in cases such as RIDGE and ELASTICNET that select all the uncorrelated characteristics and use the coefficients to give importance to each feature; and too simple models such as BSWiMS and BeSS with BIC that selecting the least number of features that allow for good performance. The most complex model was RIDGE and ELASTICNET model in case of the 11 real features and 90 random features selecting 100.10 on average; and the simplest was BSWiMS just choosing 2.90 features in the problem with 4 real features and 4 random features.

**Benchmarking** The Benchmarking process helped to realize that when we know the ground truth, the results are statistically similar. As visualized in the results, all statistics calculated in the models have values that are interlaced within the confidence intervals between them. None of the methods was statistically significant. However, there was a lot of difference in the number of features selected and in the Jaccard Index of each feature. The figure 5.2 shows the bar graph of the number of variables selected for all the methods in the first two experiments of 4 features and 11. On the other hand, a statistical difference was also found in the results of C-Index Follow-up times in all the experiments. Where the algorithms that belong to GLMNET were superior to the others, especially ELASTICNET and RIDGE. The figure 5.1 shows the bar graphs that demonstrate the difference between the methods in this statistic.

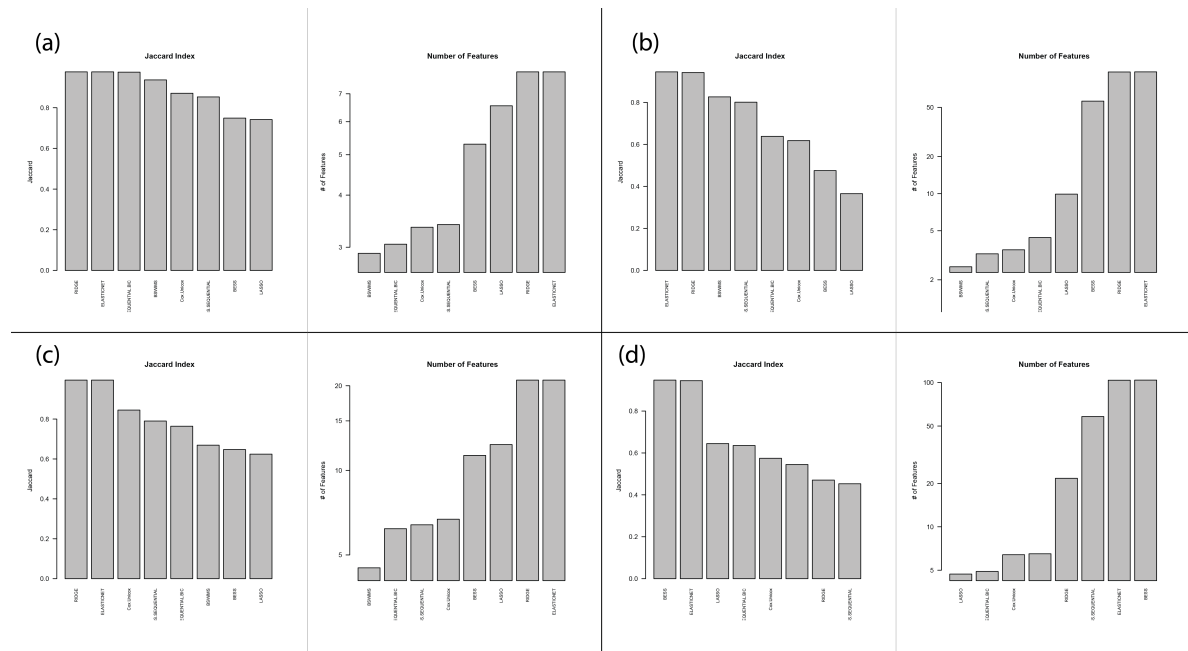


Figure 5.2: Bar plot for the number of features and Jaccard Index. Left size shows the barplot for Jaccard Index of each model and Right size shows the mean number of features (a) 4 real and 4 random features (b) 4 real and 96 random features (c) 11 real and 10 random features (d) 11 real and 90 random features.

**Limitations** These types of experiments help a lot to know how well the strategy works and if the behavior is as expected. However, sometimes the type of data used or the magnitude of the selected variables may result in the results not being as intended at the beginning. In our case, the difference in magnitude between the values made the estimation of coefficients important when simulating the probability of survival. We realized that many variables had such a small effect size that they were almost irrelevant and that is why they were hardly selected in the models. However, the two variables in which a large effect size was used intentionally were selected. The main limitation when conducting experiments with simulated data is the lack of data generation experience.

## 5.3 TADPOLE-ADNI

Results on this experiment were published in two Bioinformatics conferences. The first study were developed for the IEEE-EMBS International Conference on Biomedical and Health Informatics held from 19 to 22 May 2019 under the name "Studying MCI to AD Conversion Radiomics-Based Survival Indexes by Machine Learning". This research were expanded to be published in a journal of the same institution but it was rejected. Lately, the second study were developed for SPIE Medical Imaging Conference held on February 2020. The second study was published as "Prediction of MCI to AD risk of conversion survival models: qMRI vs CSF measures and cognitive assessments". A combination of both studies is in development for publication in a new journal.

### 5.3.1 Survival Models Associated with MCI to AD Conversion with qMRI features

In this work, we compared four different ML strategies that generated six proportional hazard models from qMRI structural analysis of MCI patients that either converted to AD or remained as MCI. The first three strategies - BSWiMS, LASSO, and BeSS - returned a Cox regression model and the set of features that were required to make an accurate estimation of the risk of conversion. The fourth strategy was a filter approach; hence selected features were used to build a standard Cox regression model. This last strategy was evaluated with the features generated by the first three methods and the features generated by a univariate Cox regression model. The performance of six proportional hazard models was evaluated using RHOCV and the most common features analyzed to report their importance in the rate of MCI to AD conversion.

**ML Method Validation** The RHOCV evaluation created 20 random splits of the dataset into training and test set. For each such split, the train fraction was 0.7 and the remaining 0.3 was used for testing. This evaluation strategy allowed the evaluation of the effect of the training set on feature selection, and, at the same time, permitted a training-set unbiased evaluation of the test performance. The reported results indicated that ML methods selected models with very different internal features. Model sizes varied from method to method and ranged from a minimum of 13 features to complex multivariate modeling based on 103 features. The six

qMRI-based models reported c-index ranging from 0.63 to 0.84. The simplest model overperformed the most complex one: 0.84 (CI 0.82,0.86) for Coxnet vs 0.63 (CI 0.60,0.66) for BeSS.

Regarding this classification performance, it is important to note that proportional hazard models were not designed for classifications task. To address this issue, we assumed that subjects predicted to have an increased risk of conversion ( $\text{Risk} > 1$ ) should correspond to true MCI to AD conversion, while subjects at low-risk prediction ( $\text{Risk} \leq 1$ ) should correspond to MCI-stable subjects. This strategy allowed us to evaluate the accuracy, sensitivity, specificity, and AUC of the Cox regression models. The reported AUC performance of the methods ranged from 0.67 to 0.73 for their potential to detect patients at risk of conversion. This performance was slightly lower to other methods based on SVM or Logistic Regression classifiers [42]. To test the impact of using all subjects in ROC AUC analysis, we conducted a post hoc experiment. In this experiment, we analyzed test prediction on MCI stable subjects whose last visit was greater than 4 years (146 no-event subjects did not meet the criteria). This change in selection criteria resulted in the ROC curve presented by Figure 5.3. We clearly see that Cox based conversion risk prediction had a similar performance (ROC AUC= 0.79) to previous works [70].

Regarding the reproducibility of feature selection, the Jaccard analysis indicated that the internal structure of the Cox models depended on the training set. The method with the largest Jaccard index (0.65) was based on the univariate filter, and also was the method with the largest set of features and with the poorest performance. The smallest models were returned by the BSWiMS strategy. It had a Jaccard Index 0.35, implying that only 35% of the features overlapped across different training sets. These results put forward that the discovery of risk factors associated with MCI to AD conversion depended on the training set and the machine learning strategy used to discover risk factors. This observation is supported by the literature, where different authors have reported a different set of features associated with MCI to AD conversion.

**Feature Relevance and Analysis** This work analyzed 316 features and their role in MCI to AD conversion risk. The RHOCV reported that 301 out of the 316 characteristics may have some association, but the detailed analysis indicated that only ten features were selected at least 50% of the time. Many of these ten features have already been reported as potential biomarkers associated with MCI to AD conversion [23, 30, 64, 28]. APOE4, a factor that has been validated several times as a biomarker indicative of the risk of conversion [23] was an important validation in our work. qMRI related features included the decreased volume of the cortical parcellation of the entorhinal, the increase in the white matter parcellation of the amygdala and increase volume and thickness standard deviation of Bankssts. These findings confirmed the results of previous studies [30, 28, 106]. Regarding novel features, our work suggests that large differences between left-right brain structures like the Pars Opercularis, Middle Temporal Lobule, and the Inferior Parietal Lobule, unlike the volumes listed for each condition and structure as mentioned in previous studies [45, 43, 125]. This work also aimed to improve the knowledge of the role of these features in the AD process; hence we reported the list of the top biomarkers along with their standardized HR associated with the conversion of MCI to AD. Reporting HR per z-units of the normal distribution may help physicians predict how far a specific patient in their MCI to AD process is.



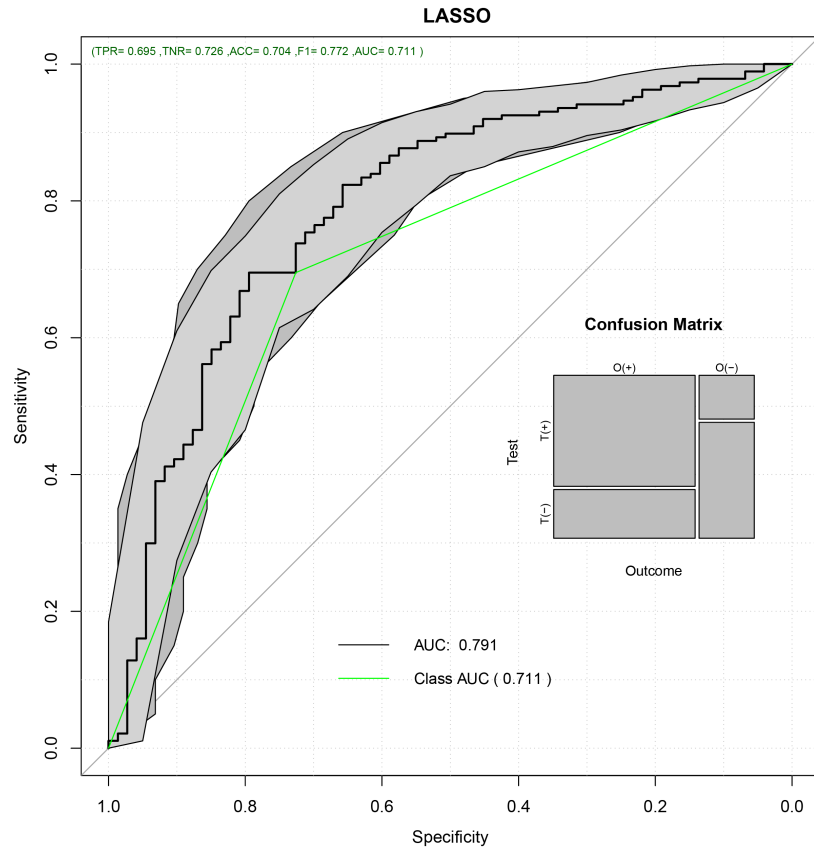


Figure 5.3: Coxnet ROC with 296 patients who suffered the conversion or have a censored event in more than 4 years.

**Limitations** The results presented in this work are limited in three key aspects. First, patient misdiagnosis is present, hence affecting feature selection and model building. The dementia diagnosis of “true” AD patients is not an exact science, hence detecting the exact time of conversion is also prone to diagnoses errors, and these two errors are present in modeled survival outcome. Second, it is based on the ADNI cohort and measurements; therefore, it is biased towards the environmental factors present in the US and the Caucasian race. Third, we assumed that all MCI will convert to AD in some point in the future. This assumption should not be a major issue if the proportion of misdiagnosed MCI is low. These key limitations indicate that the presented findings have to be confirmed on cohorts from different countries and ethnicities.

### 5.3.2 Prediction of MCI to AD Risk of Conversion Survival Models: qMRI vs CSF Measures and Cognitive Assessments

In this paper, we study seven different ML strategies that generated ten proportional risk models for each of the features groups and some combinations of them (Experiment I-VI). The

comparative evaluation of the CSF measures, the Cog evaluations and the structural characteristics of qMRI allowed us to study the conversion rate for patients with MCI who converted to AD or remained as MCI and evaluate the predictive power of the characteristics of qMRI against the other groups. The first 6 strategies (BSWiMS, LASSO, RIDGE, GPDAS, SPDAS and SPADS.BIC) returned a Cox regression model and the set of features that were required to make an accurate estimate of the conversion risk for each of the experiments. The following three strategies had a filter approach; Thus, the characteristics selected by the three default methods explored (BSWiMS, LASSO and GPDAS) were used to construct a standard Cox regression model. The last strategy evaluates the variables by using a univariate Cox regression model for characteristics that exceed a given threshold. The performance of 10 proportional risk models for each experiment was evaluated using RHOCV and the most common characteristics were analyzed in the model that combined all the feature groups to report their importance in the MCI to AD conversion rate.

**ML Method Validation** The RHOCV evaluation created 20 random splits of the dataset into training and test set. For each such split, the train fraction was 0.7 and the remaining 0.3 was used for testing. This evaluation strategy allowed the evaluation of the effect of the training set on feature selection, and, at the same time, permitted a training-set unbiased evaluation of the test performance. The reported results indicated that ML methods selected models with very different internal features. Model sizes varied considering the group of features used and from method to method. They ranged from a minimum of 3 features to complex multivariate modeling based on 294 features. The c-index on Follow-up times range depends on the type of features that were used in the model, but considering all the models it ranges from 0.59(0.56,0.62) for GPDAS on experiment IV to 0.93(0.91,0.94) for RIDGE on experiment V. In the Experiment VI (all group of features) the most complex model RIDGE has a very good c-index Follow up times 0.91(0.89,0.92) but its c-index on risks was not as good 0.63(0.61,0.66).

Regarding this classification performance, it is important to note that proportional hazard models were not designed for classifications task. To address this issue, we assumed that subjects predicted to have an increased risk of conversion ( $\text{Risk} > 1$ ) should correspond to true MCI to AD conversion, while subjects at low-risk prediction ( $\text{Risk} \leq 1$ ) should correspond to MCI-stable subjects. This strategy allowed us to evaluate the accuracy, sensitivity, specificity, and AUC of the Cox regression models. The reported AUC performance for their potential to detect patients at risk of conversion overall the models ranged from 0.69 for GPDAS on experiment III to 0.72 for BSWiMS on experiment IV. This performance was slightly lower to other methods based on SVM or Logistic Regression classifiers [42]. To test the impact of using all subjects in ROC AUC analysis, we conducted a post hoc experiment. In this experiment, we analyzed test prediction on MCI stable subjects whose last visit was greater than 3 years (52 no-event subjects did not meet the criteria). The experiment was conducted with BSWiMS using the dataset of Experiment VI. This change in selection criteria resulted in the ROC curve presented by Figure 5.4. We clearly see that Cox based conversion risk prediction had a similar performance (ROC ACU= 0.896) to previous works [70].

Regarding the reproducibility of feature selection, the Jaccard analysis indicated that the internal structure of the Cox models depended on the training set and the set of features used on each experiment. In the experiment VI, the method with the largest Jaccard index

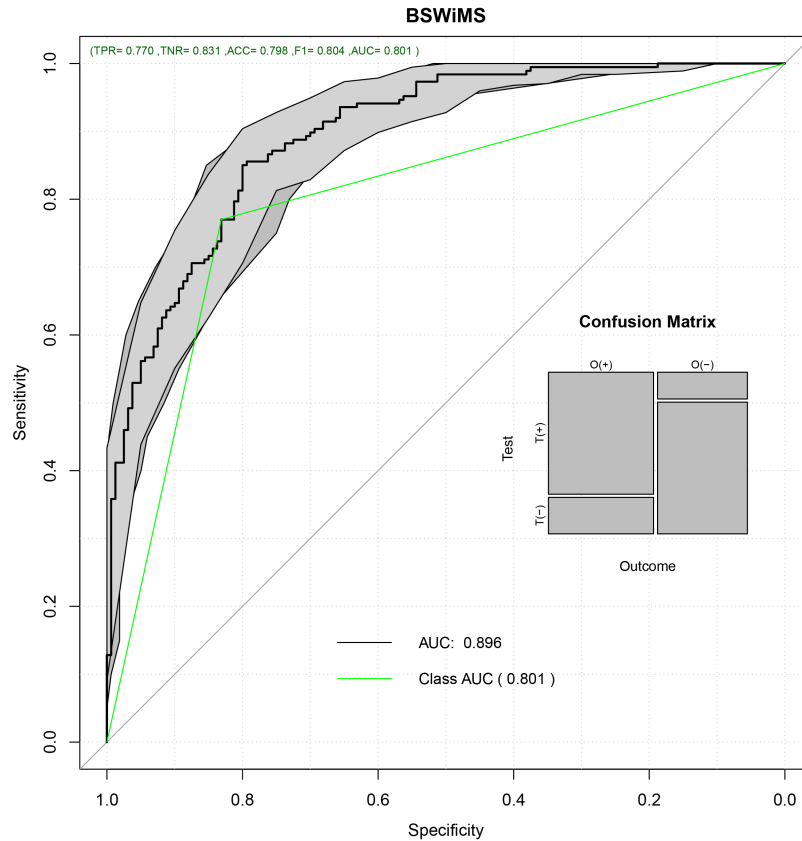


Figure 5.4: BSWiMS ROC with 347 patients who suffered the conversion or have a censored event in more than 3 years.

(0.86) was based on RIDGE selection, which also was the method with the largest set of features but with the best performance on concordance of follow-up times. The smallest models among all the experiments were returned by the SPDAS strategy. Its Jaccard Index on the same experiment was 0.51, implying that only 51% of the features overlapped across different training sets. These results put forward that the discovery of risk factors associated with MCI to AD conversion depended on the training set and the machine learning strategy used to discover risk factors. This observation is supported by the literature, where different authors have reported a different set of features associated with MCI to AD conversion.

**Feature relevance and analysis** This work studied three different types of variables -CSF measures, Cognitive assessments, Radiomics Features - used for the generation of survival models for the conversion rate of patients from MCI to AD. Ten different strategies were used over 6 different experiments, the reported stats allow discussing the importance and influence of the features in conversion. The detailed analysis on features selected in the Experiment VI indicated that only eight features out of 322, were selected at least 50% of the time. The correlation of those features are shown in the figure 5.5. Almost all of these eight features have already been reported as potential biomarkers associated with MCI to AD conversion

[118, 85, 87, 81]. qMRI related features included the decreased volume of the cortical parcel-lation of the entorhinal and temporal, and increase volume and thickness standard deviation of Bankssts. These findings confirmed the results of previous studies [30, 28, 106].

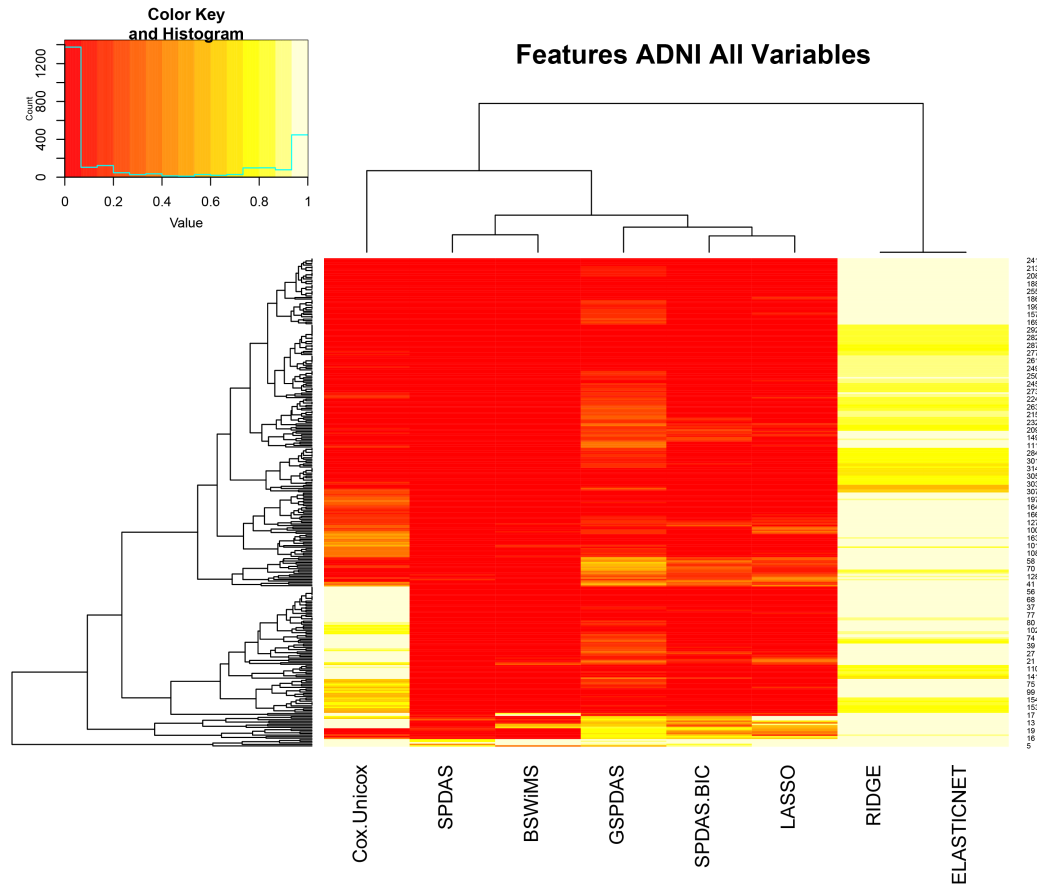


Figure 5.5: Heatmap of the features used in the Experiment VI.

CSF Measures have already been reported as a risk factor for the MCI to AD conversion. Accordingly, they had a great performance on c-index Follow-up ranging from 0.72 to 0.76. Moreover, the specificity of the model was low  $SPE = (0.53, 0.57)$ . Cog-assessments as well have already be used as a good technique to classify the conversion of the patients; Thus,  $ACC = (0.74-0.76)$   $AUC = (0.8-0.81)$ . Past work reported the role of Radiomics features on survival models [78]. Updates on that experiment result in c-index Follow-up times ranging from 0.72 to 0.92 but once again its role on classification is not as good  $ACC = (0.63-0.68)$ .

The combination of the set of features on Experiments IV, V & VI results in some better performances over all the stats but with non-significant statistical differences between them. In the context of follow-up times, the best method over these three experiments was RIDGE. Results on the experiments were: Experiment IV c-index FT = 0.92 (0.91, 0.93), Experiment V c-index FT = 0.93 (0.91, 0.94), and Experiment VI c-index FT = 0.91 (0.89, 0.92). Therefore, adding information about CSF Measures or Cog-assessments to the model of Radiomics features do not add significant information. This work also aimed to improve

the knowledge of the role of these features in the AD process; hence we reported the list of the top biomarkers along with their standardized HR associated with the conversion of MCI to AD. Reporting HR per z-units of the normal distribution may help physicians predict how far a specific patient in their MCI to AD process is.

V	FT	MT	Event Mean (SD)	No event Mean (SD)	MV HR (95% CI)	UV HR (95% CI)	M1	M2	M3	M4	M5	M6	M7
1	C	S	1.85 (0.92)	1.25 (0.69)	1.40*** (1.20,1.60)	1.8**** (1.6,2)	3	4	4	2	3	2	4
2	C	S	28.71 (8.04)	36.22 (10.39)	0.61*** (0.49,0.77)	0.45**** (0.37,0.54)	2	2	2	1	5	1	2
3	C	S	20.68 (6.25)	14.44 (5.66)	1.60*** (1.30,1.90)	2.70**** (2.30,3.20)	1	1	1	3	2	6	1
4	C	S	5.01 (4.65)	2.21 (3.3)	1.20* (1.00,1.40)	1.60**** (1.40,1.80)	5	3	3	7	6	5	3
5	M	V	1535.31 (423.91)	1783.72 (426.17)	0.79* (0.63,0.99)	0.47**** (0.39,0.57)	6	5	5	8	7	8	5
6	P	P	766.81 (715.52)	598.21 (666.43)	0.54*** (0.43,0.68)	0.44**** (0.35,0.55)	4	6	6	4	1	3	6
7	M	V	8896.89 (1803.1)	9681.64 (1617.63)	0.63*** (0.50,0.79)	0.49**** (0.40,0.61)	8	7	7	6	4	4	7
8	M	CT	0.54 (0.08)	0.51 (0.08)	1.40* (1.10,1.90)	1.70**** (1.40,2.2)	7	8	8	5	8	7	8

Table 5.4: Characteristics and ranking of eight features selected in almost the half of the iterations. The ranking was ordered based on the number of times selected, then ordered depending on the p-value of univariate cox analysis and finally, the concordance index of the univariate model. [FT = feature type; M=mean, C=Cog. Assessment, P=CSF Measure], [mt = measure type; v=volume (mm<sup>3</sup>), p = protein, ct = cortical thickness (mm)], [M1 = BSWiMS, M2 = LASSO, M3 = RIDGE, M4 = GPDAS, M5=SPDAS, M6=SPDAS.BIC, M7=Univariate Cox] P. Value significance: < 0.1, \* < 0.05, \*\* < 0.01, \*\*\* < 0.001, \*\*\*\* < 10<sup>-04</sup>

**Limitations** The results presented in this work are limited in three key aspects. First, patient misdiagnosis is present, hence affecting feature selection and model building. The diagnosis of AD is not definitive, but according to the NINCDS-ADRDA it is a probable diagnosis of AD. The error rate for this diagnosis occurs about 10% to 15% of cases [108]. Hence detecting the exact time of conversion is also prone to diagnoses errors, and these two errors are present in modeled survival outcome. Second, it is based on the ADNI cohort and measurements; therefore, it is biased towards the environmental factors present in the US and the Caucasian race. Third, we assumed that all MCI will convert to AD in some point in the future. This assumption should not be a major issue if the proportion of misdiagnosed MCI is low. These key limitations indicate that the presented findings have to be confirmed on cohorts from different countries and ethnicities.

**Conclusion** Radiomics biomarkers in the form of quantitative MRI assessments were an important source of features in the prediction of MCI to AD conversion time in models that only contained ApoE4, Cognitive Assessments and qMRI. Adding CSF biomarkers did not improve the accuracy nor the concordance of multi-source Survival Models.

## 5.4 Osteoarthritis Initiative: OAI

In this experiment, we examine 7 different ML strategies that generated 11 proportional risk models for the X-Rays features and OA Scores. We try to find the relationship between these characteristics with the total knee replacement using the information of both knees, regardless of which one is the event. The first 7 strategies (BSWiMS, LASSO, RIDGE, GPDAS, SP DAS, and SP DAS.BIC) returned a Cox regression model and the set of features that were required to make an accurate estimate of TKR on patients. The following three strategies had a filter approach; Thus, the characteristics selected by the three default methods explored (BSWiMS, LASSO, and GPDAS) were used to construct a standard Cox regression model. The last strategy evaluates the variables by using a univariate Cox regression model for characteristics that exceed a given threshold. The performance of 20 proportional risk models for each experiment was evaluated using RHOCV and the most common characteristics were analyzed in the model that combined all the feature groups to report their relation to the TKR event.

**Feature Relevance and Analysis** A total of 15 features were selected in 80 iterations out of 160 iterations and 236 features of this experiment. This features were used to compute a Cox Model and find their association to TKR Event. The most selected variable Kellgren and Lawrence grades are used right now in the OA screening from a lot of time ago [61]. Both Kellgren and Lawrence grades got a Hazard ratio bigger than one which means that when the difference of the grades between knees is bigger, the risk is also bigger. Besides, the mean of the grades is also directly proportional to the risk. The scores of KOOS were validated for the knee injury [94] and the relationship with TKR was studied [39]. In our model, both score have the HR less than one, which means that the risk is lower when the score is higher. Some studies conclude that the participation in sports had a increased prevalence rate of OA [29]. Others say that the sports was not associated with the risk of developing an incident with OA [86]. In our model, the biggest HR was found on the x position of the minimum JSW measure, but its confidence interval goes from 0.21 to 27.38 which is not a good indicator. The variability of this feature occurred because of its nature. We think that the selection of this variable has to do with the value of the minimum JSW.

**Benchmarking** We left the clinical relevance of the features for future works and we are going to take more importance into the discussion of the benchmarking process. In the figure 5.6 we show the result of the plot process of the CoxBenchmarking model for the OAI analysis. All the plots show a comparison between the stats of all the wrapper methods. We chose to compare just the wrapper methods because only those methods selected features but we will mention all the filter methods if in that stat its behavior is better. In case of Accuracy, all the models got statistically equivalent results. The best result was reported by RIDGE, ELASTIC-NET and later BSWiMS. These three methods are the only ones that show a slightly superior

behavior compared to GSPDAS and SPDAS. However, the lower tail of these three methods does run into the upper tail of the other two methods. Regarding AUC and SEN the result is the same, but the difference between the best stat and the worst is Negligible. The main difference between AUC and SEN is the order of the models. The best AUC was found by the same method than ACC (RIDGE), but the best SEN was found in the SPDAS algorithm. On the other hand, the SPE reported the same behavior than the ACC stat; the best SPE was also found in RIDGE and its difference in mean between RIDGE SPE and the the worst SPE in SPDAS is 0.9.

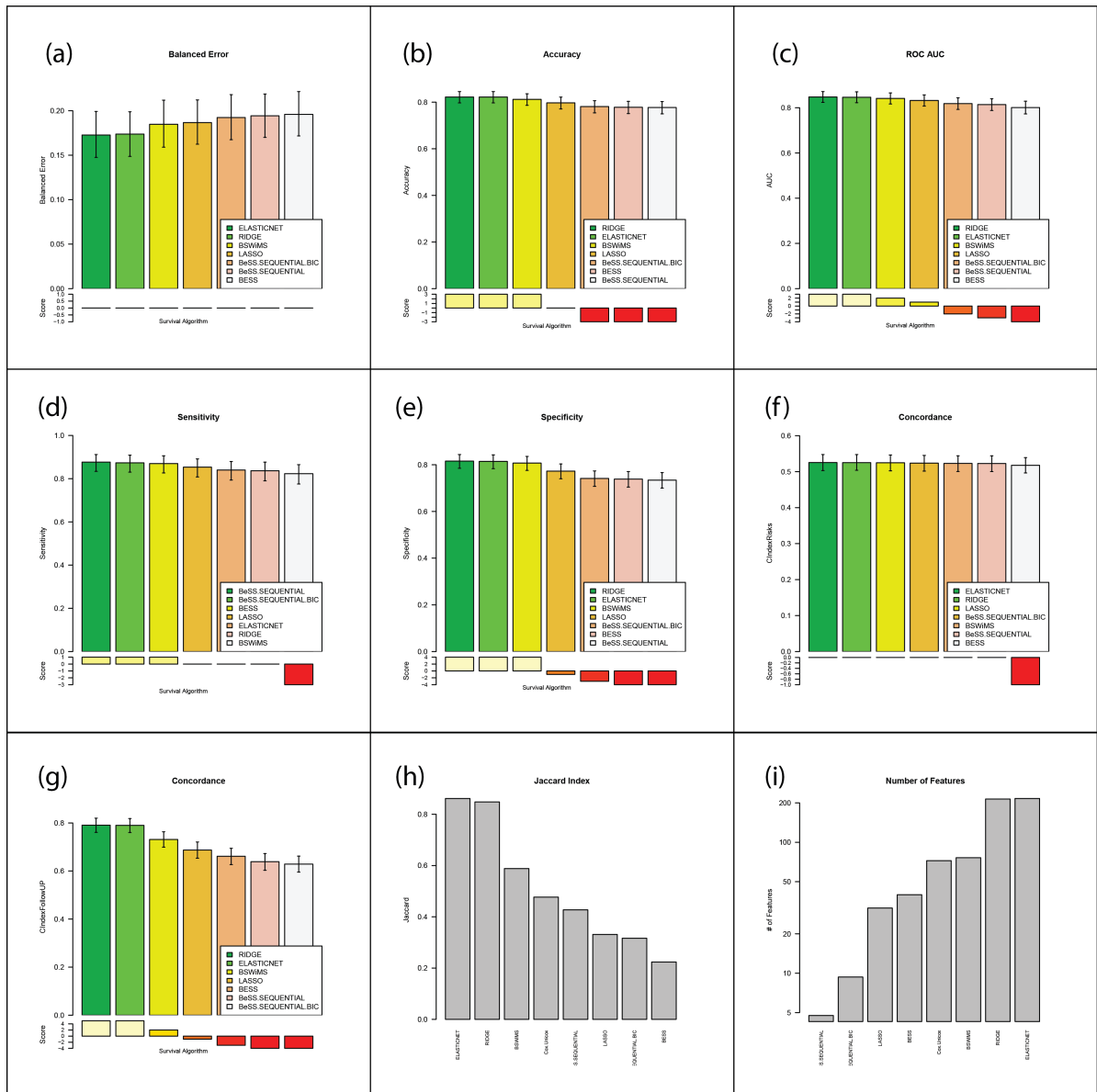


Figure 5.6: Plot of CoxBenchmarking analysis of OAI data. Barplots for (a) Balanced Error (b) Accuracy (c) ROC.AUC (d) Sensitivity (e) Specificity (f) C-Index Risks (g) C-Index Follow-Up Times (h) Jaccard Index (i) Mean of features

Regarding C-Index Risks the models show almost the same stat but this time, the filter methods got a much better result than wrappers. With the characteristics selected with BSWiMS, a C-Index was found that was far superior to the Wrappers models with 0.85 (0.82, 0.88). The model that follows is developed with LASSO whose upper tail touches the lower tail of BSWiMS. The filter with GSPDAS reduces its statistics however it is still better than its result with the wrapper method. Cox's univariate analysis also had a higher result than the models generated by wrappers. In contrast, C-Index Follow-up Times shows the most uneven statistic of all. RIDGE and ELASTICNET once again got the best C-Index FT with 0.79 (0.76, 0.82). The next model is BSWiMS whose interval is smaller than RIDGE and ELASTICNET but overlaps with LASSO which is the next statistic in order. The methods belonging to the BeSS package had similar statistics but were inferential to BSWiMS, ELASTICNET, and RIDGE, they only reached the lower tail of the LASSO result.

The classification and survival statistics in general terms show a similar performance. The main difference shown by the models is the quantity and variability of features that are used to build the Cox model. Variability is measured with the Jaccard Index and within the built models, we find that RIDGE and ELASTICNET have a very high index (close to one) which means that the models are almost the same in each iteration. On the other hand, GSPDAS has a very low index compared to the RIDGE (Jaccard Index of 0.2), which suggests that BeSS selected different models in each iteration. This statistic is better explained and interpreted by knowing the number of mean variables used by each model. In the case of RIDGE and ELASTICNET that chose more than 200 variables, they have a very high Jaccard index because they select the majority of available features. This happens because these techniques always leave all the characteristics that do not have an extreme correlation, but later they penalize the characteristics through the coefficients. In contrast, SPDAS with BIC selected less than 5 features per model but tested with different combinations in each iteration so its Jaccard Index is close to 0.4. BSWiMS and Unicox selected several quite similar characteristics, LASSO and GSPDAS found variables between 30 and 50 characteristics. The difference between SPDAS and the other methods that select more than 20 characteristics is noted.

**Limitations** This paper presents its main limitation in the lack of clinical knowledge about OA. Although the results are quite promising, we cannot infer or discuss clinical problems without having the knowledge or support of someone trained in the field. In this work, we limit ourselves to informing and comparing the statistics shown by the survival models generated. On the other hand, the order of the data and the form of distribution presented by the OAI generate a small inconvenience when carrying out the analysis. This experiment is the one that required more time for its preparation before the analysis with this tool. This limited the amount of information that can be added to use in the models. Finally, the limitation we knew to overcome is the number of events available. If we compare the number of events with the amount of data, we only had 10% of the events.



## 5.5 Prognostic Wisconsin Breast Cancer Database

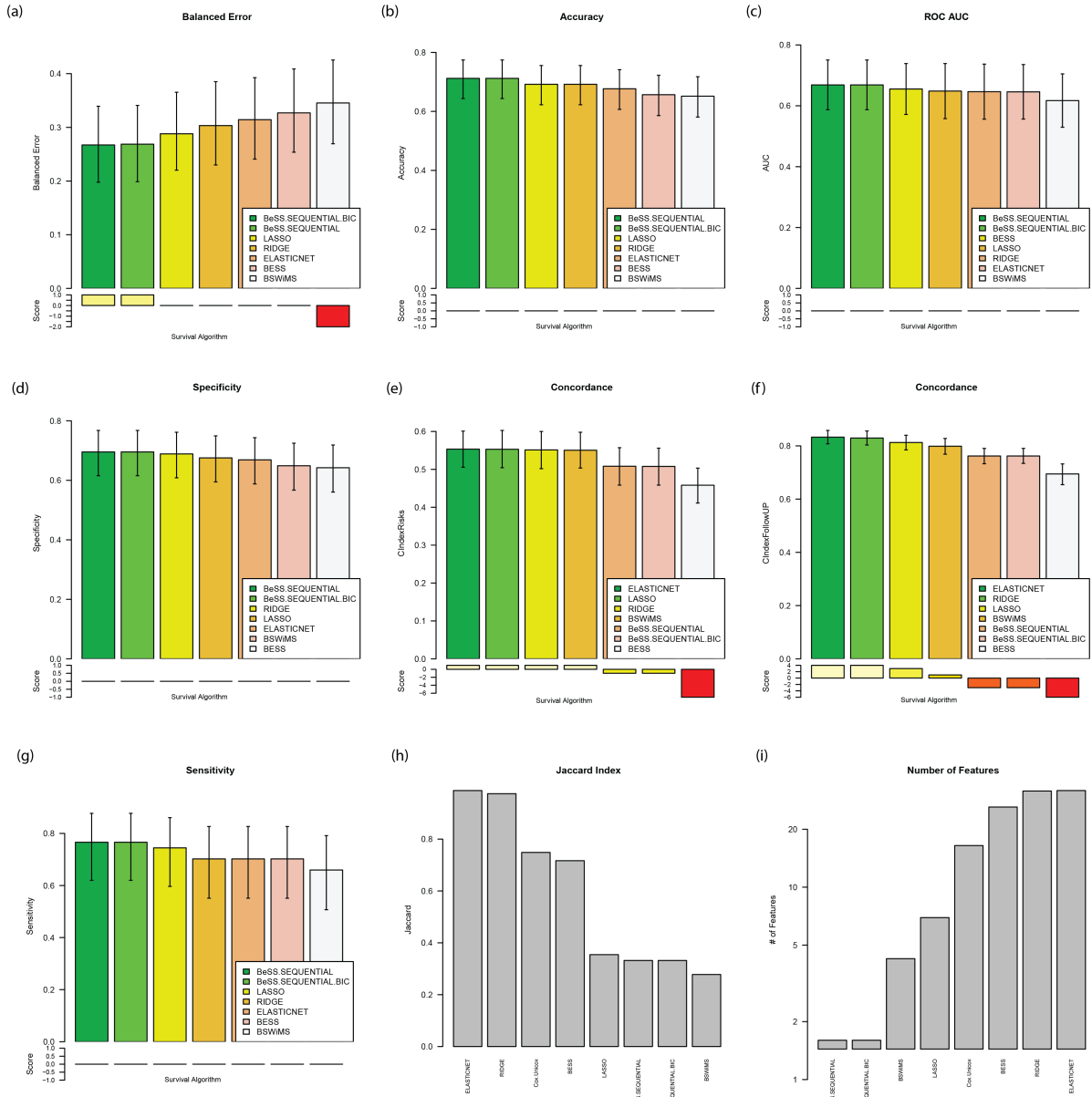


Figure 5.7: Plot of CoxBenchmarking analysis of BRCA Wisconsin data. Barplots for (a) Balanced Error (b) Accuracy (c) ROC.AUC (d) Specificity (e) C-Index Risks (f) C-Index Follow-Up Times (g) Sensitivity (h) Jaccard Index (i) Mean of features

This paper tries to estimate the recurrence time of breast cancer patients belonging to the prognosis database of the University of Wisconsin. In this experiment, the lowest statistics of all the models developed in this thesis were reported. Even so, this was expected due to the nature of the data and by previous studies that state that the Cox model does not work with this prediction [120]. On the other hand, it is important to take into account that the data was used

raw (like they were publicly provided), which may affect the ability of the methods to find a result. However, we take into account that the main objective of this thesis is the evaluation of these methods for the survival analysis, so because the results between the methods (whose final objective is to do the same analysis) were not so different and a comparison can be made, then we analyze each of the statistics.

**Benchmarking** All statistics presented statistically equivalent values. All ranking statistics have their confidence intervals intertwined and the differences between their means are negligible. In the context of survival statistics, if a comparison of the methods can be made. In the case of C-Index Risks, GSPDAS was the model that had the lowest concordance value. The highest value in its tail does not reach the lowest value of ELASTICNET, LASSO, RIDGE, and BSWiMS. However, its highest value is the statistical average of the other algorithms that belong to the BeSS package. In the case of C-Index Follow-up Times SPDAS, SPDAS with BIC and GSPDAS, in that order, were worse than the concordance found by the models created by the algorithms that are part of GLMNET. BSWiMS is a bit inferior to these models, however, its upper tail is intertwined with the tails of the other methods. The best matching model was ELASTICNET. Figure 5.7 shows the plot representation of the stats of this analysis.

## 5.6 Prognostic San Jose Breast Cancer Database

The objective of this analysis is to find characteristics that are part of the mammogram to be able to predict the recurrence time of the patients of the San Jose Hospital. The main problem of this discussion will be seen from the fact that only one model managed to reach a result, all the others did not find a solution. Next, we will discuss the importance of the results shown by this model. In this case, due to the nature of the data we only used a RHOCV that created 20 random splits of the dataset into training and test set. For each such split, the train fraction was 0.6 and the remaining 0.4 was used for testing.

**Feature Relevance and Analysis** BSWiMS selected 160 features out of 1091 features. However, of all those selected, only three were used in at least half of the characteristics. We will use these features later as the top features. On the other hand, of the remaining 157 characteristics, only 11 characteristics were used in more than 20% of the iterations and less than 10 iterations; that is, 146 features were used in almost no model. If we take into account that the number of variables analyzed was 1089, we are using approximately 15% of the characteristics for the models. The fact that 146 features are used in a few models, makes the Jaccard index much lower than what BSWiMS had accustomed us in the other experiments. BSWiMS got an Sensitivity of 0.76 which can be compared to the validation score of Oncotype REF.

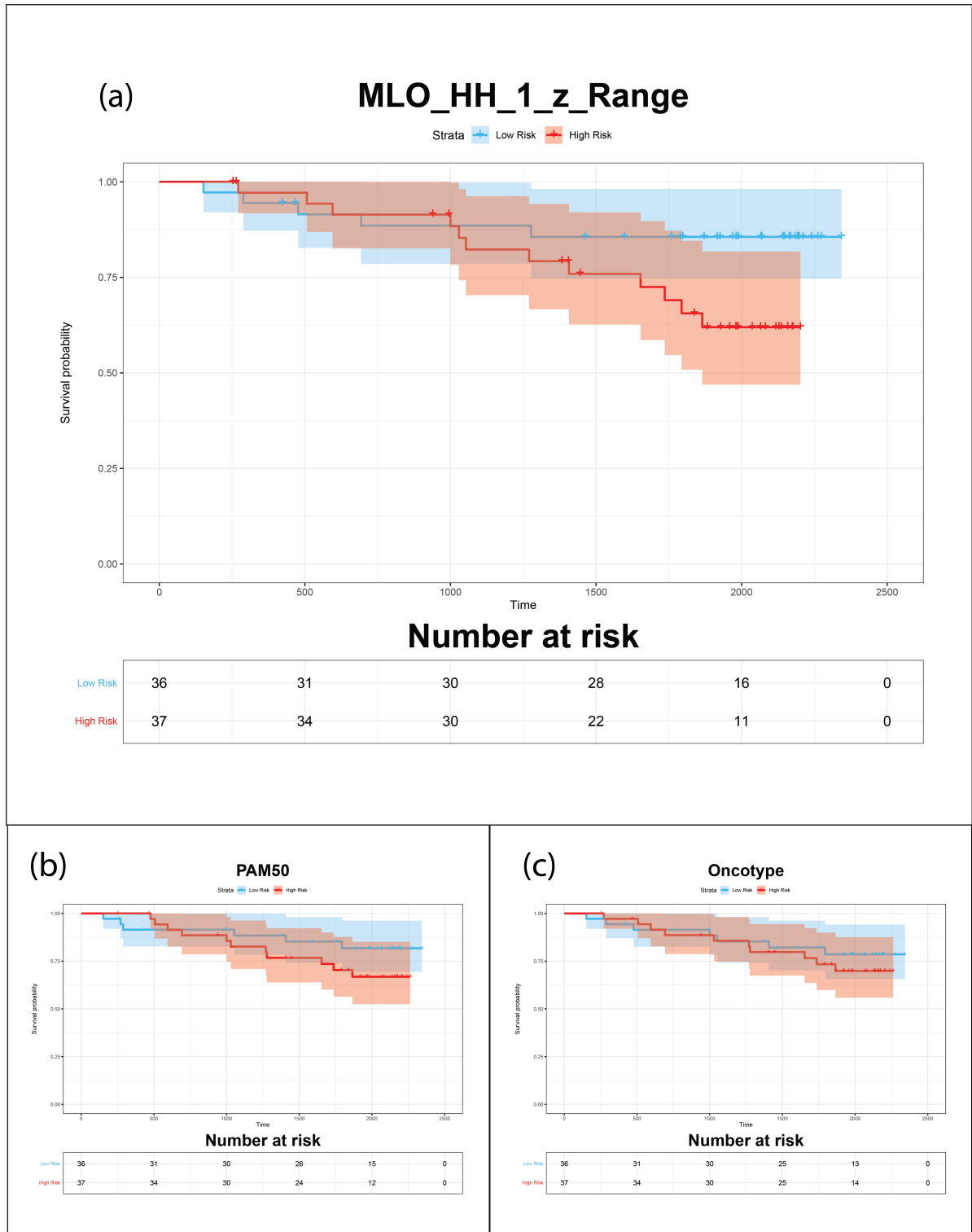


Figure 5.8: KM Curves stratified by the median of (a) HH1 Level z-Score  $p=0.063$  (b) PAM50 Score  $p=0.22$ , (c) Oncotype  $p=0.50$

Regarding the top features, we used them to predict the recurrence time of BRCA with a separate Cox Model. We relate the recurrence time to the z-range of HH1 and LH1 level and

Dynamic range of the level HH. With those features, we got a concordance index of 0.74 (se = 0.068) and the Log-rank test p-value  $p = 6e-05$ . The features result in positive HR which means more risk when the value is greater. HH1 z-range finished with HR = 1.056 (0.65, 1.71), LH1 z-range HR = 2.00 (1.04, 3.85) and Dynamic Range HR 1.004 (0.996, 1.009). Those HRs are not that important if we consider the magnitude of the value. Henceforth, to give better information about the features we create a KM Curve for the most selected variable by arranging the patients with the median of the value of that characteristic. The figure 5.8(a) shows the KM Curve generated by stratifying the data by HH1 level z-range median. The log-rank test p-value for this KM Curve is:  $p = 0.063$ . In contrast, PAM50  $p = 0.22$  and Oncotype  $p = 0.50$  KM curves shown in the figure 5.8(b)(c) despite being validated scores, they do not have such a good survival curve with these data.

**Limitations** The main limitation of this data set is the number of events we have and also the total data. However, being an investigation that belongs to the institution, it is known that more information will be obtained. For this reason, the information obtained through this research is very important for the continued development of this base. Second, the most considerable limitation is caused by the number of variables that have to be analyzed. With this amount of features and taking into account the feature selection strategies; None of the models could find a solution for the model. The number of features limited the comparison and therefore use of the CoxBenchmarking tool.

## 5.7 Conclusions

After the analysis of the results on experiments and their comparison to the ground truth and previous studies. We can assure that the machine learning algorithms selected in this thesis allow the construction of reliable survival Cox models that allowed study survival times for subject who suffers an event. In this case, especially in the application of those ML techniques on chronic degenerative diseases; on which, through the selection of clinical features such as mammograms, X-rays, MRI and PET, forms and clinical data about the patient, can partly explain the outcome. Even though some algorithms did have better performances in some circumstances and statistics, we conclude that any of them can be used in the medical context and discover relevant information in that context. The overall discrepancy between their performance can be taken as an advantage that permits get deeper information about the event. Hence, the CoxBenchmarking tool provides a strong method that provides valuable information for the survival study, making use of techniques that are already helpful by themselves and adding its benchmarking let us explore different points of view.

## 5.8 Future work

This thesis managed to answer some questions but also founded the basis for remaining high importance research. Regarding the data simulation case, the performance of CoxBenchmarking was measured with high noise datasets. The next work should concentrate on modifying the technique to help the methods somehow find a solution for these sets. Changes in the

R package implementation may be required. In other hand, the importance of the random variables was not fully studied. Those studies require some experiments to measure how the random variables change the models. Regarding the number of methods analyzed, it is still pending to analyze more techniques and implement ways in which the user can select which methods are part of the comparative analysis. In the context of ADNI results, which is the most developed experiment within this thesis, it is necessary to consider the exploration of the reported features in a clinical context. Two scientific works are already made with this experiment and more work is needed to have more clinical importance on the results. Exploration of those features considering the limitations of the population can be done on future experimentation. Regarding the BRCA Wisconsin dataset, despite literature stating that Cox could not lead to good results in this dataset, the data preparation can be different to explore something different. May be Cox with the feature selection process works better with this data if the derived information is more informative. Regarding San Jose BRCA dataset, the exploration of this survival information was not expected to have any results. In fact, the small number of cases and the vast number of features lead to the failure of almost all models. Furthermore, BSWiMS got promising results on that dataset. With the acquisition of more patient data, the results could be better and have more scientific impact. Finally regarding OAI, this experiment is the first exploration of survival analysis with Cox in the TKR event. A detailed scientific work can be performed and the features reported can be clinically analyzed. As we can note, the implementation of CoxBenchmarking resulted in invaluable information. Further, there is a lot of work to do, to socialize, check and test its use in many computer-aided diagnosis fields.



# Bibliography

- [1] AGUIRRE-GAMBOA, R., MARTINEZ-LEDESMA, E., GOMEZ-RUEDA, H., PALACIOS, R., FUENTES-HERNANDEZ, I., SÁNCHEZ-CANALES, E., CHACOLLA-HUARINGA, R., CARDONA-HUERTA, S., VILLELA, L., SCOTT, S.-P., TAMEZ-PENA, J., AND TREVINO, V. Efficient Gene Selection for Cancer Prognostic Biomarkers Using Swarm Optimization and Survival Analysis. *Current Bioinformatics* 11, 3 (jun 2016), 310–323.
- [2] AHMAD, L., AND ESHLAGHY, A. Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence. *Journal of Health & Medical Informatics* (2013).
- [3] AKAIKE, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19, 6 (dec 1974), 716–723.
- [4] AL-SHOMRANI, A. A., SHAWKY, A. I., ARIF, O. H., AND ASLAM, M. Log-logistic distribution for survival data analysis using MCMC. *SpringerPlus* 5, 1 (2016), 1774.
- [5] ALTMAN, D. G., AND BLAND, J. M. Diagnostic tests. 1: Sensitivity and specificity. *BMJ (Clinical research ed.)* 308, 6943 (jun 1994), 1552.
- [6] ALZHEIMER’S ASSOCIATION. 2018 Alzheimer’s disease facts and figures. *Alzheimer’s & Dementia* 14, 3 (mar 2018), 367–429.
- [7] ALZHEIMER’S DISEASE INTERNATIONAL. World Alzheimer Report 2018 - The state of the art of dementia research: New frontiers. Tech. rep.
- [8] ARBIZU, J., PRIETO, E., MARTÍNEZ-LAGE, P., MARTÍ-CLIMENT, J. M., GARCÍA-GRANERO, M., LAMET, I., PASTOR, P., RIVEROL, M., GÓMEZ-ISLA, M. T., PEÑUELAS, I., RICHTER, J. A., WEINER, M. W., AND INITIATIVE, F. T. A. D. N. Automated analysis of FDG PET as a tool for single-subject probabilistic prediction and detection of Alzheimer’s disease dementia. *European Journal of Nuclear Medicine and Molecular Imaging* 40, 9 (sep 2013), 1394–1405.
- [9] ARLOT, S., AND CELISSE, A. A survey of cross-validation procedures for model selection. *Statistics Surveys* 4, 0 (2010), 40–79.
- [10] ASHINSKY, B., BOUHRARA, M., COLETTA, C., LEHALLIER, B., URISH, K., LIN, P., GOLDBERG, I., AND SPENCER, R. Predicting early symptomatic osteoarthritis in the human knee using machine learning classification of magnetic resonance images

- from the osteoarthritis initiative. *Journal of Orthopaedic Research* 35, 10 (oct 2017), 2243–2250.
- [11] BAGDONAVIČIUS, V., LEVULIENÉ, R., AND NIKULIN, M. Goodness-of-fit criteria for the Cox model from left truncated and right censored data. *Journal of Mathematical Sciences* 167, 4 (jun 2010), 436–443.
- [12] BELLAMY, N., BUCHANAN, W. W., GOLDSMITH, C. H., CAMPBELL, J., AND STITT, L. W. Validation study of WOMAC: A health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. *Journal of Rheumatology* 15, 12 (1988), 1833–1840.
- [13] BENJAMINI, Y., AND HOCHBERG, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, 1995.
- [14] BENNETT, K. P., DEMIRIZ, A., AND MACLIN, R. Exploiting Unlabeled Data in Ensemble Methods. Tech. rep., 2002.
- [15] BICHINDARITZ, I., ENGLEBERT, C., REGUA, A., AND KOTULA, L. Feature Selection and Case-Based Reasoning for Survival Analysis in Bioinformatics. *The Thirty-First International Flairs Conference* (may 2018).
- [16] BIPM. International vocabulary of metrology-Basic and general concepts and associated terms (VIM) 3rd edition 2008 version with minor corrections Vocabulaire international de métrologie-Concepts fondamentaux et généraux et termes associés (VIM) 3 e édition. Tech. rep., 2012.
- [17] BOYKO, E. J. Ruling Out or Ruling In Disease with the Most sensitive or Specific Diagnostic Test. *Medical Decision Making* 14, 2 (apr 1994), 175–179.
- [18] BRADBURN, M., CLARK, T., LOVE, S., AND ALTMAN, D. Survival analysis part II: multivariate data analysis—an introduction to concepts and methods. *British journal of cancer* 89, 3 (aug 2003), 431–6.
- [19] BROWNE, M. W., AND CUDECK, R. Single Sample Cross-Validation Indices for Covariance Structures. *Multivariate Behavioral Research* (1989).
- [20] CANCER INSTITUTE, N. Mammograms - National Cancer Institute, 2016.
- [21] CHÁVARRI-GUERRA, Y., VILLARREAL-GARZA, C., LIEDKE, P., KNAUL, F., MOHAR, A., FINKELSTEIN, D., AND GOSS, P. Breast cancer in Mexico: a growing challenge to health and the health system. *The Lancet Oncology* 13, 8 (aug 2012), e335–e343.
- [22] CHEN, J., AND CHEN, Z. Extended Bayesian Information Criteria for Model Selection with Large Model Spaces, 2008.



- [23] CORDER, E. H., SAUNDERS, A. M., STRITTMATTER, W. J., SCHMECHEL, D. E., GASKELL, P. C., SMALL, G. W., ROSES, A. D., HAINES, J. L., AND PERICAK-VANCE, M. A. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* (1993).
- [24] COX, D. R. Regression Models and Life-Tables, 1972.
- [25] CRUZ, J., AND WISHART, D. Applications of machine learning in cancer prediction and prognosis. *Cancer informatics 2* (feb 2007), 59–77.
- [26] DALGAARD, P. Chapter 14: Survival analysis. In *Introductory Statistics with R*. 2008.
- [27] DE FLORA, S., IZZOTTI, A., RANDEPATH, K., RANDEPATH, E., BARTSCH, H., NAIR, J., BALANSKY, R., VAN SCHOOTEN, F., DEGAN, P., FRONZA, G., WALSH, D., AND LEWTAS, J. DNA adducts and chronic degenerative diseases. Pathogenetic relevance and implications in preventive medicine. *Mutation Research/Reviews in Genetic Toxicology 366*, 3 (dec 1996), 197–238.
- [28] DETOLEDO-MORRELL, L., STOUB, T., BULGAKOVA, M., WILSON, R., BENNETT, D., LEURGANS, S., WUU, J., AND TURNER, D. MRI-derived entorhinal volume is a good predictor of conversion from MCI to AD. *Neurobiology of Aging 25*, 9 (oct 2004), 1197–1203.
- [29] DRIBAN, J. B., HOOTMAN, J. M., SITLER, M. R., HARRIS, K. P., AND CATTANO, N. M. Is participation in certain sports associated with knee osteoarthritis? A systematic review, jun 2017.
- [30] DU, A. T., SCHUFF, N., AMEND, D., LAAKSO, M. P., HSU, Y. Y., JAGUST, W. J., YAFFE, K., KRAMER, J. H., REED, B., NORMAN, D., CHUI, H. C., AND WEINER, M. W. Magnetic resonance imaging of the entorhinal cortex and hippocampus in mild cognitive impairment and Alzheimer's disease. *Journal of neurology, neurosurgery, and psychiatry 71*, 4 (oct 2001), 441–7.
- [31] DUBITZKY, W., GRANZOW, M., AND BERRAR, D. P. *Fundamentals of data mining in genomics and proteomics*. Springer, 2011.
- [32] DURYEY, J., LI, J., PETERFY, C. G., GORDON, C., AND GENANT, H. K. Trainable rule-based algorithm for the measurement of joint space width in digital radiographic images of the knee. *Medical physics 27*, 3 (mar 2000), 580–91.
- [33] EATON, S., CORDAIN, L., AND LINDBERG, S. Evolutionary Health Promotion: A Consideration of Common Counterarguments. *Preventive Medicine 34*, 2 (feb 2002), 119–123.
- [34] FAWCETT, T. An introduction to ROC analysis. *Pattern Recognition Letters* (2006).
- [35] FELSON, D. T., MCCLAUGHLIN, S., GOGGINS, J., LAVALLEY, M. P., GALE, M. E., TOTTERMAN, S., LI, W., HILL, C., AND GALE, D. Bone Marrow Edema and Its Relation to Progression of Knee Osteoarthritis. *Annals of Internal Medicine 139*, 5 I (sep 2003).

- [36] FLICKER, C., FERRIS, S. H., AND REISBERG, B. Mild cognitive impairment in the elderly: predictors of dementia. *Neurology* 41, 7 (jul 1991), 1006–9.
- [37] FOX, N. C., AND SCHOTT, J. M. Imaging cerebral atrophy: normal ageing to Alzheimer’s disease. *The Lancet* 363, 9406 (jan 2004), 392–394.
- [38] FRAGOSO-ONTIVEROS, V., VELÁZQUEZ-ARAGÓN, J. A., NUÑEZ-MARTÍNEZ, P. M., DE LA LUZ MEJÍA-AGUAYO, M., VIDAL-MILLÁN, S., PEDROZA-TORRES, A., SÁNCHEZ-CONTRERAS, Y., RAMÍREZ-OTERO, M. A., MUÑIZ-MENDOZA, R., DOMÍNGUEZ-ORTÍZ, J., WEGMAN-OSTROSKY, T., BARGALLÓ-ROCHA, J. E., GALLARDO-RINCÓN, D., REYNOSO-NOVERON, N., ARRIAGA-CANON, C., MENESES-GARCÍA, A., HERRERA-MONTALVO, L. A., AND ALVAREZ-GOMEZ, R. M. Mexican BRCA1 founder mutation: Shortening the gap in genetic assessment for hereditary breast and ovarian cancer patients. *PLOS ONE* 14, 9 (sep 2019), e0222709.
- [39] GANDEK, B., AND WARE, J. E. Validity and Responsiveness of the Knee Injury and Osteoarthritis Outcome Score: A Comparative Study Among Total Knee Replacement Patients. *Arthritis Care and Research* 69, 6 (jun 2017), 817–825.
- [40] GAUTHIER, S., REISBERG, B., ZAUDIG, M., PETERSEN, R. C., RITCHIE, K., BROICH, K., BELLEVILLE, S., BRODATY, H., BENNETT, D., CHERTKOW, H., CUMMINGS, J. L., DE LEON, M., FELDMAN, H., GANGULI, M., HAMPEL, H., SCHELTENS, P., TIERNEY, M. C., WHITEHOUSE, P., AND WINBLAD, B. Mild cognitive impairment. *The Lancet* 367, 9518 (apr 2006), 1262–1270.
- [41] GILLIES, R., KINAHAN, P., AND HRICAK, H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* 278, 2 (feb 2016), 563–577.
- [42] GÓMEZ-SANCHO, M., TOHKA, J., AND GÓMEZ-VERDEJO, V. Comparison of feature representations in MRI-based MCI-to-AD conversion prediction. *Magnetic Resonance Imaging* 50 (jul 2018), 84–95.
- [43] GREENE, S. J., KILLIANY, R. J., AND ALZHEIMER’S DISEASE NEUROIMAGING INITIATIVE, A. D. N. Subregions of the inferior parietal lobule are affected in the progression to Alzheimer’s disease. *Neurobiology of aging* 31, 8 (aug 2010), 1304–11.
- [44] HAMBURG, M., AND COLLINS, F. The Path to Personalized Medicine. *New England Journal of Medicine* 363, 4 (jul 2010), 301–304.
- [45] HÄNGGI, J., STREFFER, J., JÄNCKE, L., AND HOCK, C. Volumes of Lateral Temporal and Parietal Structures Distinguish Between Healthy Aging, Mild Cognitive Impairment, and Alzheimer’s Disease. *Journal of Alzheimer’s Disease* 26, 4 (oct 2011), 719–734.
- [46] HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. *Springer Series in Statistics The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. 2009.

- [47] HENNEMAN, W. J. P., SLUIMER, J. D., BARNES, J., VAN DER FLIER, W. M., SLUIMER, I. C., FOX, N. C., SCHELTENS, P., VRENKEN, H., AND BARKHOF, F. Hippocampal atrophy rates in Alzheimer disease: added value over whole brain volume measures. *Neurology* 72, 11 (mar 2009), 999–1007.
- [48] HOERL, A. E., AND KENNARD, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 42, 1 (feb 2000), 80–86.
- [49] HU, T., OKSANEN, K., ZHANG, W., RANDELL, E., FUREY, A., SUN, G., AND ZHAI, G. An evolutionary learning and network approach to identifying key metabolites for osteoarthritis. *PLoS Computational Biology* (2018).
- [50] ILIOU, T., AND ANAGNOSTOPOULOS, C. Osteoporosis Detection Using Machine Learning Techniques and Feature Selection. *International Journal on Artificial Intelligence Tools* (2014).
- [51] JACCARD, P., AND ZURICH, E. Article in Bulletin de la Societe Vaudoise des Sciences Naturelles.
- [52] JACK, C. R., SHIUNG, M. M., WEIGAND, S. D., O'BRIEN, P. C., GUNTER, J. L., BOEVE, B. F., KNOPMAN, D. S., SMITH, G. E., IVNIK, R. J., TANGALOS, E. G., PETERSEN, R. C., AND PETERSEN, R. C. Brain atrophy rates predict subsequent clinical conversion in normal elderly and amnesic MCI. *Neurology* 65, 8 (oct 2005), 1227–31.
- [53] KAPLAN, E., AND MEIER, P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* (1958).
- [54] KASSAMBARA, A. Package 'survminer', 2018.
- [55] KEYS, A., ARAVANIS, C., BLACKBURN, H., VAN BUCHEM, F. S., BUZINA, R., DJORDJEVIC, B. S., FIDANZA, F., KARVONEN, M. J., MENOTTI, A., PUDDU, V., AND TAYLOR, H. L. Coronary heart disease: overweight and obesity as risk factors. *Annals of internal medicine* 77, 1 (1972), 15–27.
- [56] KHAN, A., CORBETT, A., AND BALLARD, C. Emerging treatments for Alzheimer's disease for non-amyloid and non-tau targets. *Expert Review of Neurotherapeutics* 17, 7 (jul 2017), 683–695.
- [57] KLEINBAUM, D. Survival Analysis: A Self-Learning Text. *Biometrics* (1996).
- [58] KLEINBAUM, D., AND KLEIN, M. Kaplan-Meier Survival Curves and the Log-Rank Test. Springer, New York, NY, 2012, pp. 55–96.
- [59] KNOPMAN, D. S., AND PETERSEN, R. C. Mild cognitive impairment and mild dementia: a clinical perspective. *Mayo Clinic proceedings* 89, 10 (oct 2014), 1452–9.
- [60] KOHAVI, R., AND KOHAVI, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. 1137—1143.

- [61] KOHN, M. D., SASSOON, A. A., AND FERNANDO, N. D. Classifications in Brief: Kellgren-Lawrence Classification of Osteoarthritis. *Clinical Orthopaedics and Related Research* 474, 8 (aug 2016), 1886–1893.
- [62] KOUROU, K., EXARCHOS, T., EXARCHOS, K., KARAMOUZIS, M., AND FOTIADIS, D. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal* 13 (jan 2015), 8–17.
- [63] LIU, K., CHEN, K., YAO, L., AND GUO, X. Prediction of Mild Cognitive Impairment Conversion Using a Combination of Independent Component Analysis and the Cox Model. *Frontiers in Human Neuroscience* 11 (2017), 33.
- [64] LONG, X., JIANG, C., AND ZHANG, L. Morphological Biomarker Differentiating MCI Converters from Nonconverters: Longitudinal Evidence Based on Hemispheric Asymmetry. *Behavioural neurology* 2018 (2018), 3954101.
- [65] LYNCE, F., GRAVES, K. D., JANDORF, L., RICKER, C., CASTRO, E., MORENO, L., AUGUSTO, B., FEJERMAN, L., AND VADAPARAMPIL, S. T. Genomic disparities in breast cancer among latinas. *Cancer Control* 23, 4 (oct 2016), 359–372.
- [66] MACKILLOP, W. The Importance of Prognosis in Cancer Medicine. In *TNM Online*. John Wiley & Sons, Inc., Hoboken, NJ, USA, jul 2006.
- [67] MANGASARIAN, O., STREET, W., AND WOLBERG, W. Breast Cancer Diagnosis and Prognosis Via Linear Programming. *Operations Research* (1995).
- [68] MANTEL, N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer chemotherapy reports* 50, 3 (mar 1966), 163–70.
- [69] MARINESCU, R. V., OXTOBY, N. P., YOUNG, A. L., BRON, E. E., TOGA, A. W., WEINER, M. W., BARKHOF, F., FOX, N. C., KLEIN, S., ALEXANDER, D. C., CONSORTIUM, T. E., AND INITIATIVE, F. T. A. D. N. TADPOLE Challenge: Prediction of Longitudinal Evolution in Alzheimer’s Disease.
- [70] MARTÍNEZ-TORTEYA, A., TREVIÑO, V., AND TAMEZ-PEÑA, J. G. Improved Diagnostic Multimodal Biomarkers for Alzheimer’s Disease and Mild Cognitive Impairment. *BioMed Research International* 2015 (jan 2015), 1–11.
- [71] MCLACHLAN, G. J., DO, K.-A., AND AMBROISE, C. *Analyzing Microarray Gene Expression Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA, jul 2004.
- [72] MEREDITH, J. E., SANKARANARAYANAN, S., GUSS, V., LANZETTI, A. J., BERISHA, F., NEELY, R. J., SLEMMON, J. R., PORTELIUS, E., ZETTERBERG, H., BLENNOW, K., SOARES, H., AHLIJANIAN, M., AND ALBRIGHT, C. F. Characterization of Novel CSF Tau and ptau Biomarkers for Alzheimer’s Disease. *PLoS ONE* (2013).

- [73] METZ, C. E. Basic principles of ROC analysis. *Seminars in Nuclear Medicine* 8, 4 (1978), 283–298.
- [74] MITCHELL, A. J., AND SHIRI-FESHKI, M. Rate of progression of mild cognitive impairment to dementia - meta-analysis of 41 robust inception cohort studies. *Acta Psychiatrica Scandinavica* 119, 4 (apr 2009), 252–265.
- [75] MOSIER, C. I. I. Problems and Designs of Cross-Validation 1. *Educational and Psychological Measurement* 11, 1 (apr 1951), 5–11.
- [76] NATIONAL INSTITUTE ON AGING. Osteoarthritis, 2017.
- [77] NISHII, R. Asymptotic Properties of Criteria for Selection of Variables in Multiple Regression. *The Annals of Statistics* (1984).
- [78] OROZCO-SANCHEZ, J., TREVINO, V., MARTINEZ-LEDESMA, E., FARBER, J., AND TAMEZ-PENA, J. Exploring Survival Models Associated with MCI to AD Conversion: A Machine Learning Approach. *bioRxiv* (nov 2019), 836510.
- [79] PAGANI, M., DE CARLI, F., MORBELLI, S., ÖBERG, J., CHINCARINI, A., FRISONI, G., GALLUZZI, S., PERNECZKY, R., DRZEZGA, A., VAN BERCKEL, B., OSSENKOPPELE, R., DIDIC, M., GUEDJ, E., BRUGNOLO, A., PICCO, A., ARNALDI, D., FERRARA, M., BUSCHIAZZO, A., SAMBUCETI, G., AND NOBILI, F. Volume of interest-based [18F]fluorodeoxyglucose PET discriminates MCI converting to Alzheimer’s disease from healthy controls. A European Alzheimer’s Disease Consortium (EADC) study. *NeuroImage: Clinical* 7 (jan 2015), 34–42.
- [80] PAGANI, M., NOBILI, F., MORBELLI, S., ARNALDI, D., GIULIANI, A., ÖBERG, J., GIRTLE, N., BRUGNOLO, A., PICCO, A., BAUCKNEHT, M., PIVA, R., CHINCARINI, A., SAMBUCETI, G., JONSSON, C., AND DE CARLI, F. Early identification of MCI converting to AD: a FDG PET study. *European journal of nuclear medicine and molecular imaging* 44, 12 (nov 2017), 2042–2052.
- [81] PATERSON, R. W., SLATTERY, C. F., POOLE, T., NICHOLAS, J. M., MAGDALINOU, N. K., TOOMBS, J., CHAPMAN, M. D., LUNN, M. P., HESLEGRAVE, A. J., FOIANI, M. S., WESTON, P. S., KESHAVAN, A., ROHRER, J. D., ROSSOR, M. N., WARREN, J. D., MUMMERY, C. J., BLENNOW, K., FOX, N. C., ZETTERBERG, H., AND SCHOTT, J. M. Cerebrospinal fluid in the differential diagnosis of Alzheimer’s disease: Clinical utility of an extended panel of biomarkers in a specialist cognitive clinic. *Alzheimer’s Research and Therapy* 10, 1 (mar 2018).
- [82] PELAEZ-BALLESTAS, I., SANIN, L., MORENO-MONTOYA, J., ALVAREZ-NEMEGYEI, J., BURGOS-VARGAS, R., GARZA-ELIZONDO, M., RODRIGUEZ-AMADO, J., GOYCOCHEA-ROBLES, M., MADARIAGA, M., ZAMUDIO, J., SANTANA, N., CARDIEL, M., AND GRUPO DE ESTUDIO EPIDEMIOLOGICO DE ENFERMEDADES MÚSCULO ARTICULARES (GEEMA). Epidemiology of the Rheumatic Diseases in Mexico. A Study of 5 Regions Based on the COPCORD Methodology. *The Journal of Rheumatology Supplement* 86, 0 (jan 2011), 3–8.

- [83] PETERFY, C. G., SCHNEIDER, E., AND NEVITT, M. The osteoarthritis initiative: report on the design rationale for the magnetic resonance imaging protocol for the knee. *Osteoarthritis and cartilage* 16, 12 (dec 2008), 1433–41.
- [84] PETERSEN, R. C., AISEN, P. S., BECKETT, L. A., DONOHUE, M. C., GAMST, A. C., HARVEY, D. J., JACK, C. R., JAGUST, W. J., SHAW, L. M., TOGA, A. W., TROJANOWSKI, J. Q., WEINER, M. W., AND WEINER, M. W. Alzheimer’s Disease Neuroimaging Initiative (ADNI): clinical characterization. *Neurology* 74, 3 (jan 2010), 201–9.
- [85] POPP, J., WOLFSGRUBER, S., HEUSER, I., PETERS, O., HÜLL, M., SCHRÖDER, J., MÖLLER, H. J., LEWCZUK, P., SCHNEIDER, A., JAHN, H., LUCKHAUS, C., PERNECZKY, R., FRÖLICH, L., WAGNER, M., MAIER, W., WILTFANG, J., KORNHUBER, J., AND JESSEN, F. Cerebrospinal fluid cortisol and clinical disease progression in MCI and dementia of Alzheimer’s type. *Neurobiology of Aging* 36, 2 (feb 2015), 601–607.
- [86] QIN, J., BARBOUR, K. E., NEVITT, M. C., HELMICK, C. G., HOOTMAN, J. M., MURPHY, L. B., CAULEY, J. A., AND DUNLOP, D. D. Objectively measured physical activity and risk of knee osteoarthritis. *Medicine and Science in Sports and Exercise* 50, 2 (2018), 277–283.
- [87] RAGHAVAN, N., SAMTANI, M. N., FARNUM, M., YANG, E., NOVAK, G., GRUNDMAN, M., NARAYAN, V., AND DIBERNARDO, A. The ADAS-Cog revisited: Novel composite scales based on ADAS-Cog to improve efficiency in MCI and early AD trials. *Alzheimer’s and Dementia* 9, 1 SUPPL. (feb 2013).
- [88] RASCHKA, S. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. Tech. rep., 2018.
- [89] RAYKAR, V., STECK, H., KRISHNAPURAM, B., DEHING-OBERIJE, C., AND LAMBIN, P. On ranking in survival analysis: Bounds on the concordance index. In *Advances in Neural Information Processing Systems 20* (2008).
- [90] REUTER, M., SCHMANSKY, N. J., ROSAS, H. D., AND FISCHL, B. Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage* 61, 4 (jul 2012), 1402–1418.
- [91] ROBIN, X., TURCK, N., HAINARD, A., TIBERTI, N., LISACEK, F., SANCHEZ, J.-C., AND MÜLLER, M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics* 12 (mar 2011), 77.
- [92] RODRÍGUEZ, G. Survival Models. New Jersey, 2010, ch. Seven.
- [93] RODRIGUEZ-ROJAS, J., GARZA-MONTEMAYOR, M., TREVINO-ALVARADO, V., AND TAMEZ-PENA, J. Predictive features of breast cancer on Mexican screening mammography patients. C. L. Novak and S. Aylward, Eds., vol. 8670, International Society for Optics and Photonics, p. 867023.

- [94] ROOS, E. M., AND LOHMANDER, L. S. The Knee injury and Osteoarthritis Outcome Score (KOOS): From joint injury to osteoarthritis, nov 2003.
- [95] RUPERT, M. *Survival analysis*. University of Standford - Division of biostatistics, Standford - California, 1980.
- [96] SALVATORE, C., CERASA, A., AND CASTIGLIONI, I. MRI Characterizes the Progressive Course of AD and Predicts Conversion to Alzheimer's Dementia 24 Months Before Probable Diagnosis. *Frontiers in aging neuroscience* 10 (2018), 135.
- [97] SCHWARZ, G. Estimating the Dimension of a Model. *The Annals of Statistics* 6, 2 (mar 1978), 461–464.
- [98] SIMON, N., FRIEDMAN, J., HASTIE, T., AND TIBSHIRANI, R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software* (2011).
- [99] SPERLING, R. A., JACK, C. R., BLACK, S. E., FROSC, M. P., GREENBERG, S. M., HYMAN, B. T., SCHELTENS, P., CARRILLO, M. C., THIES, W., BEDNAR, M. M., BLACK, R. S., BRASHEAR, H. R., GRUNDMAN, M., SIEMERS, E. R., FELDMAN, H. H., AND SCHINDLER, R. J. Amyloid-related imaging abnormalities in amyloid-modifying therapeutic trials: recommendations from the Alzheimer's Association Research Roundtable Workgroup. *Alzheimer's & dementia : the journal of the Alzheimer's Association* 7, 4 (jul 2011), 367–85.
- [100] STEHMAN, S. V. Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment* 62, 1 (oct 1997), 77–89.
- [101] STRITTMATTER, W. J., AND ROSES, A. D. Apolipoprotein E and Alzheimer's Disease. *Annual Review of Neuroscience* 19, 1 (mar 1996), 53–77.
- [102] SUK, H.-I., AND SHEN, D. Deep learning-based feature representation for AD/MCI classification. *Medical image computing and computer-assisted intervention : MIC-CAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention* 16, Pt 2 (2013), 583–90.
- [103] TABLEMAN, M. *Survival Analysis Using S/R*. 2005.
- [104] TAMEZ-PENA, J., MARTINEZ-TORTEYA, A., AND ALANIS, I. Package 'FRESA.CAD' Feature Selection Algorithms for Computer Aided Diagnosis, 2016.
- [105] TAMEZ-PENA, J., RODRIGUEZ-ROJAS, J., GOMEZ-RUEDA, H., CELAYA-PADILLA, J., RIVERA-PRIETO, R., PALACIOS-CORONA, R., GARZA-MONTEMAYOR, M., CARDONA-HUERTA, S., AND TREVIÑO, V. Radiogenomics analysis identifies correlations of digital mammography with clinical molecular signatures in breast cancer. *PLOS ONE* 13, 3 (mar 2018), e0193871.

- [106] TANG, X., HOLLAND, D., DALE, A. M., MILLER, M. I., AND ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE, F. T. A. D. N. APOE Affects the Volume and Shape of the Amygdala and the Hippocampus in Mild Cognitive Impairment and Alzheimer's Disease: Age Matters. *Journal of Alzheimer's disease : JAD* 47, 3 (2015), 645–60.
- [107] TAYLOR, J. R. J. R., AND JOHN. *An introduction to error analysis : the study of uncertainties in physical measurements*. University Science Books, 1997.
- [108] THAL, L. J., KANTARCI, K., REIMAN, E. M., KLUNK, W. E., WEINER, M. W., ZETTERBERG, H., GALASKO, D., PRATICO, D., GRIFFIN, S., SCHENK, D., AND SIEMERS, E. The role of biomarkers in clinical trials for Alzheimer disease. 6–15.
- [109] THE AMERICAN CANCER SOCIETY MEDICAL AND EDITORIAL CONTENT TEAM. *What Is Breast Cancer?*, 2017.
- [110] THE SUSAN G. KOMEN BREAST CANCER FOUNDATION. *PAM50 (PROSIGNA)*, 2017.
- [111] THERNEAU, T. M. *A Package for Survival Analysis in S*, 2015.
- [112] THERNEAU, T. M., AND GRAMBSCH, P. M. Estimating the Survival and Hazard Functions. 2000, pp. 7–37.
- [113] THERRIAULT, J., BENEDET, A. L., PASCOAL, T. A., MATHOTAARACHCHI, S., SAVARD, M., CHAMOUN, M., THOMAS, E., KANG, M. S., LUSSIER, F., TISSOT, C., SOUCY, J. P., MASSARWEH, G., REJ, S., SAHA-CHAUDHURI, P., POIRIER, J., GAUTHIER, S., AND ROSA-NETO, P. APOE $\epsilon$ 4 potentiates the relationship between amyloid- $\beta$  and tau pathologies. *Molecular Psychiatry* (2020).
- [114] TIBSHIRANI, R. Regression Shrinkage and Selection via the Lasso Robert Tibshirani. *Journal of the Royal Statistical Society. Series B (Methodological)* (1996).
- [115] TIULPIN, A., THEVENOT, J., RAHTU, E., LEHENKARI, P., AND SAARAKKALA, S. Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach. *Scientific Reports* (2018).
- [116] US FOOD & DRUG ADMINISTRATION. *For Women - Mammograms*, 2018.
- [117] WEN, C., ZHANG, A., QUAN, S., AND WANG, X. BeSS: An R Package for Best Subset Selection in Linear, Logistic and CoxPH Models.
- [118] WESTMAN, E., MUEHLBOECK, J.-S., AND SIMMONS, A. Combining MRI and CSF measures for classification of Alzheimer's disease and prediction of mild cognitive impairment conversion. *NeuroImage* 62, 1 (aug 2012), 229–38.
- [119] WIBMER, A., HRICAK, H., GONDO, T., MATSUMOTO, K., VEERARAGHAVAN, H., FEHR, D., ZHENG, J., GOLDMAN, D., MOSKOWITZ, C., FINE, S., REUTER, V., EASTHAM, J., SALA, E., AND VARGAS, H. Haralick texture analysis of prostate



- MRI: utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different Gleason scores. *European Radiology* 25, 10 (oct 2015), 2840–2850.
- [120] WOLBERG, W. H., STREET, W. N., HEISEY, D. M., AND MANGASARIAN, O. L. Computerized Breast Cancer Diagnosis and Prognosis From Fine-Needle Aspirates. *Archives of Surgery* 130, 5 (1995), 511–516.
- [121] WORLD HEALTH ORGANIZATION. WHO — Chronic rheumatic conditions, 2016.
- [122] WORTMANN, M. Dementia: a global health priority - highlights from an ADI and World Health Organization report. *Alzheimer's Research & Therapy* 2012 4:5 4, 5 (sep 2012), 40.
- [123] WU, J., CHEN, Y., AND GREENES, R. Healthcare technology management competency and its impacts on IT–healthcare partnerships development. *International Journal of Medical Informatics* 78, 2 (feb 2009), 71–82.
- [124] WYMAN, B. T., HARVEY, D. J., CRAWFORD, K., BERNSTEIN, M. A., CARMICHAEL, O., COLE, P. E., CRANE, P. K., DECARLI, C., FOX, N. C., GUNTER, J. L., HILL, D., KILLIANY, R. J., PACHAI, C., SCHWARZ, A. J., SCHUFF, N., SENJEM, M. L., SUHY, J., THOMPSON, P. M., WEINER, M., JACK, C. R., AND ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE. Standardization of analysis sets for reporting results from ADNI MRI data. *Alzheimer's & Dementia* 9, 3 (may 2013), 332–337.
- [125] ZHAO, H., LI, X., WU, W., LI, Z., QIAN, L., LI, S., ZHANG, B., AND XU, Y. Atrophic Patterns of the Frontal-Subcortical Circuits in Patients with Mild Cognitive Impairment and Alzheimer's Disease. *PloS one* 10, 6 (2015), e0130017.

