Ana Sofia Fachada Fernandes

PROGNOSTIC MODELLING OF BREAST CANCER PATIENTS – A BENCHMARK OF PREDICTIVE MODELS WITH EXTERNAL VALIDATION

Dissertação apresentada para obtenção do Grau de Doutor em Engenharia Electrotécnica e de Computadores – Sistemas Digitais e Percepcionais pela Universidade Nova de Lisboa, Faculdade de Ciências e Tecnologia.

LISBOA
2010

Sumário

Existem inúmeros modelos de prognóstico clínico na área médico e particularmente no prognóstico do cancro da mama. Previamente à sua utilização clínica, os modelos de prognóstico necessitam de ser aplicados a pacientes provenientes de diferentes centros médicos, de forma a que sejam submetidos a uma rigorosa validação. Esta tese avalia a precisão preditiva de um modelo flexível com uma regularização Bayseana, o PLANN-ARD. Para tal utiliza uma base de dados composta por 4016 registos de pacientes com cancro da mama, diagnosticados de 1989 a 1993 e identificados pelo BCCA, Canada, com um *follow-up* de 10 anos. Este método é comparado com a regressão de Cox, sendo considerado o modelo mais utilizado neste género de análise.

Ambos os métodos foram ajustados a 931 pacientes cujos dados de rotina foram adquiridos e diagnosticados entre 1990 e 1994 no Christie Hospital, UK, com um *follow-up* de 5 anos. Nesta tese foram desenvolvidos avanços metodológicos significativos que suportam a validação externa desta rede neuronal com dados clínicos, nomeadamente: imputação dos dados em falta em ambas as bases de dados, treino e validação e um índice de prognóstico que permite a estratificação dos pacientes em grupos de risco diferentes. A precisão preditiva dos modelos foi medida empiricamente utilizando o índice de descriminação standard, C^{td} e uma medida de calibração utilizando o teste estatístico, Hosmer-Lemeshow.

Verificou-se que ambos os modelos, regressão de Cox e PLANN-ARD têm uma discriminação semelhante, sendo que a rede neuronal demonstrou uma precisão preditiva marginalmente superior durante o período de 5 anos de *follow-up*. Para além desta melhoria, a

rede neuronal regularizada tem a vantagem substancial de ser adaptada para efectuar previsões da função de risco e sobrevivência de pacientes individuais.

São propostas quatro diferentes abordagens de estratificação de pacientes em grupos de risco, cada qual com um fundamento diferente. Embora tenha sido verificado que as quatro metodologias são concordantes entre elas, foram identificadas diferenças importantes entre elas. Foram extraídas e comparadas as regras das duas metodologias de estratificação, o "log-rank bootstrap" e a aplicação directa das árvores de regressão, e para as duas metodologias de extracção de regras, OSRE e CART, respectivamente.

Os índices de prognóstico clínico de cancro da mama, como o NPI, TNM e *St. Gallen consensus rules* foram também comparados com os modelos de prognóstico propostos representados como árvores de regressão, onde se pôde concluir que as abordagens propostas podem melhorar a prática clínica corrente.

Por fim, é proposto um sistema clínico Web de suporte à decisão para médicos oncologistas e para pacientes com cancro da mama, onde é efectuada uma avaliação prognóstica, adaptada às características particulares de cada paciente.

Abstract

There are several clinical prognostic models in the medical field. Prior to clinical use, the outcome models of longitudinal cohort data need to undergo a multi-centre evaluation of their predictive accuracy. This thesis evaluates the possible gain in predictive accuracy in multi-centre evaluation of a flexible model with Bayesian regularisation, the (PLANN-ARD), using a reference data set for breast cancer, which comprises 4016 records from patients diagnosed during 1989-93 and reported by the BCCA, Canada, with follow-up of 10 years. The method is compared with the widely used Cox regression model.

Both methods were fitted to routinely acquired data from 743 patients diagnosed during 1990-94 at the Christie Hospital, UK, with follow-up of 5 years following surgery. Methodological advances developed to support the external validation of this neural network with clinical data include: imputation of missing data in both the training and validation data sets; and a prognostic index for stratification of patients into risk groups that can be extended to non-linear models. Predictive accuracy was measured empirically with a standard discrimination index, C^{td}, and with a calibration measure, using the Hosmer-Lemeshow test statistic.

Both Cox regression and the PLANN-ARD model are found to have similar discrimination but the neural network showed marginally better predictive accuracy over the 5-year follow-up period. In addition, the regularised neural network has the substantial advantage of being suited for making predictions of hazard rates and survival for individual patients.

Four different approaches to stratify patients into risk groups are also proposed, each with a different foundation. While it was found that the four methodologies broadly agree, there

are important differences between them. Rules sets were extracted and compared for the two stratification methods, the log-rank bootstrap and by direct application of regression trees, and with two rule extraction methodologies, OSRE and CART, respectively.

In addition, widely used clinical breast cancer prognostic indexes such as the NPI, TNM and St. Gallen consensus rules, were compared with the proposed prognostic models expressed as regression trees, concluding that the suggested approaches may enhance current practice.

Finally, a Web clinical decision support system is proposed for clinical oncologists and for breast cancer patients making prognostic assessments, which is tailored to the particular characteristics of the individual patient. This system comprises three different prognostic modelling methodologies: the NPI, Cox regression modelling and PLANN-ARD. For a given patient, all three models yield a generally consistent but not identical set of prognostic indices that can be analysed together in order to obtain a consensus and so achieve a more robust prognostic assessment of the expected patient outcome.

Symbols and Notations	Description
AIC	Akaike's Information criterion
AJCC	American Joint Committee on Cancer
ANN	Artificial neural network
ARD	Automatic Relevance Determination
BCCA	British Columbia Cancer Agency
CART	Classification and Regression Trees
CCI	Crude cumulative incidence
DFS	Disease-free survival
EPV	Events per variable
ER	Oestrogen receptor
FIGO	International Federation of Gynecology and Obstetrics
HER2	Human epidermal growth factor receptor 2
KM	Kaplan Meier
MAR	Missing at random
MCAR	Missing completely at random
MI	Multiple Imputation
MLP	Multilayer perceptron
MNAR	Missing not at random
NPI	Nottingham Prognostic Index
OS	Overall Survival
OSRE	Orthogonal Search Rule Extraction
PgR	Progesterone receptors
PI	Prognostic Index
PLANN-ARD	Partial Logistic Artificial Neural Networks with Automatic Relevance Detection
PLSPL	Partial logistic spline
PVI	Peritumoural vascular invasion
ROC	Receiver operating characteristic
TNM	Tumour, Nodes, Metastasis

Index of Contents

Index of Figures

Index of Tables

Chapter 1 -

Introduction

Predictions of the prognosis for a patient diagnosed with cancer has an important clinical role in informing decisions on the choice of adjuvant therapy, in particular to minimise the risk of under- or over-treatment. Current interest in the development of personalised medicine increasingly requires the specialisation of clinical outcome predictions at the level of the individual patient. Moreover, patient empowerment for "shared decision-making" makes physicians and patients both active participants in deciding on the choices for therapeutic interventions. This requires accurate communication and transparent explanations about the prognosis of the disease, in order to permit a well-founded assessment of the risks and benefits of particular treatment choices. In clinical practice, prediction models inform patients and their physician on the probability of a diagnosis or a prognostic outcome as well as stratifying the patients according into risk groups, which can be useful for communication between physicians. Stratification is also important for promoting consistent care protocol between physicians and in the design and assessment of clinical trial outcomes by comparing like-for-like patients.

However, it is imperative for the application of the model that its predictions are reliable, quantifying how accurate the predictions from the model are, which is the "Model Performance". Furthermore, the gold standard for performance evaluation is to carry out an external validation, that is to say predicting for patients from a clinical centre that is different from the clinical centres from which patient data were acquired for model development (Lisboa, 2002).

There are several clinical prognostic classification schemes proposed for breast cancer patients, some of which discriminate between the survival of different risk groups defined from the patient characteristics. A general criterion defined by the World Health Organisation is the so-called Tumour, Nodes and Metastasis (TNM) staging system, which is a rule-based filter whose arguments are the ordinal representations of tumour size (T), the extent of spread of the disease to the lymph nodes (N) and the presence of clinical signs of metastatic spread (M). The strength of this index is that it only requires clinical information, which is obtained by the clinician without resort to laboratory tests.

A limitation of TNM staging is that its discrimination power is best for separating severe from early-stage disease. Clinically, it is important to differentiate between the severity of illness of patients with non-metastatic disease, sometimes referred to as the 'operable group' since, for them, there is the possibility of completely removing the cancer. However, this requires the addition of histological information about the stage of advancement of the cancerous tissue cells. This is provided in one of the most widely used early stage clinical indices for breast cancer, the Nottingham prognostic index (NPI) (Haybittle, Blamey, Elston, Johnson, Doyle, Campbell, Nicholson, Griffiths, 1982). This index combines the pathological size of the tumour, measured in cm, together with an integer index of *histological grade* in three discrete groups (good, moderate and poor differentiation) and the number of axillary lymph nodes affected, also in three groups.

Note that this index comprises a linear combination of discrete categories, forming an analytical scoring index which can be represented as $\beta.x$ in a Cox regression model. The second point to note about this robust and time honoured score is that is relies on a careful non-uniform discretisation of the categorical variables of histology and nodes affected. It is therefore a non-linear index, albeit expressed linearly in terms of discrete indicators. The model was fitted to explain the variation in survival, lending itself naturally to the derivation of a discrimination index that has since undergone extensive multi-centre validation (Galea, Blamey, Elston, Ellis, 1992).

Prognostic indexes have been proposed both as a continuous score, providing a risk estimate for individual patients, and on a discrete basis, for defining a limited number of risk groups from which non-parametric grouped survival estimates can be obtained, for instance using Kaplan-Meier, or actuarial, methods. A further score is the consensus rules agreed and periodically reviewed by the St. Gallen group (Harbeck, Jakesz, 2007).

More recently, there was much interest in deriving predictive indices to estimate the likely survival of individual patients, rather than discriminating between different risk groups. One such model is AdjuvantOnline (Ravdin, Sminoff, Davis, Mercer, Hewlett, Gerson, Parker, 2001), which goes further than pure prognosis and reports also the impact on 10-year survival from different choices of adjuvant therapy. This model is derived from meta-analysis, rather than by fitting a single empirical data set from a longitudinal cohort study, although it has been validated using the same cohort with 10-year follow-up that serves as the reference data set for this study (Olivotto, Bajdik, Ravdin, Speers, Coldman, Norris, Davis, Chia, Gelmon, 2005). However, it does not report confidence intervals for its predictions.

Although the final power of any index is limited by the substantial unexplained prognostic heterogeneity and the data accuracy of the adopted retrospective databases, it is relevant to ask whether a generic non-linear database methodology can be developed. This will remove the need to limiting assumptions such as the proportionality of hazards in Cox regression and to remove also the need to discretise continuous variables in order to obtain piecewise linear models using linear-in-the-parameter methodologies. These models have the potential to provide better accuracy than traditional methodologies because of their flexibility due to a semi-parametric formulation using distributed nodes, in the case of artificial neural networks. Clearly the generalisation of the model needs to be controlled with a robust regularisation framework and evaluated empirically with a carefully designed external validation study.

This thesis evaluates the possible advantages of a non-linear, time dependent neural network methodology for survival modelling of a single risk, or outcome. This is the Partial Logistic Artificial Neural Network regularized with a Bayesian network using a Laplace approximation of the evidence, known as Automatic Relevance Determination (MacKay, 1995), hence PLANN-ARD. In particular, the thesis reports the first large-scale performance evaluation of this methodology for an external data set, using imputation and risk stratification. This is benchmarked with the stalwart linear model for survival, Cox regression. According to an easier interpretation of covariate effects, the non-linearities that are inherent in outcome modelling of medical data are often naïvely taken into account implicitly by collecting real-valued covariates binned into discrete groups. Consequently, in such a framework, a flexible model such as the neural network can only improve modelling accuracy by implicitly modelling non-additive (interactions) effects between covariates and non-proportional covariate effects, which in linear models need to be parameterised explicitly.

The PLANN-ARD model takes account of censorship that occurs when a patient drops out of the study before the event of interest is observed, for instance being lost to follow-up without a definitive outcome being recorded.

A related matter of clinical relevance is the choice of outcome to model, i.e. the choice of risk. In general there are three possibilities, for breast cancer. The most generic and best defined risk is that of overall mortality, since this a well-defined endpoint in prognostic assessment. It has the limitation of combining together the effects of breast cancer and age-related mortality. However, the alternative of specifying breast specific death suffers from potential inaccuracies in the attribution of the cause of death, for instance in the case of patients who die from a heart attack which may potentially reflect the health load resulting from radiotherapy to the left side of the chest or the toxicity of chemotherapy. A third possibility is to track local and distal recurrences of the disease. This represents multiple competing risks and suffers from potential bias because of the absence from the available databases of follow-up for the occurrence of second primary tumours. Taking all of these factors into consideration, current prognostic decision support systems such as Adjuvant report overall or relative mortality. This thesis focuses on the primary modelling stage for both of these outcome indicators, which is the risk of overall mortality as a function of age and clinical and histological measurements specific to breast cancer.

This thesis also takes in account with the important issue of variable selection by making a carefully analysis in order to incorporate the best predictors of the outcome variables given the cohort sample size.

In addition to the above issues, routinely acquired data from clinical practice commonly contain missing values. Moreover, the data is not missing at completely at random since there may be a reason for missing, for instance because the variables concerned are of no consequence to the choice of therapy. So it is that the survival of the patient group with a particular variable missing do not always lie between the survival groups for extant values, but sometimes are close to the extremes of survival. However, the data used is missing at random, in the sense that the missing can be reasonably imputed from ancillary information, such as the eventual choice of therapy, by representing missing values as random variables (Fernandes, Jarman, Etchells, Fonseca, Biganzoli, Bajdik, Lisboa, 2008). Missing values were incorporated in the benchmark linear model, Cox proportional hazards and the non-linear model, the PLANN-ARD and both were compared.

When considering out-of-sample predictions it is necessary to distinguish between modelling the training data, which may have missing values; then predicting on an out-of-sample cohort which may also have missing values, possibly in a different set of covariates than the training data, or missing with a different distribution; and predicting for a single new patient, for which data imputation may not be possible. Therefore, this work analyses how these issues can be overcome.

Both prognostic models being compared in this study define a different prognostic index that ranks patient data by severity of the illness that incorporate all the issues mentioned in model developing and were extended to the out-of-sample cohort.

A critical performance indicator in the validation of prognostic models, is the assessment of discrimination and calibration accuracy, both of which are highly relevant in decision support. This performance assessment is carried out in two different ways. First, by a detailed analysis of the predicted *vs* observed survival over the five years of follow-up for the training data. Second, a prognostic index is defined, with which to assess the discrimination between patient groups, to be evaluated against the crude empirical event rate over the full follow-up period for the validation data, which has a 10 year timeframe.

Once the risk score is defined, the population of patients at risk needs to be stratified for the purpose of tailoring adjuvant therapy and to enable comparisons to be made between patient cohorts from different clinical centres, or subject to different clinical interventions, to be made between patients at similar risk by outcome. This involves the application of significance tests, which in survival analysis is usually the log-rank statistic. However this statistic finds the different patient risk groups by thresholding only the Prognostic Index, making an assumption that the threshold separates distinct patient populations, while in practice it may be cutting across a single patient population. It would be desirable to stratify by identifying distinct patient populations directly from the prognostic factors. Therefore, this thesis make an analysis between 4 different stratification methodologies: the first is a log-rank bootstrap aggregation methodology, which uses the log-rank statistic at its core but carries out bootstrap re-sampling of the population of prognostic indices in order to gain robustness over a maximum significant search. The second methodology is based on regression trees, applied to the continuous value prognostic scores. The third methodology is a robust unsupervised clustering methodology that uses k-means where only patient covariates without any

knowledge of outcome are considered. The fourth one uses informed clustering based on the principle of learning metrics.

Models were trained on a cohort of patients with operable breast cancer recruited at Christie Hospital, Manchester, between 1990-93 (n=743) and were the subject of an external predictive evaluation for overall mortality by applying the model to a database acquired by the British Columbia Cancer Agency (BCCA), Vancouver, during the period 1989-93 (n=4,016).

This study also enhances the existing classification schemes proposed for breast cancer previously mentioned, by comparing them with the PLANN-ARD and Cox proportional hazards modelling followed by a robust stratification methodology. The survival for patient sub-groups was compared for the different methods, revealing the heterogeneity among the prognostic groups of the existent classification schemes. An advancement achieved by PLANN-ARD is to reliably stratify the NPI group 3 of breast cancer patients, which is thought to contain a mixture of patient groups with different severity of illness. The results reported in the thesis identify three sub-groups with statistically different 5 year survival.

Finally, a web decision support system for breast cancer patients was developed in order to incorporate patient's risk group models, such as the known NPI, the Cox proportional hazards prognostic index and the PLANN-ARD prognostic index both followed by a stratification methodology. The derived explanatory rules, the different treatments received by patients and the KM survival curves were also incorporated to help on the visualisation of relevant existing patient data an interpretation of inferences in clinical terms. It is important to mention that the aim of the proposed decision support system is to enhance the oncologists' current practices, rather than to replace them. In this decision support framework, the predictive model represents an analytical window into the evidence base comprised of historical patient records. In particular, the model serves to provide a context for the patient, using the risk score and risk group strata, as well as an individualised inference of the expected prognostic outcome, through the predicted hazard for that patient.

Chapter 2 of this thesis presents an overview of the classical survival models and flexible models, including the semi-parametric linear model and PLANN-ARD model, explaining how it can be applied to prognostic modelling. It also describes the existing and proposed stratification methodologies as how Boolean rules extraction can be achieved. In addition, chapter 2 gives a description of the existing classification schemes and their importance.

Chapter 3 starts by outlining the main issues that must be taken in account while developing prediction models which were introduced and analysed in this work. Secondly, it defines how the predictive accuracy and validation must be analysed and the modelling strategy that must be taken in account when developing prognostic models.

Chapter 4 describes the validation results obtained with the flexible modelling approach. It begins with an explanation of the two datasets used for training and external validation and a study of the missing data and the application of multiple imputation. Second, the integration of the multiple imputation into the linear and non-linear modelling methodologies is described, leading to the evaluation of the predictive performance for the two alternative models, each applied to an external validation data set which was not used at all during the optimisation of model fitting to the training data, and in fact from a completely different patient sample, recruited in a different country. The results from the different stratification methodologies, applied to the proposed prognostic models are presented and compared. Finally the relative validation performance of the existent and new prognostic and risk stratification schemes is presented, explaining how these new methodologies can improve those currently used in clinical practice.

Finally, chapter 5 presents an overview of the existing online breast cancer prognostic models, with particular emphasis the AdjuvantOnline. It also defines the web clinical decision support system for breast cancer patients that incorporates the prognostic models proposed in this thesis.

1.1 - Contribution of the thesis

The present thesis makes an important contribution to both technical innovation and clinical application as several important novelties were added or changed to current practice. Currently there are several survival models which are in use described previuosly, such as NPI and other Cox proportional hazards models. It is intended to augment NPI by adding more variables considered to be important in the prognostic model. Moreover, it was intended to define a prognostic model to become predictive rather than explanatory as well as modelling non-linear dependences, with PLANN-ARD.

This thesis also takes account of missing data and censorship within principled frameworks, applying multiple imputation in combination with neural network models for time-to-event modelling, where a new prognostic index was also considered, which able the stratification of patients into risk groups.

Moreover a new stratification methodology was developed, based on decision trees, which adds a more robust path to identify the patient's risk group and the explanatory rules that characterize risk group membership, based on patient's characteristics. Flexible modelling, incorporating the missing data was also subjected to a very accurate validation.

Finally, a new web decision support system contributes to technical innovation as it implements both the previously mentioned models, where all can be compared.

1.2 - Articles accepted

- **"**Assessment of benefit *vs.* risk of drug therapy: the potential for outcome analysis with flexible models", accepted at "2010 International Joint Conference on Neural Networks", Barcelona (Spain), 18-23 July 2010, Lisboa, P.J.G, Fernandes, A.S., Fonseca, J.M., Bajdik, Chris, Biganzoli, Elia.
- "Cohort-based Kernel Visualisation with Scatter Matrices", accepted at "2010 International Joint Conference on Neural Networks", Barcelona (Spain), 18-23 July 2010, Romero, Enrique, Fernandes, A.S., Mu, Tingting, Lisboa, P.J.G.
- "A clinical decision support system for breast cancer patients", accepted at "Doctoral conference on computing, electrical and industrial systems", Lisboa (Portugal), 22-24 February 2010, Fernandes, A.S., Alves, Pedro, Jarman, Ian H., Etchells, Terence A. , Fonseca, José M., Lisboa, Paulo J. G..
- "p-Health in breast oncology: a framework for predictive and participatory e-systems", accepted at "International Conference on "Developments in eSystems Engineering" (DeSE '09)", Abu Dhabi (United Arab Emirates); 14-16 December 2009; Fernandes, A.S., Bacciù, D., Etchells, T.A., Jarman, I.H., Fonseca, J.M., Lisboa, P.J.G.
- "Different methodologies for patient stratification using survival data", accepted at "Sixth International meeting on computational intelligence methods for bioinformatics and biostatistics", special session "Intelligent systems for medical decisions support (ISMDS)"; Génova (Italia); 15-17 October 2009; Fernandes, A.S., Bacciù, D., Etchells, T.A., Jarman, I.H., Fonseca, J.M., Lisboa, P.J.G.
- "Evaluation of missing data imputation in longitudinal cohort studies in breast cancer survival"; Int. J. Knowledge Engineering and Soft Data Paradigms, Vol. 1, No. 3, 2009, pp. 257-276; Fernandes, A.S., Etchells, T.A., Jarman, I.H., Fonseca, J.M., Biganzoli, Elia, Bajdik, Chris, Lisboa, P.J.G.
- "Stratification methodologies for neural networks models of survival", Proceedings of the 10th International Work-Conference on Artificial Neural Networks (IWANN2009), Salamanca, Spain; vol. 5517 – Pág. 989-996; Springer (2009); Fernandes A.S., Etchells T.A., Jarman, I.H., Fonseca, J.M., Biganzoli, Elia, Bajdik, Chris, Lisboa, P.J.G.

- "Missing data imputation in Longitudinal Cohort studies – Application of PLANN-ARD in Breast cancer Survival"; The 7th International Conference on Machine Learning and Application (ICMLA'08), Vol. 5177, pp. 214-221, Springer (2008); San Diego, California; Fernandes, A.S., Etchells, T.A., Jarman ,I.H., Fonseca, J.M., Biganzoli, Elia, Bajdik, Chris, Lisboa, P.J.G.
- "Stratification of severity of illness indices: a case study for breast cancer prognosis"; Proceedings of the 12th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES'08), Zagreb,Croatia; vol. 5178 – Pág 214-221; Springer; Etchells T.A., Fernandes, A.S., Jarman, I.H., Fonseca J.M.,Lisboa P.J.G.
- "Stratification of severity of illness índices and out-of-sample validation: a case study for breast cancer prognosis", chapter in "Computational Intelligence in Human Cancer Research"; KES Rapid Research Results Series; Etchells, T.A., Fernandes, A.S., Jarman, I.H., Fonseca, J.M., Lisboa, P.J.G.

Chapter 2 -

Analytical methodologies

This chapter introduces the survival models and artificial intelligence techniques used in this thesis and it is divided in four different sections. The first section gives an overview of classical survival methods, which explains in detail the benchmark model used in this thesis, the Cox proportional hazards modelling. The second section describes flexible models and how these can incorporate survival analysis, introducing the PLANN-ARD model, which is used in this thesis. This chapter also provides a description of the stratification methods that can be applied to a prognostic index obtained with survival modelling. Finally, an overview of clinical prognostic indexes is presented, making a higher emphasis on the existing prognostic indexes in breast cancer field.

2.1 - An overview of classical survival methods

Survival Analysis is composed by statistical methods used to study the occurrence and timing or the events. It analyses the data from a specific time of origin until the occurrence of a particular event. These methods were designed to apply in the study of deaths. However, the survival analyses can be applied in other kind of events, such as equipment failure, automobile accidents, stock market crashes, job terminations, births, marriages, divorces, arrests, and other.

In medical research, the time origin corresponds to the recruitment of the individual and the end-point can be the death of the patient, relief of pain, recurrence of symptoms. The result of the first end-point referred is literally survival times.

Survival analysis was designed for longitudinal data on the occurrence of events, and it is very important to clarify the events in the study. The event can be characterized as a qualitative change (transition from one state to another, such as being alive to being death) and a quantitative change (a stock market crush could be defined as a single day life loss of more than 20% in the market index). However, in survival analysis it is needed to know more than a qualitative or a quantitative change. You also have to situate this event on time.

The survival analysis can be performed considering only the time of events, but a usual aim of these methods is to estimate causal or predictive models in which the risk of an event depends on the covariates These covariates can be constant over the period of study, such as sex and race, or can change over the period of study, such as age, blood pressure, marital status. There are some reasons to consider why survival data can't be obtained with conventional statistical procedures. The first one is related with the fact that survival data is not symmetrically distributed and consequently this type of data does not have a normal distribution. Transforming the data to give a more symmetric distribution could solve this feature. However, a more satisfactory approach is to adopt alternative distributional model to the original data.

Besides this feature, the survival data has also other two features that are difficult to handle with conventional statistical procedures, which are censoring and time-dependent-covariates. The following example illustrates both problems: *a sample of X prisoners was followed one year after release*. The event of interest is the first arrest and the aim of the study is to determine how the occurrence and timing of arrest depends on some covariates. Some of these covariates are constant during the period of study, such as sex or race, others could change ate any time during the period of follow-up. How can this data be analysed using conventional methods? The conventional methods ignore the information about timing of arrest. Ignoring this information should reduce the precision of the estimates. One solution to this problem is to make the dependent variable the length of time between the release and first arrest and then estimate a conventional linear regression model. But what can be made with the persons who were not arrested during the follow-up period? Such cases are called **censored**.

Another problem is how to deal with a time-dependent variable, such as employment status. This variable can be incorporated in the data estimating 52 indicator variables (one variable for each week indicating the employment status). This can lead to a computational awkwardness and statistical inefficiency. Aside to this, there is a more fundamental problem, such as the employment indicators for weeks after an arrest might be a consequence of the arrest rather than the cause. In particular, someone who is jailed after an arrest is not likely to be working full time in subsequent weeks.

In conclusion we can say that conventional methods are not efficient dealing either with censoring and time-dependent covariates. However, the survival analysis methods allow censoring, and many also allow time-dependent covariates.

2.1.1. Censorship and their importance

Censorship is the main feature of survival analysis. The survival time of an individual is said to be censored, not only when the end-point of interest has not been observed for that individual but also, if at the end of the study the individual has not experienced the event. During the follow-up it is necessary to know if the event has occurred or not and when did it occur. Sometimes, the survival status of an individual cannot be known, as that individual has been lost to follow-up. For example, consider that the recruited individual, after being recruited, moves to another country and his survival experience cannot be traced. The only information available is the last date he or she was known to be alive.

Another reason for a survival time being considered censored is when the event experienced is different from the event of interest, which happens due to a cause that is known to be unrelated to the specific study. However, this type of censoring is difficult to be sure that is not related to the study. For example, consider a patient in a clinical trial to compare alternative therapies for prostatic cancer who experiences a fatal road traffic accident. The accident could have resulted from a side effect of the treatment. If so, the death is related with the treatment and cannot be considered censored. A breast cancer patient who dies due to a cause unrelated with the disease, has to be considered censored. This kind of event is known as **competing risk events**. Suppose also that a breast cancer patient undergoes a prophyletic oophorectomy after surgery to breast cancer. This prophyletic treatment substantially reduces the probability of developing ovarian cancer. So it is considered a competing risk event when calculating ovarian cancer incidence. A competing risk may

preclude the onset of the event of interest, or may modify the probability of the event of interest. An individual who experiences a competing risk event is censored in an informative manner. However it is necessary to analyse the data and the event of interest to conclude if the informative censoring makes any influence in the estimate of the probability of the event of interest. If it does not make any influence, the informative censoring can be ignored. For example, death to other causes may not be related to having breast cancer unlike breast cancer-specific mortality. Here, the informative censoring does not influence the estimates of breast cancer mortality. It is important to verify if censoring is noninformative, because otherwise a bias is introduced in survival analysis methods. Resuming, good patient follow-up and avoidance of unnecessary drop-out is the best solution (Bradburn, Clark, Love, Altman, 2003).

2.1.2. Survivor function and Hazard function

In survival analysis there are two functions of central interest such as the **survivor function** and the **hazard function**. The survivor function represents the probability that an individual survives from the time origin to some time beyond t:

$$S(t) = P(T \geq t) \quad \textbf{(1)}$$

The hazard function is widely used to express the risk or hazard of an event at time t, and is obtained from the probability that an individual experience the event at time t, conditional on he or she having survived to that time:

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T < t + \delta t \mid T \geq t)}{\delta t} \right\} = \frac{f(t)}{S(t)} \quad \textbf{(2)}$$

where

$$f(t) = \frac{dF(t)}{dt} \quad \textbf{(3)}$$

and

$$F(t) = \Pr(t \leq T) = \int_{t=0}^{T} f(u)dt = 1 - S(t) \quad \textbf{(4)}$$

This function can also be called hazard rate, instantaneous death rate or force of mortality and represents the approximate probability that an individual dies in the interval (t,t+δt), conditional on that person having survived to time t. The hazard rate is known as the conditional rate of failure. This is the rate of an event, given that a person has survived up to

that time. It can vary from 0 to infinity. It can increase or decrease or remain constant over time. The function H(t), called cumulative hazard, can be obtained from the Survivor function, since:

$$H(t) = -\log S(t) \quad \textbf{(5)}$$

Supposing that in a single sample of survival times none of the observations are censored, the survival function S(t) can be estimated by the *empirical survival function*, given by:

$$S(t) = \frac{Number\ \ of\ \ individuals\ \ with\ \ survival\ \ times\ \geq\ t}{Number\ \ of\ \ individuals\ \ in\ \ the\ \ data\ \ set} \quad \textbf{(6)}$$

The estimated survivor function is assumed to be constant between two adjacent death times. The overall survival probability is the probability of being event-free at least up to a given time. The cumulative incidence of an event at a given time is one minus the overall survival probability at that time.

Consider, for example, the event of interest to be death. Suppose that 100 patients lived for at least 1 year and 5 patients died. The estimated survival at 1 year is 95%. Suppose, at 2 years, 10 patients died. The estimated survival at 2 years is 85/95=89,5%. The estimated overall survival probability up to 2 years is the probability of having survived to first and second year, which is 95*89,5=85%. The cumulative incidence of mortality at 2 years is the sum of mortality at first and second years, which is, in the previous example (95/100)+((95/100)*(10/95)) =15%.

2.1.3. Actuarial/Descriptive Model – Kaplan Meier

There are some methods for the estimation of the survival function and the hazard function. Methods for estimating these functions from a single sample of survival data are said to be non-parametric or distribution free, since they do not require a specific assumption to be made about the distribution of the survival times.

The Kaplan-Meier approach provides a non-parametric estimate of the overall survival probability of an event of interest. It adequately deals with censored data, and provides attractive graphs on the relationship between predictor values and the outcome over time. The cumulative incidence is calculated as 1 minus this survival probability. Every individual in the data set has a follow-up time and status (event or censored). In order to estimate the survivor function, the follow-up times where an event has occurred are ordered from the

smallest to the largest. Then a series of time intervals are constructed. These intervals begin at the time when an event has occurred and end at the time before the next event occurs. No intervals begin at a censored time. It can be noticed that there can be ties, since more than one individual can die simultaneously.

Considering n_j the number of event free individuals up to time t_j and d_j the events occurred at time t_j, the estimated survival probability at time t_j is given by $(n_j-d_j)/n_j$. The overall survival probability up to time t_j, denoted $S(t_j)$ is the probability of surviving up to t_j, including time t_j. Therefore the overall survival probability up to t_j is the product of the probability of surviving trough the interval t_j-1 to t_j and all preceding intervals. Then it can be said that a product of series of estimated probabilities forms the Kaplan-Meier estimate:

$$S(t_j) = S(t_j - 1) \times \frac{n_j - d_j}{n_j} \quad \textbf{(7)}$$

$$S(t) = \prod_{i=1}^{k} \frac{n_i - d_i}{n_i} \quad \textbf{(8)}$$

A plot of the Kaplan-Meier estimate of the survival function is a step-function, where the estimated survival probabilities are constant between adjacent death times and decrease at each death time. In the absence of censoring the Kaplan-Meier estimate is simply the empirical survivor function. Therefore, it can be concluded that the Kaplan-Meier estimate is a generalisation of the empirical survivor function that accommodates censored observation.

In the following example it can be observed how the Kaplan-Meier curve is calculated:

Time interval	Hazard at the beginning of interval	Censored during the interval	Hazard at the end of interval	Deaths in the interval	Survival at the end of interval	Overall survival at the end of interval
0-1	7	0	7	1	6/7=0,86	0,86
1-4	6	2	4	1	¾=0,75	0,86*0,75=0,64
4-10	3	1	2	1	½=0,5	0,86*0,75*0,5=0,31
10-12	1	0	1	0	1/1=1	0,86*0,75*0,5*1=0,31

Table 2.1 – Example for a Kaplan-Meier curve calculation

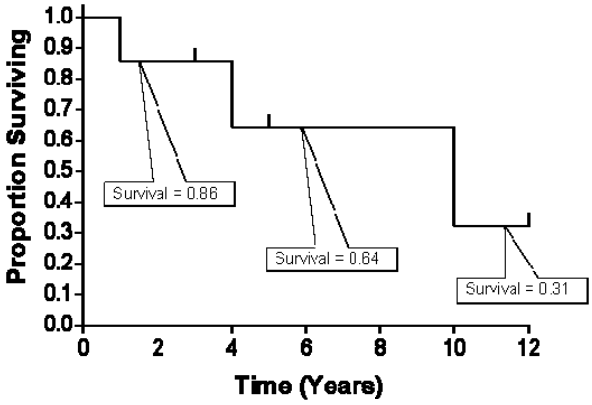

Figure 2.1 – Example of a Kaplan-Meier curve

When competing risk events are present in the data, it is necessary to make some considerations about the Kaplan-Meier estimate. Satagopan et al (Satagopan, Ben-Porat, Berwick, Robson, Kutler, Auerbach, 2004) makes some illustrations of a non-parametric estimation of the cumulative incidence function for an event of interest in the presence of competing risk events

2.1.4. Piecewise Linear Models – Proportional Hazards (Cox regression Model)

The actuarial or descriptive methods described can be useful in the analysis of a single sample of survival data or in the comparison of two or more groups of survival times. However, in most studies, supplementary information (explanatory variables) is also recorded for each recruited individuals. The analysis, using this information is much more complex than the analysis considered before. In order to analyse the relationship between the survival experience and the explanatory variables of the individuals it is used an approach based on statistical modelling. There are two main reasons for modelling survival data. First, one objective of the modelling process is to determine how the explanatory variables affect the hazard function. Another objective is to obtain an estimate of the hazard function for an individual, in order to estimate the median survival time. This value can be estimated for current or future individuals with particular values of the explanatory variables.

When the survival times are assumed to follow a statistical distribution, it should be used a fully parametric model. There are different distributions and the identification of a suitable one is a crucial step in modelling the survival data. What distinguishes between the existing

parametric models is the shape of the hazard they assume the data follows. If the hazard is always increasing and decreasing, the Weibull and Gompertz distributions are appropriate. If the hazard rises to a peak and then decreases or always decreases, then the Log-Logistic distribution should be used. Log-Normal or Generalised Gamma models are preferable used when the hazard rises to a peak and then decreases. In the Exponential model, the hazard is constant over time.

If there is no need to assume a particular form of the probability distributions for the survival times S(t), then the Cox regression model is used. In medical studies, the Cox proportional hazards model it the most often used method of survival outcomes. This model is based on the assumption of proportional hazards, that is, assumes that the log(-log (S(t))) for different subjects are equidistant over time or equivalently that the hazard function for any two subjects are proportional over time. Therefore, this model is referred as semi-parametric model.

As stated before, the Cox regression model is based on the assumption of proportional hazards. The following example can explain this property of the model.

Suppose two patients are randomised to receive a standard treatment or a new treatment. $h_S(t)$ is the hazard of dead at time t the first treatment and $h_N(t)$ is the hazard of the second. According to the proportional hazards model:

$$h_N(t) = \psi h_S(t) \quad \textbf{(9)}$$

This assumption implies that the corresponding true survivor functions for individuals on the new and standard treatment do not cross. The ψ value corresponds to the ratio of the hazards of death at any time for an individual on the new treatment relative to an individual on the standard treatment. The ψ value is known as the relative hazard or hazard ratio.

In the Cox regression model a reference group called the baseline population specifies all time dependence, which is characterized by a zero covariate vector ($h_0(t)$). The values of these covariate will be assumed to have been recorded at the time origin of the study. The set of values of the explanatory variables for each individual will be represented by $\boldsymbol{x}$. The hazard function can be written as:

$$h_p(x_p, t) = \psi(x_p) h_0(t) \quad \textbf{(10)}$$

As the relative hazard cannot be negative, it can be written has $e^{\eta p}$, where η_p is the linear combination of the explanatory variables of the individuals. This quantity is the linear component of the model and it is also known as the risk score or prognostic index for the i^{th} individual. The dependence of the covariate variables is then aggregated into the scalar $\beta^T x_{p.}$ Consequently, the general proportional hazards model is as follow:

$$h_p(x_p, t) = e^{(\beta^T x_p)} h_0(t) \tag{11}$$

This model can be re-expressed in:

$$\log\left\{\frac{h_i(t)}{h_0(t)}\right\} = \beta^T x_p \tag{12}$$

The hazard function may depend on two types of variables namely variates and factors. A variate is a variable that takes numerical values that are often on a continuous scale of measurement, such as age. A factor is a variable that takes a limited set of values which are known as the levels of the factor, such as the variable gender. If we have a situation where the hazard function depends on two variables, X_1 and X_2 the proportional hazards for the i^{th} individual, can be written as:

$$h_i(t) = e^{(\beta_1 x_1 + \beta_2 x_2)} h_0(t) \tag{13}$$

Using this model, the logarithm of the hazard ratio is considered linear, due to:

$$\log\left\{\frac{h_i(t)}{h_0(t)}\right\} = \beta_1 x_1 + \beta_2 x_2 \tag{14}$$

Considering the ratio of the hazard of death for an individual with the value x+1 for X relative to one with value x, it obtains:

$$\frac{e^{\{\beta(x+1)\}}}{e^{\beta x}} = e^{\beta} \tag{15}$$

This ratio shows that when a variable with a single beta is included in the model, the hazard ratio when the value of X is changed by r units does not depend on the actual value of X. This means that the hazard ratio for an individual with value 60 for a variable, related to one with value 55 for the same variable, is the same for an individual with value 20 related to one with value 15 for the same variable. This is a result of fitting X as a linear term in the proportional hazards model.

Supposing that the dependence of the hazard function on a single factor A is to be modelled where A has *a* levels, the proportional hazards model for an individual with factor A at level j is $e^{(\alpha_j)}h_0(t)$. Here, the baseline hazard function has to be defined as the hazard for an individual with values of all explanatory variables equal to zero. Consequently, one of the α_j must be taken to be zero. If the constraint $\alpha_1=0$ is adopted, the term α_j can be included defining a-1 indicator variables, $X_2,X_3,\ldots.X_a$, which take the values shown below:

Level of A	**X2**	**X3**	**….**	**Xa**
1	0	0	….	0
2	1	0	….	0
3	0	1	….	0
….	….	….	….	0
a	0	0	….	1

For each ith individual it is only possible to have one variable for each factor, and the proportional hazards model can be written as:

$$h_i(t) = e^{(\beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_a x_a)} h_0(t) \quad \textbf{(16)}$$

Using this model, the logarithm of the hazard ratio is considered piece-wise linear, because each covariate has the number of betas equals to the number of levels that exist.

However, it is important to mention that equation 12 represent the Cox Regression Model in continuous time. When the model is in discrete time equation 12 is re-expressed as in the following equation:

$$\log\left(\frac{h(x)}{1-h(x)}\right) = \log\left(\frac{h_0(t)}{1-h_0(t)}\right) + \beta x \quad \textbf{(17)}$$

2.2 - Flexible Models

2.2.1. Generally of Artificial Neural Networks

An artificial neural network (ANN) is a mathematical model or computational model originally inspired on the central nervous system and neurons.

As described in (Bar-Yam, 1997), the basic computational unit in the nervous system is the nerve itself, or neuron. A neuron has dendrites, cell body and axon. A biological neuron receives input from other neurons. The input zone is composed by the dendrites. When the potential reaches a threshold, the cell fires and an action potential propagates along the axon (output) to other neurons, through an electrical signal. Transmission of an electrical signal from one neuron to the next is affected by neurotransmitters chemicals, which are released from the first neuron and bind to receptors in the second. This link is called a synapse.

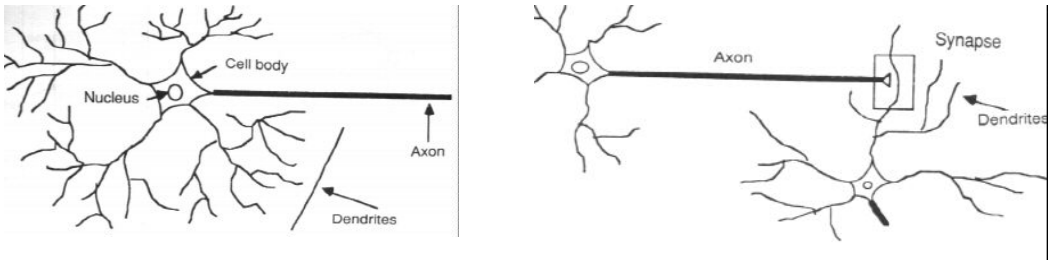

Figure 2.2 – Constitution of a neuron (Computation in the brain).

Brain learns by altering the strengths of connections between neurons, and by adding or deleting connections between neurons. Brain learns based on experience, which means that connexions between neurons are dynamic. Connexions highly used are strengthened and connexions less used tend to disappear. The same phenomenon happens in neural networks.

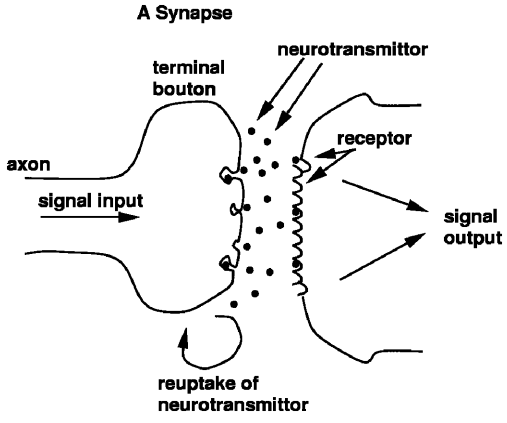

Figure 2.3 – Constitution of a synapse (Computation in the brain).

There are several properties of the nervous system that are of particular interest in the biologically inspired Neural Networks models, which are:

1. Parallel and distributed information processing
2. High degree of connectivity among basic units
3. Connections are modifiable based on experience
4. Learning is a constant process
5. Learning is based on local information
6. Performance degrades gracefully if some units are removed

The basic computational element (model neuron) is often called a node or unit. It receives input from other units or from an external source. Each input has an associated weight *w*, that is modified while the model learns. The "electrical" information is simulated using numerical values stored in these weights. The unit computes a function f of the weighted sum of its inputs. This resembles the perceptron model of Rosenblatt (1962), which is a linear discriminant model. In a simplistic neural network the summed value is compared with a certain threshold, in order to propagate the signal or not. However nowadays, instead of a threshold activation functions are used. Using the perceptron of Rosenblatt the nonlinear *f* function is given by the step function and the algorithm used to determine the parameters *w* of the perceptron is an error function known as the perceptron criterion.

$$y_i = f(\sum w_{ij} x_j) \tag{18}$$

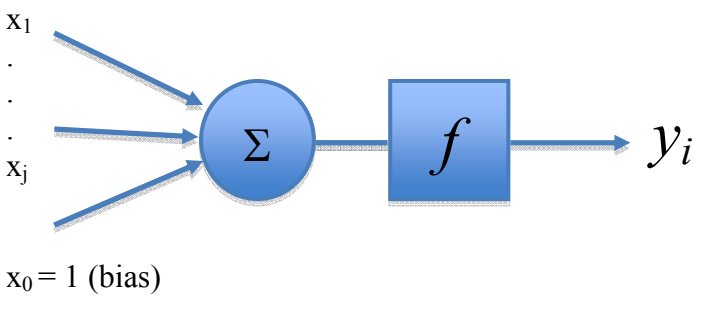

Figure 2.4 – Example of a perceptron.

The output can serve as the input to other units. The function *f(x)* is the unit's activation function. This activation function describes the output behaviour of a neuron. There are several activation functions that can be used. The choice of the activation function is determined by the nature of the data and the assumed distribution of target variables. Activation functions for the hidden units are needed to introduce nonlinearity into the network. Without nonlinearity, hidden units would not make nets more powerful than just plain perceptrons (which do not have any hidden units, just input and output units). The reason is that a linear function of linear functions is again a linear function. However, it is the nonlinearity (i.e., the capability to represent nonlinear functions) that makes multilayer networks so powerful. Almost any nonlinear function does the job, except for polynomials. Following are some of the commonly used activation functions:

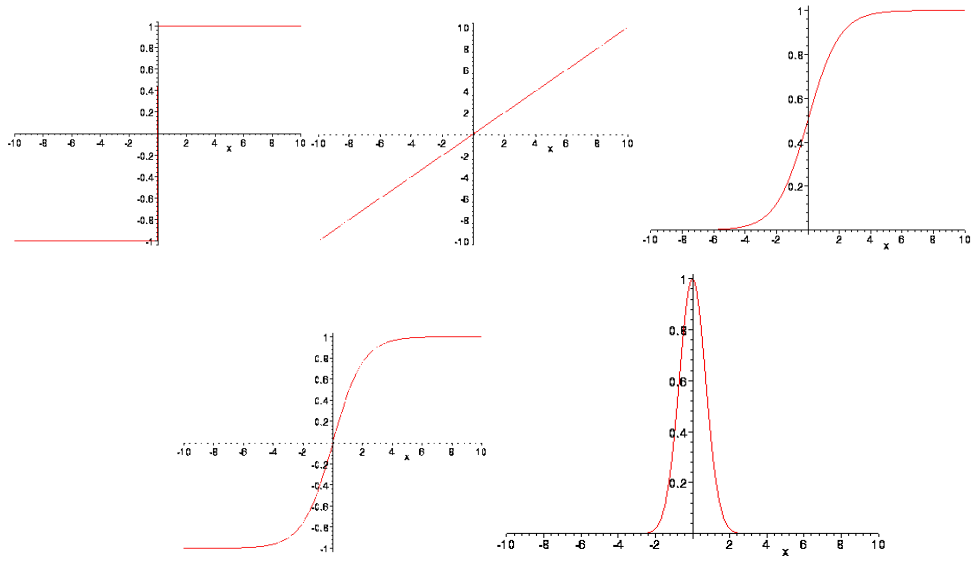

Figure 2.5 – Representation of possible neural networks activation functions.
The top pictures are from left to right, the step, identity and sigmoid function, respectively. The bottom pictures are the symmetric sigmoid function and the radial basis functiony.

If several perceptrons or weighted neurons are connected to each other, i.e., the output of a perceptron is the input of another perceptron, there will be a neural network model, or also known a multilayer perceptron or MLP. In a MLP the neurons are organized in layers, which number differs from network to network. The input nodes receive signals from "outside" the network and the output nodes send the signals "outside" the network. If there is a layer

constituted by several nodes, but their signals are not received from outside the network nor are sent to outside the network, then it is called hidden layer. A multilayer perceptron have several hidden layers, some input nodes and some output nodes, which is represented by the following figure:

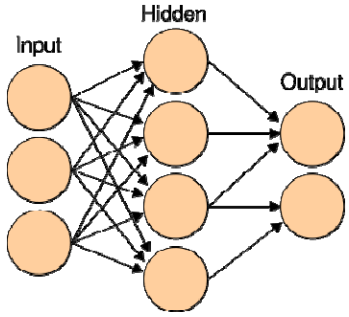

Figure 2.6 – Example of a multilayer perceptron.
It has 1 hidden layer, 3 input nodes and 2 output nodes

The weight parameters are represented by the links between the nodes and the arrows denote the direction of information that flows through the network during forward propagation. There was however introduced a new element, called "bias", which defines the neuron trend, subject to modifications during the network training. Bias units can also be weighed and connect an unitary input to each neuron.

The number of neurons on each layer is always equal to the number of variables. Although the number of hidden layers may differ, it is almost always equal to two or three. This is a result of a study done by Bishop (Bishop, 2006), where he shows that any network with two hidden layers can approximate any function independently of its complexity. One the one hand we should consider to use always a powerful network (three layers), on the other hand this network can create overfitting problems (there is a lost of a generalisation capacity).

Combining these various stages to give the overall network function, using sigmoidal output activation, the final network takes the form:

$$y_k(x,w) = \sigma\left(\sum_{j=1}^{M} w_{kj} h\left(\sum_{i=1}^{N} w_{ji} x_i + w_{j0}\right) + w_{k0}\right) \quad \textbf{(19)}$$

where

$$\sigma(a) = \frac{1}{1+e^{-a}} \quad \textbf{(20)}$$

There is however a key difference between these model and the perceptron model, that is the neural network uses continuous sigmoidal nonlinarities in the hidden unit, whereas the perceptron uses step function nonlinearities. This means that the neural network function is differentiable with respect to the network parameters, and this property will play a central role in network training.

2.2.2. Neural Network Training

Neural networks try to simulate the learning ability of the human brain. However, unlike the human brain, the neural network structure is fixed, not modifiable and constituted by a fixed number of neurons and connexions between them, which have some values (weights). What changes, on neural networks' learning process are the weights' values, increasing if the information is to be transported and decreasing otherwise. There is no indication of what should be the weights values in the beginning of the network training, so they are initialized randomly. Then these values are adjusted after processed one individual or at the end of all individuals processing.

Training a neural network essentially means selecting one model from the set of allowed models, or in a Bayesian framework determining a distribution over the set of allowed models that minimizes the cost criterion. There are numerous algorithms available for training neural network models; most of them can be viewed as a straightforward application of optimization theory (weights adjustment in order to minimize the error) and statistical estimation. In the following picture it can be observed the purpose of network training, that is through the adjustments done in weightings, minimize the error produced by the network.

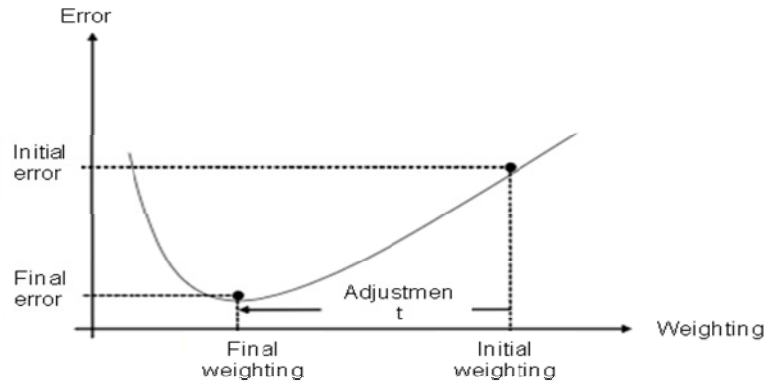

Figure 2.7 – Neural network weighting versus error

Most of the algorithms used in training artificial neural networks employ some form of gradient descent. This is done by simply taking the derivative of the cost function with respect to the network parameters and then changing those parameters in a gradient-related direction. With this method, the adjustment can be calculated at each point in order to minimize the network error function, that is given a training set comprising a set of input vectors $\{x_n\}$,where n= 1…N, together with a corresponding set of target vectors $\{t_n\}$, the error function *E(w)* must be minimized:

$$E(w) = \frac{1}{2}\sum_{n=1}^{N} \|y(x_n, w) - t_n\|^2 \qquad \textbf{(21)}$$

The value of *w* found by minimizing this function corresponds to the maximum likelihood solution. As the function error *E(w)* is a smooth continuous function of *w*, its smallest value will occur at a point in weight space such that the gradient of the error function is equal to 0. However it is difficult to find the solution to the previous equation, iterative numerical procedures are used to find a solution. Moreover, the use of the gradient information can lead to significant improvements in the speed with which the minima of the error function can be located. The simplest approach to using gradient information is to choose the weight update to comprise a small step in the direction of the negative gradient, so that:

$$w^{(\tau+1)} = w^{(\tau)} - \eta \nabla E(w^{(\tau)}) \qquad \textbf{(22)}$$

where the parameter η is the learning rate. After each update the gradient is re-evaluated for the new weight vector and the process is repeated. At each step the weight vector is moved in the direction of the greatest rate of decrease of the error function, and so this technique is known as gradient descent.

In order to find a good minimum it may be necessary to run a gradient-based algorithm multiple times, each time using a different randomly chosen starting point and comparing the resulting performance on an independent validation data set.

If the gradient descent technique is used to train a multi-layer network, there will be a difficulty that is the absence of the target value for the hidden units. Therefore it was found an efficient technique for evaluating the gradient of an error function for a feed-forward neural network, which is the **backpropagation algorithm**. The name of this algorithm is based on the idea of backpropagate the error obtained in the output of the network.

The backpropagation learning process works in small iterative steps:

1. The example cases are applied to the network producing some output based on the current state of its synaptic weights (initially, the output will be random).
2. The output is compared to the desired output, and a mean-squared error signal is calculated.
3. The error value is then propagated backwards through the network, and small changes are made to the weights in each layer. The weight changes are calculated to reduce the error signal for the case under study.
4. The whole process is repeated for each example in the training set, then back to the first case again.
5. The cycle is repeated until the overall error value drops below some pre-defined threshold.

At this point we say that the network has learned the problem. It is important to refer that the network will never exactly learn the ideal function, but rather it will asymptotically approach it.

For backpropagation learning, the activation function must be differentiable, and it helps if the function is bounded; the sigmoidal functions (such as logistic and tanh) and the Gaussian function are the most common choices. Functions such as tanh or arctan that produce both positive and negative values tend to yield faster training than functions that produce only positive values such as logistic, because of better numerical conditioning.

For hidden units, sigmoid activation functions are usually preferable to threshold activation functions. Networks with threshold units are difficult to train because the error function is stepwise constant, hence the gradient either does not exist or is zero, making it impossible to use backpropagation or more efficient gradient-based training methods. With sigmoid units, a small change in the weights usually produces a change in the outputs, making possible to tell whether the change in the weights is good or bad. With threshold units, a small change in the weights will often produce no change in the outputs.

For the output units, the activation function should be chosen to suit the distribution of the target values:

- For binary (0/1) targets, the logistic function is an excellent choice (Jordan, 1995).
- For categorical targets using 1-of-C coding, the softmax activation function is the logical extension of the logistic function.
- For continuous-valued targets with a bounded range, the logistic and tanh

functions can be used, provided either scale the outputs to the range of the targets or scale the targets to the range of the output activation function ("scaling" means multiplying by and adding appropriate constants).

- If the target values are positive but have no known upper bound, the exponential output activation function can be used.
- For continuous-valued targets with no known bounds, the identity or "linear" activation function can be used.

Multilayer networks can approximate any smooth function as long as there are enough hidden nodes. However, having this great flexibility can cause the network to learn the noise in the data and be over-trained or over-fitted. There are several ways to control the complexity to avoid this over-fitting. One way is to add a regularization term to the error function, also known as *weight decay*, giving a regularized error of the form:

$$\tilde{E}(w) = E(w) + \frac{\lambda}{2} w^T w \qquad \textbf{(23)}$$

An alternative to regularization as a way of controlling the effective complexity of a network is the *early stopping* procedure. The training of nonlinear network models corresponds to an iterative reduction of the error function defined with respect to a set of training data. However, the error measured with respect to independent data often shows a decrease at first followed by an increase when the network starts to over-fit. Training can therefore be stopped at the point of smallest error with respect to validation data set.

2.2.3. ANN application in prognostic modelling

Artificial neural networks are non-linear, semi-parametric models that have been considered as alternative methods for prognostic models in the presence of censorship (Lisboa, 2002). The most common applications of neural networks have been for diagnosis or prognosis, that is to decide which class k, $k \in \{0,\ldots,K\}$, in terms of prognostic risk, an individual belongs by using information on a set of *p* covariate values $x=(x_1,\ldots x_p)$. The usual aim is to construct a decision rule for individuals with known covariate values but unknown class level. The usual approaches to construct such decision rules are based on estimating the

conditional probability of observing an individual with a certain class level given the covariates:

$$p(Y = class \quad level \mid X = x) = f(x, \beta) \tag{24}$$

β is the vector of unknown parameters, which are called "regression coefficients" in statistics and weights in neural networks modelling. In neural networks modelling, the input layer corresponds to the covariates. The hidden units are the result of applying the activation function to a weighted sum of the input units plus a constant (w_0). The value of a hidden unit h_j is given by:

$$h_j = \phi(w_{0j} + \sum_{i=1}^{p} w_{i_j} x_i) \tag{25}$$

where ϕ is the activation function, w_{0j} the bias.

The value of that output unit y is calculated by applying another activation function, as follows:

$$y = f(x, w) = \phi\left(W_0 + \sum_{j=1}^{r} W_j \phi\left(w_{0j} + \sum_{i=1}^{p} w_{i_j} x_i \right) \right) \tag{26}$$

There is a large literature on the use of neural networks for other kinds of classification tasks have been published in the area of medicine (Lisboa, 2002). The application of feed-forward neural networks to survival data has been discussed in the past years and it is an extension of the previous equation. Here, the output y_K corresponds to the conditional probability of dying in the k^{th} time interval I_k. Data for the n^{th} individual consists of a vector of covariate variables $y=(y_1,...,y_{kn})$ and a vector which indicates the interval I where the individual has died. Thus $y_1,...,y_{Kn-1}$ are all zero and y_{Kn} is equal to 1 if the individual died in I_{kn} and equal to zero if the individual was censored. This implies that the network has a randomly varying number of output nodes according to those time intervals where the individual is at risk.

Some studies have showed that by treating the time interval as an input variable in a standard feed forward network with logistic activation and entropy error function, it was possible to estimate smoothed discrete hazards as conditional probabilities of failure. This proposed artificial neural network (ANN) approach can be applied to the estimation of the functional relationships between covariates and time in survival data to improve model

predictivity in the presence of complex prognostic relationships (Biganzoli, Boracchi, Mariani, Marubini, 1998).

There have been other proposals for analysing survival data using feed-forward neural networks, as using the network with only one output unit using the number k of the time interval as additional input and consider the unconditional survival probability of dying before t_k rather than the conditional as output (Ravdin, Clark, 1992), (Ravdin, Clark, Hilsenbeck, Owens, Vendely, Pandian, McGuire, 1992). Here, time was entered as a predictor, and each patient had as many entries in the model as the number of intervals during which it was alive. The intervals were derived from Kaplan-Meier estimates. It was introduced some bias, due to the introduction of coding time as covariate. This work was one of the first studies, which addressed the use of neural networks for survival analysis using real clinical data, producing accurate estimations for survival of breast cancer patients and raising the important issue on how to deal with censored data in neural network implementations for survival analysis. Another form of neural networks that can be applied to survival data is the called "single time-point models". Here a single time point t is fixed and the network is trained to predict the t year survival probabilities. This approach is used by (McGuire, Tandon, Allred, Chamness, Ravdin, Clark, 1992), (Kappen, Neijt, 1993), (Burke, 1994). The common drawback of these approaches is that they do not allow incorporating censored observations. Neither omission of the censored observations nor treating censored observations as uncensored is a valid approach.

Other approach to use neural networks in survival analysis has been the use of hierarchical neural networks, which predict the survival in a stepwise manner. This approach predicts for the first time interval, than for the second interval and so on. The system produces a survival estimate for patients at each interval, given relevant covariates and it is able to handle continuous and discrete variables, as well as censored data. They can predict absolute, cumulative survival as well as instantaneous, conditional survival. Ohno-Machado (Ohno-Machado, Walker, Musen, 1995) compared three neural network models for survival analysis and for AIDS patients. He concluded not only that the hierarchical neural-networks models for survival analysis could learn infrequent patterns faster than could a non-hierarchical model, but also that they provide better accuracy in predicting death for the used cohort. However, this can lead to inconsistent answers such as give a higher predicted probability for

death in year 1 or 2 than for deaths in years 1,2 or 3.Other approach is to model conditional probabilities:

$$p(\text{die in ith interval} \mid \text{survive first } i - 1 \text{ intervals}, x) = g(\eta_i) \quad \textbf{(27)}$$

where g is usually the logistic function. The patient dying in the i^{th} interval contributes with $log(g(\eta_i)(1-g(\eta_i-1))\ldots(1- g(\eta_1)))$ to the likelihood, and a patient lost to follow up interval $log((1-g(\eta_i-1))\ldots (1- g(\eta_1))$. The scores $\eta_1, \ldots, \eta_k$ are given by the output of a neural network with k linear outputs.

Ripley and Ripley (Ripley, Ripley 1998) tried several methods to compare neural networks and linear methods to classify binary outcomes, 1year period proportional odds, regression, proportional hazards, Weibull survival and log-logistic survival. They obtained a neural network with a specificity, sensitivity and accuracy higher for almost the methods than linear methods.

Delen et al (Delen, Walker, Kadam, 2005) reported a research where they developed several prediction models for breast cancer survivability, especially ANN. They used a binary categorical survival variable where survival is represented with value of "1" and non-survival is represented with "0". They measured their accuracy, which was very good, and compared with other methods. The conclusion was that ANN are better than linear methods, such as logistic regression.

2.2.4. Misuses in Applications of ANN for prognostic models

Schumacher et al (Schwarzer, Vach, Schumacher, 2000) concluded that most applications of artificial neural networks for prognostic and diagnostic models suffer from methodological deficiencies, such as:

1. Biased or inefficient estimation or misclassification due to inappropriate splitting of the data set.
2. The size of the test set is usually very small, leading to an inefficient estimation.
3. Insufficient number of events for the number of variables used.
4. Regularization terms to avoid over-fitting (tune the neural network to the peculiarities of the examples rather than to extract the salient dependencies of the whole population) are rarely used.

5. The ANN performance must be compared with adequate statistical competitors, like nearest neighbour, CART, generalized additive models or logistic regression and performance must be well calculated with relevant measures.
6. Some applications of ANN did not guarantee monotonicity of the estimated survival curves.
7. Some applications of ANN used the number of the time interval as additional input unit.
8. Naïve application of ANN models due to inappropriate handling of censoring.

2.2.5. Advantages of using Neural Networks in Prognostic Modelling

Previous studies (Neal, 2001), (Lisboa, Wong, Harris, Swindell, 2003), (Lisboa, 2002), (Taktak, Antolini, Aung, Boracchi, Campbell, Damato, Ifeachor, Lama, Lisboa, Setzkorn, Stalbovskaya, Biganzoli, 2007) have found several advantages of using artificial neural networks in prognostic modelling, which gives confidence in their use, such as:

1. ANN are capable of modelling extremely complex non-linear functions.
2. ANN overcomes the limitation of proportional hazards modelling assumption that the time development of the hazards is proportional to a fixed baseline population.
3. ANN overcomes the limitation of proportional hazards modelling assumption that the covariates influence the model through explicit linear terms.
4. The overfitting overcoming of ANN can be avoided using Bayesian methods.
5. ANN do not rely on the availability of prior knowledge.

2.2.6. Bayesian Regularisation framework

In Neural network training section, the learning framework was focused on the use of maximum likelihood. In a Bayesian framework the network must be marginalized over the distribution of parameters in order to make predictions. The aim is to construct a probability distribution over the possible values of the parameters in the network. There are several advantages of Bayesian learning over maximum likelihood methods (Neal, 2001), (Antolini, Boracchi, Biganzoli, 2005), (Bishop, 2006), (Mackay, 1992):

1. Model overfitting is unlikely in contrast with the maximum likelihood approach that is prone to overfitting when the number of network parameters is large in relation to the number of training cases (Biganzoli, Boracchi, Marubini, 2002).

2. The model is automatically regularized.
3. The uncertainty in predictions can be obtained.

In the conventional maximum-likelihood approach, a single weight vector is found, which minimizes an error function. In contrast, the Bayesian framework considers a probability distribution over the network weights, which is described by a prior distribution *p(w)* and it is modified when the data $D=\{(x,t)\}$ is observed. The process can be expressed by the Bayes theorem, which aim is to approximate the posterior distribution by a Gaussian distribution:

$$p(w \mid D,\alpha,H) = \frac{p(D \mid w,\alpha,H)\,p(w \mid \alpha,H)}{p(D \mid \alpha,H)} \tag{28}$$

where w denotes the set of weights, D the data set, α the penalty parameters and H the model hypothesis. Using Bayes formula, the prior and the data likelihood can be transformed into a posterior distribution.

To evaluate the previous equation it is necessary to find the expressions for the prior $p(w|\alpha,H)$ and for the likelihood $p(D|w,\alpha,H)$. The prior over the weights reflect the knowledge we have about the network we want to build, if there is any. The function must have a multivariate normal density with zero mean and diagonal covariance matrix with elements $1/\alpha_K$. The weights that are centered in zero have higher probability:

$$p(w \mid \alpha) = N(w \mid 0,\alpha^{-1}I) \tag{29}$$

$$p(w \mid \alpha,H) = \frac{e^{-E(w,\alpha)}}{Z_w(\alpha)} \tag{30}$$

where

$$E(w,\alpha) = \frac{1}{2}\sum_{m=1}^{N_\alpha}\alpha_m \sum_{n=1}^{N_m} w_{mn}^2 \tag{31}$$

α (inverse variances) are the regularization parameters, called hyper-parameters, because they control the distribution of the network parameters (Neal, 2001). These hyper-parameters should be different for each input, as some inputs have more influence than others. Hidden layers must look at different subsets of the inputs and the hyper-parameters must control how much each layer contributes to the function (Bakker, Heskes, 1999). $Zw(\alpha)$ is a normalisation

constant calculated from a product of univariate normal distribution, also called the evidence. The index n indicates a group of weights w_{mn} sharing a common regularization parameter α_m of which there are N_m. These weights correspond to attributes from a single field or variable. As the training progresses, the α_m for variables with little predictive power increase in size, forcing the corresponding weights towards zero, adjusting the effect of each covariate according to its relevance to the model protecting against overfitting of the data. The term "weight decay" is commonly used for this regularization method.

The likelihood given the weights can be obtained for discrete or continuous time cases by using the hazard function. For the continuous time case:

$$L_c = \sum_{p=1}^{cardinality\ of\ the\ set} \sum_i d_{pi} \log h_c(x_p, t_i) - \int_0^{t_i} h_c(x_p, u)du \qquad \textbf{(32)}$$

For discrete time case, the log-likelihood is:

$$G = -\sum_{p=1}^{sample\ size} \sum_{k=1}^{t_1} \left[d_{pk} \log\left(h_p(x_p, t_k)\right) + (1 - d_{pk}) \log\left(1 - h_p(x_p, t_k)\right)\right] \qquad \textbf{(33)}$$

In the latter case the likelihood term reflects the status of the patient p at a time t_k. This is achieved with an indicator label or target d_{pk} which assumes the value of 0 if the patient is observed to be alive and 1 to indicate a death attributed to breast cancer in the time interval t_k. $t_1 < t \leq t_k$. t_1 is the month when the patient was last observed.

Previous research (Eleuteri, Aung, Taktak, Damato, Lisboa, 2007) has been done in order to compare both models, in the continuous and discrete time. The first approximates the logarithm of the hazard rate function, and the other models the log-odds ratio of the hazard probability. It was concluded that both models exhibit good discrimination and calibration capabilities, although the continuous time model has the advantage of providing error bars on its estimates. From now on it will be used the discrete-time model. With this $p(D \mid w, \alpha, H) = e^{L_p}$, where L_p is the log-likelihood function.

Once the expressions for the prior and the noise model are given, the posterior distribution can be evaluated as:

$$p(w \mid D,\alpha,H) = \frac{e^{-G}e^{-E(w,\alpha)}}{Z(\alpha)} \quad \textbf{(34)}$$

The posterior distribution is usually considerably complex and multimodal. A solution is to integrate out the parameters separately from the hyper-parameters by making a Gaussian approximation for the mode with respect to the hyper-parameters. (Bishop, 2006) and (Mackay, 1992) mentioned that this procedure gives a good estimation of the posterior probability, particularly for distributions over high-dimensional spaces. The posterior and evidence are maximized until a consistent solution $\{w_{MP}, \alpha_K\}$ is found.

Generally hyper-parameters can be adjusted with empirical measures such as cross-validation. However, this approach is computationally very expensive and may not be robust. A solution is to use the Bayesian framework:

$$p(\alpha \mid D,H) = \frac{p(D \mid \alpha,H)p(\alpha \mid H)}{p(D \mid H)} \propto \frac{e\left(-S(w^{MP},\alpha)\right)}{Z_w(\alpha)}(2\pi)^{N_w/2}\det(A)^{-1/2} \quad \textbf{(35)}$$

where

$$S(w,\alpha) = G + E(w,\alpha) \quad \textbf{(36)}$$

A is the Hessian of S with respect to the weights. The posterior of the hyper-parameters can be calculated analytical approximating the evidence $p(D|\alpha,H)$ with a Taylor expansion about the current value of the weights, which are known as the "most probable" values w^{MP}.

2.2.6.1 Marginalization of the network predictions

In survival modelling the distribution of the target is extremely skewed, due to the scarcity of the events and the large number of time steps used in the analysis. Therefore, there is usually a highly unbalanced distribution of the target values. This conflicts with the implicit prevalence of the output of the network, where high levels of uncertainty shifts the network output to 0.5. Using the bayesian regularization framework, it is assumed that the weight values have a posterior distribution and the predicted hazard must be marginalised over this distribution, which is parameterised by a normal density function for the output node activation:

$$a \sim N(a_{MP}, s^2) \sim N(a_{MP}, g^t A^{-1} g) \quad \textbf{(37)}$$

where a_{MP} is the most probable output and g is the activation function of the output node.

The predicted hazard is the mean calculated from the distribution of the activation function:

$$h_g(x,t) = \int g(a)P(a \mid x_p,t,D)da = \int\left(\frac{1}{1+e^{-a}}\right)\frac{e^{-a(-a_{MP})^2/2s^2}}{\sqrt{2\pi}s} \tag{38}$$

where

$$g(a) = \frac{1}{1+e^{-a}} \tag{39}$$

However, this integral is not analytical, but it can be approximated using:

$$h_g(x,t) \approx g\left(\frac{a_{MP}(x,t)}{\sqrt{1+(\pi/8)g^t A^{-1} g}}\right) \tag{40}$$

With the implicit prevalence of the previous equation, high levels of uncertainty, characterized by high values of s^2, as it was noticed previously, the network output $h_g(\cdot) \to_{s\to\infty} 1/2$. To overcome this situation it is necessary to update the standard regularisation framework to take into account the prevalence of the targets (P_τ) by re-scaling the log-likelihood:

$$G = -\sum_{p=1}^{sample}\sum_{k=1}^{size\ t_1}\left[\frac{1}{2P_\tau}\tau_{pk}\log\left(h_p(x_p,t_k)\right) + \frac{1}{2(1-P_\tau)}(1-\tau_{pk})\log\left(1-h_p(x_p,t_k)\right)\right] \tag{41}$$

The marginalised network prediction $\tilde{h}_g(x_p,t)$ needs to be compensated so the maximum uncertainty in the network predictions is the empirical estimate of the prevalence:

$$h_g(x_p,t) = \frac{\tilde{h}_g(x_p,t)P_\tau}{\tilde{h}_g(x_p,t)P_\tau + (1-\tilde{h}_g(x_p,t))(1-P_\tau)} \tag{42}$$

2.2.7. PLANN-ARD in prognostic modelling

ANN models are considered alternative methods for survival analysis in the presence of censorship. These models are an extension of the discrete time proportional hazards. The partial logistic artificial neural network with automatic relevance determination (PLANN-ARD) (Lisboa, Wong, Harris, Swindell, 2003) is a development of the PLANN method proposed by Biganzoli et al. (Biganzoli, Boracchi, Marubini, 2002), (Biganzoli, Boracchi, Mariani, Marubini, 1998). The network structure is represented as follows:

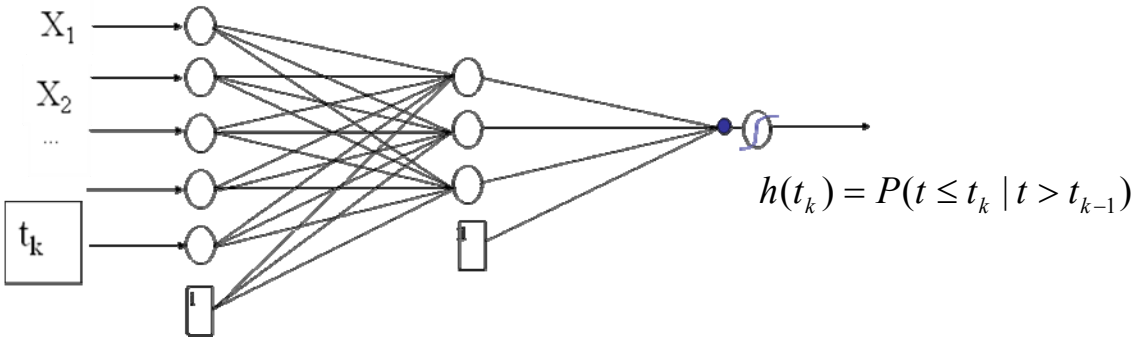

Figure 2.8 – Partial logistic artificial neural network structure.

In figure 2.8, X_1, $X_{2...}$ are the covariates and t_K is the time value observed. The outcome observed corresponds to the event indicator at each time k, that is, if the event occurs at a certain time t, then the outcome is 1, otherwise is 0. The covariates are replicated for each observation period, alongside a monotonically increasing interval measure of the discrete time interval, until the event indicator is observed. The output of the network is the hazard at each time t. As an example, for an individual sample, if the event occur in t=4 and not previously, the covariates for this sample are imputed to the network 4 times, where the time value changes. The event indicator is 0 for t=1,2 and 1 for t=4.

Attributes	Time value	Event indicator
$X_1 \ldots X_j$	0,5	0
$X_1 \ldots X_j$	1,5	0
$X_1 \ldots X_j$	2,5	0
$X_1 \ldots X_j$	3,5	1

Table 2.2 – Example of s inputs and outputs for the PLANN-ARD model.

The PLANN-ARD method uses the MLP structure with time as input, where the discrete time implementation is strictly a proportional model for the odds-ratio. The associated logistic link (activation) function makes it natural to extend the feedforward model by adding flexibility in the form of a multi-layer perceptron neural network. As result the analytical expression for the network becomes:

$$\frac{h_p(x_p,t_k)}{1-h_p(x_p,t_k)} = \exp\left(\sum_{h=1}^{N_h} w_h g\left(\sum_{i=1}^{N_i} w_{ih}x_{pi} + wt_k + b_h\right) + b\right) \tag{43}$$

where i and h denote the input and hidden node layers respectively, b_h is the bias (intercept) term in the hidden layer, b is the bias (intercept) term in the output layer and the non-linear function $g(.)$ is the sigmoidal function at the hidden nodes:

$$g(a) = \frac{1}{1+e^{-a}} \tag{44}$$

In addition, the model must take into account the loss to follow-up, or right censorship. This is reflected, under the assumption of independent censoring, in the form of the objective function, which is the log-likelihood summed over the observed status of the patient with a binary indicator when the patient status is observed alive or has died, using target values τ_{pk} as indicator labels and t_l as the time index:

$$G = -\sum^{No.\ patients}\sum^{t_1}\left[\tau_{pk}\log\left(h_p(x_p,t_k)\right) + (1-\tau_{pk})\log\left(1-h_p(x_p,t_k)\right)\right] \tag{45}$$

If the sigmoidal function was replaced by a linear function the argument of the exponential would be $\beta^T x_p + \theta_1 t_k + \theta_2$. This represents the factorisation of the dependence of the discrete time hazard on the explanatory variables and time. The previous function can be compared with proportional hazards modelling because, for discrete time intervals, it parameterises the odds of survival, as follows:

$$\frac{h(x_i,t_k)}{1-h(x_i,t_k)} = \frac{h_0(t_k)}{1-h_0(t_k)} e^{\beta^T x_p} \tag{46}$$

$$h(x_i,t_k)=\frac{h_0(t_k)e^{\beta^T x_p}}{h_0(t_k)(e^{\beta^T x_p}-1)+1}=\frac{1}{1+\frac{1-h_0(t_k)}{h_0(t_k)}e^{-\beta^T x_p}}=\frac{1}{1+e^{-\beta^T x_p-\beta_0}} \quad \textbf{(47)}$$

where

$$\beta_0=\log\left(\frac{h_0(t_k)}{1-h_0(t_k)}\right) \quad \textbf{(48)}$$

As $-\beta^T x_p-\beta_0$ corresponds to a summed output of the hidden nodes multiplied by the weights $\sum_{h=1}^{N_h} w_h g\left(\sum_{i=1}^{N_i} w_{ih}x_{pi}+wt_k+b_h\right)+b$, the hazard will be calculated applying the sigmoidal function to this value.

Once the hazard estimate is available, the probability of the event occurring up to a time threshold, that is to say, the survival probability can be directly estimated for discrete data by the successive products of the conditional estimates of survival in each individual time interval (Marubini, Valsecchi, 1995).

$$S_p(x_p,t_k)=P(t\le t_k\mid x_p)=\prod_{l=1}^{t_k}\left(1-h(x_p,t_l)\right) \quad \textbf{(49)}$$

Once the modelling process is complete it is necessary to define a prognostic index for each patient (as it can be done using Cox proportional hazards modelling). This index identifies the patients with a lower and higher mortality risk. Equivalent to the linear prognostic index βx, previous studies has showed that the prognostic index obtained with PLANN-ARD was:

$$\text{Prognostic index}(x_p)=\frac{\sum_{i=1}^{T}\left(\frac{h_p(x_p,t_k)}{1-h_p(x_p,t_k)}\right)}{T} \quad \textbf{(50)}$$

where T is the number of time intervals (Lisboa, Wong, Harris, Swindell, 2003). An improvement of this prognostic index was done in (Fernandes, Jarman, Etchells, Fonseca, Biganzoli, Bajdik, Lisboa, 2008), where better results were obtained:

$$\text{Prognostic index}(x_p)=\frac{1}{T}\sum_{i=1}^{T}S(t_{i-1})\ln\left(\frac{h_p(x_p,t_k)}{1-h_p(x_p,t_k)}\right) \quad \textbf{(51)}$$

As this index must be as accurate as possible, there have been other studies that justify the improvement of the previous index, too. These studies were based on using competing risk methodologies. Although competing risks has not been mentioned in PLANN-ARD modelling, the prognostic index to use must be the same, in single risk and competing risk modelling. Therefore, the prognostic index proposed is:

$$\text{Prognostic index } (x_p) = \ln(-\ln(1 - CCI)) = \ln(-\ln(S(t))) \quad \textbf{(52)}$$

where the CCI is the crude cumulative incidence, identified as the probability of the occurrence of a specific event of interest (Ambrogi, Biganzoli, Boracchi, 2008) and *S(t)* is the estimated survival at the end of the follow up period, i.e. 60 in this study.

This flexible model accounts implicitly for non-linear and non proportional covariate effects. The neural network does not seek merely to explain the observed variation in survival, as a function of covariate effects. Instead, it fits the hazard function directly, without resort to proportionality assumptions about the covariate effects. In this way, it is suited to making individual predictions of the event rate. Automatic Relevance Determination (ARD) has the effect of suppressing covariate effects that are least informative about the predicted outcome. An essential feature of flexible models is the requirement to control model complexity in order to prevent over-fitting to the training data.

This was done using the well-known Bayesian framework of Automatic Relevance Determination (MacKay, 1995), where a separate weight–decay index is allocated to each covariate, taking care to apply the same value of the regularisation hyper-parameter to all of its binary coded attributes. A Laplace approximation of the evidence is calculated in the usual way (MacKay, 1995), to allow an iterative estimate of the most likely value of each weight decay to be obtained.

The application of this principled regularisation framework brings two advantages to the model. First, it enables a set of hyper-parameter values to be optimised efficiently. Second, this framework explicitly models the activation of the output nodes in a probabilistic way, which forms a second, natural extension, of linear models – this time in respect to the estimates of uncertainty, represented by the variance of the model predictions, from which confidence intervals can be obtained. This extension defines the Partial Logistic Artificial Neural Network regularised with Automatic Relevance Determination (PLANN-ARD) (Lisboa, 2002). This has the important advantage of automatically adjusting the effect of each

covariate according to its relevance to the model, protecting against over-fitting of the data without requiring hard model selection.

2.2.7.1 PLANN-ARD evaluation

A. Taktak et al. (Taktak, Antolini, Aung, Boracchi, Campbell, Damato, Ifeachor, Lama, Lisboa, Setzkorn, Stalbovskaya, Biganzoli, 2007) provided a double-blind evaluation and benchmarking of the accuracy in out-of-sample prediction of mortality from two generic non-linear models, using artificial neural networks (Partial logistic neural networks model with auto-relevance determination and Partial logistic basis function networks) against a partial logistic spline, log-normal and Cox-regression model. In this study, it was concluded that the recent and flexible modelling algorithms show a comparative predictive performance to that of more established methods from the medical and biological literature.

Moreover it has showed that PLANN-ARD obtained overall the best calibration performance. PLSPL, LOGN and PLANN-ARD showed similar performance on both model and test data set. P.J.G. Lisboa et al. (Lisboa, Etchells, Jarman, Aung, Chabaud, Bachelot, Perol, Gargi, Bourdès, Bonnevay, Négrier, 2008) has also showed that PLANN-ARD seemed to generalise better than Cox model, and appear to be more specific to identify patients at the extremes of high and low risk.

2.2.7.2 Individual prognostic predictions with confidence intervals

There has been a very high interest in predictive inference of prognosis for individual patients. The individual prediction of survival with confidence intervals can be obtained using either PLANN-ARD or Monte Carlo methods (Jarman et al, 2008).

As described previously PLANN-ARD model provides a prediction of smooth estimates of the discrete time hazard. At time t_i the estimated summed weights to each output unit is approximated by a Gaussian distribution $N(a_i, \sigma^2_i)$.

The individual prognosis for a patient x is calculated by first taking a random sample $\tilde{a}_i$ from $N(a_i, \sigma^2_i)$, calculating $\tilde{h}_i = g(\tilde{a}_i)$ and finally estimating survival $\tilde{S}(t_k)$. Regarding imputation, $\tilde{h}$ must be computed for each trained network, applying the following equation:

$$h(x_p, t_l) = \int h(x_p, t_l \mid \mu) P(\mu) d\mu \sim \frac{1}{T} \sum_{\mu_i}^{10} h(x_p, t_l \mid \mu_i) \qquad \textbf{(53)}$$

and the survival estimate may then be obtained.

These steps are repeated *n* times until there are enough survival points for each patient in order to build a distribution. The personalised prognosis with 95% posterior density intervals is the mean survival with 95% intervals determined by omitting the upper and lower 2.5% of the sample estimates.

To better represent the individual prognostic predictions with its confidence intervals, the pseudocode is as follows, for an individual patient with covariate set 'x':

i. Sample the output node activation in the PLANN-ARD model, a~N(a_most_probable(x,t), variance) ~ $N(a_i, \sigma^2_i)$ for each trained network.
ii. Propagate this value of 'a' through the sigmoid function to obtain a sample of *h(x,t)* for each trained network.
iii. Obtain the final *h_final(x,t),* being the average of the obtained samples of *h(x,t)* for each network.
iv. Repeat this procedure over time and calculate *S(x,t)=Product_t (1-h_final(x,t))*.
v. Repeat the complete procedure 1000 (necessary to stabilise the distribution of *S(x,t)*).
vi. Sample the mean and confidence intervals from this distribution.

2.3 - Prognostic index stratification and Boolean Rules extraction methodology

In clinical environment, the comparison of two survival distributions is frequently used in the evaluation of treatments or on the impact of prognostic factors on survival, especially to stratify patients by risk. As an example, Boracchi et al. (Boracchi, Coradini, Antolini, Oriana, Dittadi, Gion, Daidone, Biganzoli, 2008) presents a case study concluding that reliable outcome prediction is necessary for treatment allocation, exploiting the predictive potential of consolidated clinical and biological variables.

There are a variety of parametric and non-parametric methods for comparing distributions in the complete data, but there are fewer options for comparing two survival distributions in the presence of censored data. For this kind of data, the most widely used test is the log-rank test from which the statistical significance for pairwise data partitions can be measured. Given that, the test only applies in a pairwise manner, that is to say, for separating two cohorts at a time. This requires a search for the most appropriate threshold to divide the distribution of prognostic index scores. However, in the literature the approach to splitting risk indices into

risk groups is not always stated clearly, sometimes stating the cut-off points of the respective risk scores without a clear indication of how these were obtained (Guerra, Algorta, Diaz de Otazu, Pelayo, Farina, 2003), (Sebastian, Gonzalez, Paricio, Perez, Flores, Madrona, Romero, Tebar, 2000). Where the split of the indices is at all explained, expert knowledge has been a factor as in the case for the widely used NPI.

In another approach the indices are split into equal sized groups as suggested by Harrell et al. (Harrell Jr., Lee, Mark, 1996). This tutorial in biostatistics suggests using deciles as a starting choice and in a prognostic model for ovarian cancer Clark et al. (Clark, Stewart, Altman, Gabra, Smyth, 2001) used quartiles to partition the risk score. A suggestion for an automated method is to use successive top-down splits by maximising the log-rank test statistic (Williams, Mandrekar, Mandrekar, Cha, Furth, 2006). This approach can be called as "minimum P-value approach". However, the optimization of these cut-points results in an overestimation of the relative risk between the two prognostic groups. Usually, there is instability of the p-value in the minimum p-value approach, as there are some cut-points that can be considered minimum, because their p-values are all significant, so choosing the optimal cut-point is not the best approach (Altman, Lausan, Sauerbrei, Schumacher, 1994). Moreover, it should be kept in mind that the cut-point obtained is highly data dependent and it would be expected that this value vary markedly between different data sets, which is not the main goal for the model validation.

In order to improve the existing methodologies, a new one is proposed to make the stratification of risk indices more robust. This technique based on bootstrapping re-sampling technique can be used applying it to the prognostic indices, as these are calculated from the original data set.

Bootstrapping is a re-sampling method, which has a computer-intensive approach. This method is a general approach to statistical inference based on building a sampling distribution for a statistic by re-sampling from the original data.

There are some studies that use bootstrapping methods (Heller, Venkatraman, 1996) to compare four classes of test procedures and two survival distributions. Two of these tests use bootstrap methods, and they recommend using them when the log-rank test statistic is employed. M. Schumacher et al (Schumacher, Hollander, Sauerbrei, 1997) also has explored at what extended the result bias of the cut-points can be reduced using bootstrap re-sampling and cross-validation techniques, applied to a model with one factor, where an optimal cut-off

value in this factor is selected to define a low-risk group and a high-risk group. This single factor can be a prognostic index, which is a weighted sum of several covariates. Their results have shown that the bootstrapping approach they studied is capable of correcting the overestimation of the cut-off point.

However, there are some concerns about using the log-rank statistic at all, as a means of identifying patient cohorts, since thresholding by a prognostic index does not necessarily separate patients into groups with different clinical characteristics. This can result in mixed populations within single risk groups, which the application of an automatic rule extraction method has to be used to obtain coherent rules of variables for each identified risk group. Therefore, some statisticians prefer the use of a clustering method, based on the patients variables and not only in the prognostic indices obtained, as regression trees or K-means clustering, which results in a stratification with coherent patient groups but with relatively poor specificity for outcome, as measured by the separation between the group means of the overall mortality rates. This section, first presents the log-rank statistic, followed by the minimum p-value methodology description, robust log-rank bootstrap methodology, regression tree methodology, a clustering methodology and a clustering methodology based on learning metrics. Regressing the distribution of prognostic scores with rule-based trees (CART) succeeds in separating patient groups with statistically different mean survival but coherent membership in each group.

Many clinicians refer the important issue of explaining individual inferences by the modeling and stratification used. This is a key stage in evaluating the clinical plausibility of inferences made by analytical models to enable clinicians to apply these inferences with confidence. When the stratification methodology used is the regression tree, the tree itself explains the rules. However, when another stratification methodology, among the ones mentioned previously, is used a rule extraction algorithm is required. A previously published methodology designed to extract low-order Boolean rules from data, the orthogonal search rule extraction (OSRE) algorithm (Etchells, Lisboa, 2006) will be after explained.

2.3.1. Log-rank Test

Two groups of survival data can be compared looking at the survivor function of both groups. Plotting these survivor functions some information can be obtained. For example, if the survivor function of one group (Group I) is always greater than the survivor function of

the other group (Group II), it can be concluded that at any time t, the estimated survival probability of group I is greater than group II.

However, there are two explanations for this conclusion. First is that there is a real difference between the survival times of the two groups of individuals. The other reason for this is that there is no difference between the survival times in each group. The difference that has been observed is due to chance variation. To distinguish between these two explanations, the hypothesis testing is used. There is a non-parametric procedure, which is used to compare between two groups of survival data, namely Log-rank test (Collet, 2003).

The log-rank method is used to compare the survival of groups, which takes the whole follow up period into account. It is not required to know about the shape of the survival curve or the distribution of survival times. This test summarizes the extent to which the observed survival times in two groups of data deviate from those expected under the null hypothesis of no group differences. This means that the null hypothesis here is that there is no difference between the populations in the probability of an event (e.g. death due to breast cancer or death by any cause) at any time point. The larger the value of the statistic, the greater the evidence against the null hypothesis, because this statistic approximates the χ^2 distribution with one degree of freedom. The p-value associated with the test statistics can be obtained from the distribution function of a chi-square random variable. This statistic is obtained by calculating the U_L value, which is the difference between the total observed and expected number of deaths in Group 1, and the V_L value, variance of U_L.

$$U_L = \sum_{j=1}^{r} (d_{1j} - e_{1j}) \tag{54}$$

where

$$e_{1j} = \frac{n_{1j} d_j}{n_j} \tag{55}$$

$$V_L = \text{var}(U_L) = \sum_{j=1}^{r} V_{1j} \tag{56}$$

where

$$v_{1j} = \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)} \tag{57}$$

e_{1j} represents the expected number of individuals who died at time $t_{(j)}$ in group 1. Under the null hypothesis, the probability of dying at time $t_{(j)}$ does not depend on the group that an individual is. As the probability of death is d_j/n_j, multiplying this value by n_{1j} (number at risk before $t_{(j)}$), gives e_{1j}, which is the expected number of deaths in Group I at $t_{(j)}$.

The variance of U_L is the sum of the variances of d_{1j}, which are represented by v_{1j}, because the statistic U_L has a zero mean, and the death times are independent one from another. Moreover, the U_L value has a normal distribution when the number of death times is not too small. Then, the value $U_L / \sqrt{V_L}$ has a normal distribution with zero mean and unit variance. As the square of a standard random variable has a chi-square distribution on one degree of freedom, we can have:

$$\frac{U_L^2}{V_L} = \chi^2 \qquad \textbf{(58)}$$

2.3.2. Minimum p-value methodology

The minimum p-value methodology is an accepted strategy and was implemented in SAS (Williams, Mandrekar, Mandrekar, Cha, Furth, 2006). It starts by sorting all the records by the value of the prognostic index. Next, the total data are divided into two groups at a threshold value that sweeps the full range of prognostic indices from minimum to maximum. For each threshold, the log-rank statistic is calculated and hence a p-value results. The maximum of the log-rank statistic determines the first cut-off point. The same method is then repeated in each of the separated cohorts until no further partitioning exceeds a pre-set confidence level which, for this study, is as p-value of 0.01 (99% of confidence), corresponding to a test statistic value of around seven.

In practice, the test statistic very much exceeds this value across a wide range of thresholds with the associated p-values forming a plateau indicating that there are a wide range of candidate cutpoints in addition to the maximum log rank statistic that has been selected as can be seen in Figure 2.9 . Here, the significance of data partitions in the top-down approach that is generally applied to stratify patient data in medical statistics detects the global maximum as it can be observed in the left picture of Figure 2.9 , but this does not take into account that the statistical significance is high for a wider range of possible cut-off thresholds as shown in the right picture of Figure 2.9 .

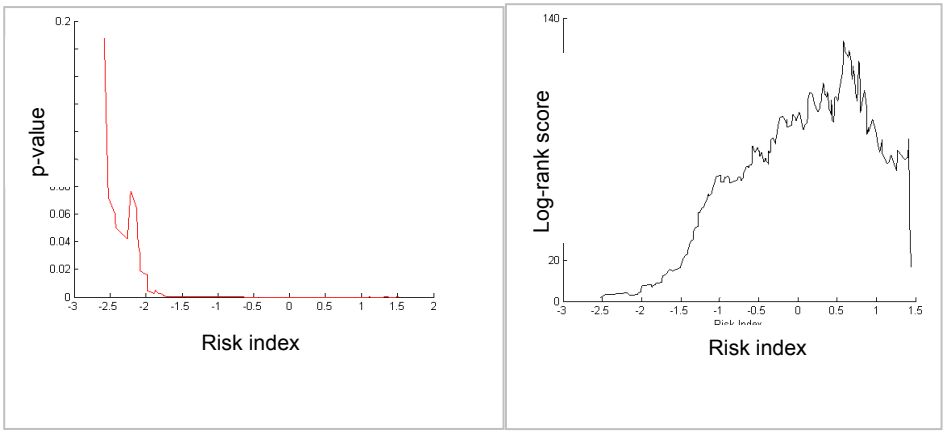

Figure 2.9 – Distribution of Risk index versus the log-rank score and p-value. The left picture represents the distribution of Risk index versus the log-rank score and the right picture represents the distribution of Risk Index versus the p-value. For each risk index there is a log-rank score as well as a p-value. The higher the log-rank score the lower the p-value.

2.3.3. Log-rank bootstrap methodology

The log-rank bootstrap methodology is proposed to make the stratification of risk indices more robust (Etchells, Fernandes, Jarman, Fonseca, Lisboa, 2008). The new approach is bottom-up according to the following procedure, which involves two nested loops: the inner loop and the outer loop.

Inner loop:

1. Bin the risk indices into discrete intervals each containing a minimum number of cases (e.g. minimum number=10).
2. Calculate the log-rank statistic for each pair of adjacent cells and aggregate together the two cells with the smallest value of this test statistic.
3. Repeat the process until the long-rank statistic is significant for all remaining cell pairs.

Outer loop:

1. Draw a sample of the risk indices, with replacement, of size equal to the original data size – this is a bootstrap re-sample of the data.
2. Apply the inner loop to convergence using the re-sampled data.
3. Allocate each value in the full range of the risk index to a risk group, from 1..Ngroups
4. Repeat from i a given number of times (e.g. number samples =3000).

5. Identify the distribution of values of Ngroups and discard all group assignments different from the mode of this distribution.
6. For each value in the full range of the risk index, build a distribution of risk group allocations – this clearly indicates the cases that fit firmly into a risk group and those that are near the boundary between adjacent groups.
7. Allocate each case in the original sample to the mode of the distribution of risk groups.

With this methodology, the training data is bootstrapped a number of times and the group allocation algorithm is applied to each bootstrap sample. Different bootstrap samples may produce different number of groups. Therefore, the most popular number of groups from all the bootstrap samples is chosen. Then, for the risk index value assigned to the training data, a distribution of group membership is derived from the bootstrap risk group allocation, as in Figure 2.10.

If the number of bootstrap samples is sufficiently large, then a probability of group membership can be assigned to each risk index score. This probability can then be used to indicate whether new data presented to the model are clearly in a particular risk group or are on or near the borders of two adjacent risk groups. This can be especially useful in the clinical context where the inference model allocates a patient to a particular group with a particularly aggressive treatment, but the patient may be in the crossover region between this more aggressively treated group and in adjacent risk group that is assigned to less aggressive treatment. An indication that a patient is in a crossover region of two groups may influence the decision of a clinician differently from that of a patient situated firmly in one of the groups.

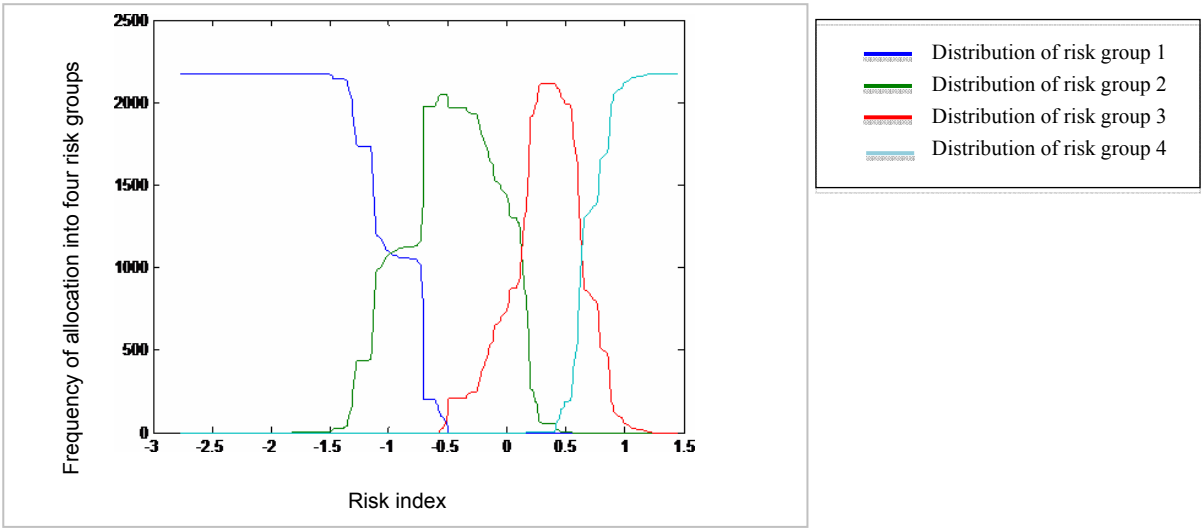

Figure 2.10 – Distribution of group membership.
This was derived from the bootstrap log-rank risk group allocation. Different colours represent each risk group. The interception of each risk group should be the cutpoint for risk group belonging.

2.3.4. Regression Tree Methodology

There are some algorithms that can be used to build the regression tree methodology. However, the algorithm used in this study, to build the tree, was CART. Here it is assumed the existence of a single output variable, which is the prognostic index obtained with a prognostic model, which was in this study, either with Cox or PLANN-ARD, and one or more predictor variables. The output variable is numerical, and the predictor variables may be a mixture of continuous and categorical variables. Regression trees are a recursive partitioning algorithm, which aim is to determine the optimal cutpoints for predictors. Therefore the resultant groups are most homogenous with regard to the outcome (i.e., minimum deviance). The terminal nodes of the tree contain the predicted output variable values. The starting level (complete dataset) is referred to as the root, each split is referred to as a branch, and the data subset resulting from the split is called a node; the terminal or ending nodes are referred to as leaves. Regression tree is built through a process known as binary recursive partitioning. This is an iterative process of splitting the data into partitions, and then splitting it up further on each of the branches. Initially, all the records in the training set are together and the algorithm tries breaking up this data, using every possible binary split on every field. The algorithm chooses the split that partitions the data into two parts such that it minimizes the sum of the squared deviations from the mean of the target in the separate parts. This splitting or partitioning is then applied to each of the new branches. The process continues until each

node reaches a user-specified minimum node size and becomes a terminal node. (If the sum of squared deviations from the mean in a node is zero, then that node is considered a terminal node even if it has not reached the minimum size.)

It is important to refer, while building a tree, that a full tree is complex and can yield an overly optimistic goodness of fit. Thus, methods to reduce the tree size have been developed so that the model is predictive in other cohorts. Moreover it is essential that the leaves contain statistically significant patients, as they need to be compared to each other in order to identify the belonging to a risk group. Therefore, it is necessary to define a minimum of records for the nodes, a minimum of records for the leaves or both.

After obtaining the final tree, it was developed a "pruning method", which aim was to find the leaves that have patients with a significantly similar survival of and join them. Therefore, leaves must be ordered by their final average and for each pair of leaves the log-rank test must be computed. The pair that has the minimum of this statistic, which means that the records existing in these leaves are the most similar ones in terms of survival are grouped and the new final average is computed, according to new grouped records. After it, all leaves are again ordered and the log-rank test is computed for the new grouped leaves with the ones which have the closest predictor average value. This algorithm is made until there is no significantly difference in terms of survival for the records belonging to each leaf. After the "pruning method", the regression tree remains a classification tree.

2.3.5. Clustering Methodology

The clustering method is an orthogonal approach, which aim is to cluster the clinical data first, then to produce a cohort tree with the leaves organized in order of mean group survival, where the prognostic index is not used at all. It is a k-means algorithm based, with a Euclidean metric, which used Monte Carlo methods to overcome the initialization problems. However, clustering with a Euclidean metric often provides insufficient discrimination for prognostic purposes.

2.3.6. Clustering methodology based on learning metrics

The Multivariate Fisher Distance is a metric based on learning metrics that measure local distances with respect to the distribution of one or more prognostic indices, which are proportional to the risk of mortality. This metric is embedded into a clustering model that can

estimate the most likely number of sample partitions within the data, providing a stratification of patients intro groups characterized by different survival curves. The Learning Metrics model (Kaski, Sinkkonen, Peltonen, 2001) provides a mean for estimating the distance function directly from the data, exploiting prior information concerning the distribution of the samples with respect to some auxiliary information. Such an auxiliary information is typically modeled as a random variable *c* that is bound to the input samples *x* by a conditional distribution $P(c \mid x)$, providing information regarding relevant aspects of the data. The scenario used comprises a set of samples (i.e. patient profiles) $x \in X$ that are associated to a multivariate auxiliary variable PI (prognostic index). In principle, what it has to be done is to learn a Fisher metric that can solve the variable's categorical bias problem, while taking into consideration the information brought by the distribution of samples with respect to the prognostic indices. The learning metrics, then, measures distances in terms of changes in the distribution $P(c \mid x)$ as x varies; such changes can be measured by the local Kullback-Leibler divergence that is a non-symmetric measure of the difference between two probability distributions as:

$$D(P(c|x)\|P(c|x+dx)) = dx^T J(x)\, dx \qquad \textbf{(59)}$$

where *J(x)* is the Fisher information matrix, that is

$$J(x) = E_{P(c|x)}\left(\frac{\partial \log(P(c|x))^T}{\partial x}\frac{\partial \log(P(c|x))}{\partial x}\right) \qquad \textbf{(60)}$$

where $E_{P(c|x)}$ is the expectation over *c*. The tensor *J(x)* is a positive semidefinite function defining a local scaling of the input space at the point *x*.

The Fisher information is a way of measuring the amount of information that an observable random variable *x* carries about an unknown parameter θ upon which the likelihood function of θ, $L(\theta) = f(X;\theta)$, depends. The likelihood function is the joint probability of the data, conditional on the value of θ, as a function of θ. The score function is:

$$\frac{\partial \log f(X;\theta)}{\partial \theta} \qquad \textbf{(61)}$$

Since the expectation of the score is zero, the variance is simply the second moment of the score, the derivative of the log of the likelihood function with respect to θ. Hence, the Fisher information can be written

$$L(x) = E\left(\frac{\partial \log f(X;\theta)^2}{\partial \theta} | \theta \right) \quad \textbf{(62)}$$

which implies $L(x) \in [0, \infty]$. The Fisher information is thus the expectation of the squared score. If a random variable is carrying high Fisher information then the absolute value of the score is often high. The new metric that is used to cluster the data in place of the Euclidean distance is as follows:

$$d^2(x, m_i) = (x - m_i)^T J(x)(x - m_i) \quad \textbf{(63)}$$

where m_i is the prototype of the *i-th* cluster and *J(x)* is the Fisher matrix at the point *x*.

For the purpose of our study, it was derived the informed metric based on the Fisher information matrix of the conditional distribution of the prognostic indices obtained by survival analysis. In particular, we will only focus on the prognostic indices obtain with Cox and with PLANN-ARD. Therefore it was considered a set of samples (i.e. the patient profiles) $x \in X$ that are associated to two independent auxiliary variables PI_{Cox} and $PI_{PLANNARD}$, through the respective conditional probabilities $P(PI_{Cox}|x)$ and $P(PI_{PLANNARD}|x)$. In addition, it was also considered the joint conditional probability of the two independent prognostic indices, that is $P(PI_{Cox}, PI_{PLANNARD}|x)$. To obtain an analytical formulation for the Fisher information matrix in our survival analysis scenario, we need to compute the derivative $\partial \log(P(PIx)) \partial x$ for each of the three distributions of the prognostic indices. To do so, it was considered PI_{Cox} and $PI_{PLANNARD}$ to be Normally distributed as

$$P(PI|x) \sim \exp - \left\{ \frac{(CCX^T \sum^{-1} CCX)}{2} \right\} \quad \textbf{(64)}$$

where CCX is a short form for

$$CCX = Bx + \beta_0 - \mu \quad \textbf{(65)}$$

where *B* is the *1* × *K* vector (*2* × *K* matrix for the joint distribution $P(PI_{Cox}, PI_{PLANNARD}|x)$) of the linear parameters of Cox survival model. The term μ refers to the Normal expectation and *x* is the *K* × *1* sample vector. To calculate the Fisher matrix with we need to compute

$$\frac{\partial \log(P(PI|x))}{\partial x} = \frac{\partial\left\{-\frac{(CCX^T \sum^{-1} CCX)}{2}\right\}}{\partial x} \quad \textbf{(66)}$$

Given the CCX , we can solve the previous equation using the product rule, yielding

$$\frac{\partial \log(P(PI|x))}{\partial x} = -(Bx + \beta_0 - \mu)^T \sum^{-1} B \quad \textbf{(67)}$$

By inserting this result in the equation $d^2(x,m_i)$, the Fisher distance for the prognostic indices can be obtained. To complete the derivation of the model, we need to fit the linear parameters B and β_0 to the values of the prognostic indices PI predicted by the Cox and PLANN-ARD model for each sample x. This entails solving a linear system with respect to $[B;\beta_0]$, which can be straightforwardly done by using the pseudo inverse matrix (see (Bacciu, Jarman, Etchells, Lisboa, 2009) for details).

Once the Fisher distance has been estimated, it can be embedded within the clustering algorithm at the stage where unit activation is computed. Following on the approach in (Bacciu, Jarman, Etchells, Lisboa, 2009), we focus on the CoRe clustering model (Bacciu, Starita, 2008), an algorithm that performs cluster number identification by exploiting an information compression mechanism of the visual cortex, named *repetition suppression*. Starting from an initial overestimation of the actual cluster number, the CoRe algorithm iteratively suppresses neurons whenever they fire un-selectively for the input patterns, eventually pruning unselective units from the network. The neurons retained at the end of the learning phase encode the clusters found by CoRe within the input data. Hence, the estimated cluster number is equal to the network size at convergence. The response of CoRe units is determined by a multivariate Gaussian: to embed the Fisher metric within such units, we consider the following activation function

$$\varphi_i(x, m_i, \sum_i^{-1}) = \exp\left\{-\frac{1}{2}\left[R(x)(x - m_i)\right]^T \sum_i^{-1^T} \sum_i^{-1}\left[R(x)(x - m_i)\right]\right\} \quad \textbf{(68)}$$

where $R(x)$ is the right Cholesky decomposition of the Fisher matrix $J(x)$. Regards learning, the original CoRe algorithm updates unit means and variances in the direction given by the gradient ($\partial\varphi_i/\partial m_i$) and ($\partial\varphi_i/\partial\Sigma_i$).

Since the Fisher metrics is a Riemannian metrics, such steepest descent can be computed by means of the natural gradient (Amari, 1998). The application of this rule to the prototype vectors m_i yields to the same update rules as for the Euclidean case; on the other hand, the learning equations for the scale matrix Σ_i need to be slightly modified to enclose the contribution from the Fisher matrix $J(x)$ (see (Bacciu, Jarman, Etchells, Lisboa, 2009) for the details of the learning equations). In the experimental phase, the CoRe algorithm with the Fisher metrics has been applied to discover clusters within the patient population, by exploiting the information from the distribution of the prognostic indices PI_{Cox} and $PI_{PLANNARD}$ in isolation and jointly. The simulation setup is the same described in (Bacciu, Jarman, Etchells, Lisboa, 2009): the CoRe network has been initialized with 30 units and the cluster number estimates are based on 50 repeated runs of the algorithm, with random initial prototype assignments.

2.3.7. OSRE rule extraction algorithm

OSRE is an orthogonal search rule extraction (OSRE) algorithm that provides a practical and efficient tool to explain the predictions from different prognostic models, after applying a stratification methodology. The main goal of OSRE is to produce rules that are comprehensible to a human analysis. The OSRE methodology searches for rules using multivariate descriptions of data sub-sets, and provides an efficient alternative to methods that follow sequences of univariate searches such the well-known method of Classification and Regression Trees. The OSRE methodology transcribes the RULENEG algorithm developed by Pop. et al (Pop, Hayward, Diederich, 2009) for binary data, as if there are categorical or ordinal variables RULENEG cannot be applied to the data (Etchells, Lisboa, 2006).

The OSRE algorithm searches through the Boolean space in order to generate rules in response to a certain risk group classified previously. The rules' search is restricted to the data predicted to be within class and searching in successive orthogonal directions. The algorithm searches the multi-variable space for changes in the networks response, sweeping each variable over its possible values whilst keeping all other variables constant. Therefore, if there are m variables, each with up to n values, there are at most n^m points to evaluate, which is not viable to handle.

A disadvantage of this algorithm is that it can sometimes produce as many rules as there are data. This happens because the orthogonal search is performed relative to a data point and

the distance from a particular data point to the decision boundary may be unique to that data point. Therefore, a rule refinement technique must be applied, which reduces the number of rules. This must delete the rules which are covered by other rules and incorporate a single rule allocation to a patient by ordering into a hierarchy were a rule's position depends on how accurate it is, analysing the sensitivity and the specificity of the rule.

Summarising, after obtaining a risk group belonging for each patient in order to obtain a set of rules from the training data, a neural network, MLP, is run in turn for each prognostic group. For each MLP, the risk group of interest is assigned the target 1 and the remaining risk groups are assigned to 0. For each risk group the rules are placed into a rule hierarchy with the aim to assign one rule only to a patient. If a patient is assigned to a rule for more than one risk group it is placed into the more conservative of these survival groups. It can also be flagged as a patient who is in the cusp of decision boundaries. If a patient is not classified by any of the obtained rules, than this patient is considered as an outlier of the rule extraction survival model.

2.4 - Clinical Prognostic Indices

Single items of patient data, such as age or smoking history are widely used in making clinical decisions. Prognostic models are more complex tools for helping decision making that combine two or more items of patient data to predict clinical outcomes. They are of potential value when doctors are making difficult clinical decisions (such as ordering invasive tests, selecting which patients should benefit from scarce resources or selecting the most appropriate treatment), or selecting uniform groups of patients for clinical trials. Risk prediction models can play an important role in decision making and future management of individual or groups of patients with a particular medical condition. These models are usually designed to predict the risk of a patient developing some future clinical event based on a number of patient and disease characteristics. Unfortunately, most of the published indices of risk have no clinical impact and disappear into archives. There are however some prognostic indexes that are widely used and accepted in clinical practice. Therefore, the new obtained indices must be compared with the existent ones and demonstrate that make better predictions in order to be potentially accepted in clinical practice.

As an example, QRISK, a new cardiovascular disease risk has been developed and internally validated (Hippisley-Cox, Coupland, Vinogradova, Robson, May, Bringle, 2007).

After some critiques of the original papers, other studies have been carried out to validate externally this index and compare it with existing ones (Collins, Altman, 2009), (Hippisley-Cox, Coupland, Vinogradova, Robson, Bringle, 2008), providing evidence to support the use of the index in favour to the existing ones.

In terms of breast cancer disease, there are different prognostic indices that can be derived in different risk groups that are widely used, namely NPI, TNM and St. Gallens, which are presented in this section. It is very important for both, breast cancer patients and clinicians, to provide some form of interpretation of the prognostic groups in terms of clinically relevant variables. This would be highly appreciated when predicting more accurately the clinical course of the disease at the time of initial treatment. When defining a prognostic group, the survival for each group at a certain time scale, can be obtained as well as the more adequate treatment.

2.4.1. NPI (Nottingham Prognostic Index)

In practice, oncologists frequently use an algorithm commonly referred to as Nottingham Prognostic Index (Haybittle, Blamey, Elston, Johnson, Doyle, Campbell, Nicholson, Griffiths, 1982), which was derived using the proportional hazards modelling, linear in the parameters. The model was fitted to explain the variation in survival, lending itself naturally to the derivation of a discrimination index that has since undergone extensive multi-centre validation (Galea, Blamey, Elston, Ellis, 1992). It was concluded that the use of NPI allows to accurately predict the prognosis of patients with breast cancer and carried out surgical and systemic adjuvant procedures that are appropriate for the individual patient (D'Eredita, Giardina, Martellotta, Natale, Ferrarese, 2001). This identified three factors from nine firstly recorded for each patient as being significant indicators of survival prognosis, namely Tumour size (in cm), grade of tumour (coded as 1,2 or 2) and number of axillary nodes affected (coded as 1 (no nodes involved), 2 (1 to 3 nodes involved) or 3 (more than 3 nodes involved)) in the form:

$$NPI\ score = 0.2*Tumour\ size + Grade\ of\ tumour + Number\ of\ axillary\ nodes\ affected \quad \textbf{(69)}$$

The split of this index in different prognostic groups is at all explained by expert knowledge. These values can be splited in 3, 4 or 5 different groups. The division in five

different risk groups is as following: excellent prognosis group (NPI ≤ 2.4), good prognosis group (NPI > 2.4 and NPI ≤ 3.4), moderately prognosis good group (NPI > 3.4 and NPI ≤ 4.4), moderately prognosis poor group (NPI > 4.4 and NPI ≤ 5.4) and pour prognosis group (NPI > 5.4). The moderately good prognosis group and the moderately poor prognosis group can be combined resulting in four risk groups. If the excellent group is combined with the good prognosis group, 3 risk prognostic groups can be obtained. For each risk group it was identified the 5, 10 and 15 years survival.

This index has the advantage of utilising information about histological differentiation, which makes it more specific for patients with early-stage disease, which is non-metastatic.

2.4.2. TNM prognostic index

TNM is a widely used staging system in breast cancer patients. TNM is developed and maintained by the International Union Against Cancer (UICC). The TNM classification is also used by the American Joint Committee on Cancer (AJCC) and the International Federation of Gynecology and Obstetrics (FIGO). In 1987, the UICC and AJCC staging systems were unified into a single staging system.

TNM stands for "tumour, nodes, metastasis". TNM staging takes into account the size of the tumour, whether the lymph nodes are affected and whether cancer has spread to other parts of the body (metastasis). The size of the tumour can be divided in four stages: stage 1 represents a tumour with less than 2 cm, stage 2 a tumour with more than 2 cm and less than 5 cm, stage 3 a tumour bigger than 5 cm, stage 4 when the tumour has spread into the chest wall, skin, is fixed to the skin and chest wall or when it is a inflammatory carcinoma. The lymph nodes can be divided in 4 different stages: N0 if no cancer cells were found, N1 if cancer cells are in nodes in the armpit but the nodes are not stuck to surrounding tissues, N2 if there are cancer cells in the lymph nodes in the armpit, which are stuck to each other and to other structures or if there are cancer cells in the lymph nodes behind the breast bone (the internal mammary nodes), N3 if there are cancer cells in lymph nodes below the collarbone, in the armpit and under the breast bone or above the collarbone. The metastases stage can be divided in two stages: M0 if there is no sign of cancer spread and M1 if the cancer has spread to another part of the body, apart from the breast and lymph nodes under the arm.

Once the T, N, and M categories have been determined, this information is combined in order to obtain a stage grouping, 4 stages were considered from stage I (the least advanced

stage) to stage IV (the most advanced stage). Non-invasive cancer is listed as stage 0. Stage I incorporates T1N0M0 and T0N1M0. Stage 2 incorporates: T1N1M0, T2N0M0, T2N1M0, T3N0M0. Stage 3 incorporates: T0N2M0, T1N2M0, T2N2M0, T3N1M0, T3N2M0, T4N0M0, T4N2M0, T1N3M0, T2N3M0, T3N3M0, T4N3M0. Stage 4 incorporates any T, any N and M1. According to these group stages the 5-year relative survival rate can be obtained. The strength of this index is that it only requires clinical information, which is obtained by the clinician without recourse to laboratory tests. However, its discrimination power is best for separating severe from early-stage disease.

2.4.3. St. Gallen Classification

Another approach for choosing the best treatment options for early breast cancer has been proposed by an international panel of experts in a report from the St. Gallen conference. The report represents the consensus on early breast cancer treatment that emerged from the conference (Harbeck, Jakesz, 2007). The consensus maintains an emphasis on targeting adjuvant systemic therapies according to subgroups defined by predictive markers. It further refines the treatment algorithm by identifying 'thresholds for indication' of each type of systemic treatment modality (endocrine therapy, anti-HER2 therapy, chemotherapy) based on criteria specific to each modality. The report emphasises the importance of identifying which type of breast cancer a patient has and which treatment, or combination of treatments, are most likely to be successful.

Low risk	Node negative AND all of the following:	pT ≤ 2 cm AND grade=1 AND no extensive PVI AND ER AND/OR PgR expression AND neither HER2 over expression nor amplification AND Age≥35 years
Intermediate risk	Node negative AND at least one of the following:	pT > 2 cm OR grade=2-3 OR extensive PVI OR lack of ER AND PgR expression OR HER2 over expression or amplification OR Age<35 years
	Node positive (1-3 N+) AND	ER AND/OR PgR expression AND neither HER2 over expression nor amplification
High risk	Node positive (1-3 N+) AND	Lack of ER and PgR expression OR HER2 over expression or amplification
	Node positive (≥ 4 N+)	

Table 2.3 – St. Gallen risk categories 2007.

St. Gallen classification defined 3 risk categories—low, intermediate and high—using a combination of nodal status, tumour size, histological and nuclear grade, oestrogen receptor, progesterone receptors, the status of extensive peritumoural vascular invasion (PVI), human epidermal growth factor receptor 2 (HER2) and age. The definition of risk categories is presented in Table 2.3. Recommendations for adjuvant systemic therapy were based on the three categories of risk, menopausal status, and steroid hormone receptor status.

The recommendations of the St. Gallen consensus panel provide a minimal standard for up-to-date breast cancer treatment, which are based on expert opinions as well as published trial data.

Concluding, in this chapter it has been presented the essential analytical methodologies, which are required for the development of this thesis. It has also been shown the improvement of some of the existing methodologies that contributed to this thesis' innovation.

Chapter 3 - Study Design for Prediction Models

There are several issues in the design of studies for prediction models. These include the main subjects such as the selection of patients for a cohort study, choosing predictors and outcome variables, the cohort sample size and how to deal with missing values. Finally, an appropriate adequacy of the model must be done in order to verify the choices done. All these important issues are presented in this chapter, defining at the end of the chapter a section, which defines a modelling strategy that was followed in this thesis.

3.1 - Choice of covariates

In order to obtain a well-performing prediction model it is mandatory to have strong predictors. It is essential therefore, to make a careful model selection which depends on the study aim and the responsibility of the analyst. Bradburn et al (Bradburn M.J., Clark, Love, Altman, 2003) suggests three possible scenarios as to why a study may use a multivariate model and how to deal with each one:

1. *A single factor is under investigation for its association with survival, but several other factors exist.*
2. *A collection of factors of known relevance is under investigation for their ability to predict survival.*
3. *Where a collection of factors are under investigation for their potential association with survival, possibly with additional known factors.*

While choosing a significant model as well as the significant covariates it is worthwhile to consider the specific question under investigation and the opinion of a non-statistical specialist of the study.

When the aim of the study is to find a set of potential explanatory variables that the hazard function depends on and combine them to find the best models, the statistic AIC (Akaike's Information criterion), can be used (Collet, 2003). This statistic measures the extent to which the data are fitted by a particular model.

The AIC of a model can be defined as:

$$AIC = -2\log L + \alpha q \qquad \textbf{(70)}$$

where q is the number of unknown β parameters in the model, α is a predetermined constant and L is the optimised log-likelihood function for the proportional hazards. The smaller the value of this statistic, the better the model.

The value of AIC will increase when an unnecessary variable is added to the model. If the only difference between the two models is that one includes unnecessary variables, the values of AIC of both models will not be very different.

When the number of variables is relatively large it can be computationally expensive to compare all the models that can be derived. To avoid this situation there are some automatic routines for variable selection. These are based on forward selection, backward selection or a combination of both, frequently called a stepwise procedure.

In forward selection, variables are added one at a time. At each step the variable added is the one that most decreases the value of AIC. The process ends when no other variable decreases the AIC value by a statistical value. In backward elimination variables are excluded one at a time and the variable that is omitted is the one that increases the value of AIC. The process ends when there is no candidate variable for deletion that increases the AIC value more than a statistical amount.

In stepwise procedure, probably the most widely used procedure in medical applications, the variables are added to the model one at a time. The variable added is the one that gives the largest decrease of the AIC value. The process ends when the next candidate for inclusion in the model doesn't reduce the AIC value by more than a predefined amount. After selecting the best variables, the routine may exclude some of these variables if the variable omitted is the one that increases the value of AIC.

There are some advantages and disadvantages about using these procedures. The advantages reside on the following: they are straightforward to apply in modern statistical packages; they are relatively objective and usually reach their goal of making a model smaller (this helps to eliminate variables that have no true relationship to the outcome as noise variables). Unfortunately, these automatic procedures have some disadvantages as they identify solely one model instead of a set of statistical significant models. They depend on the variable selection process (forward selection, backward selection, stepwise selection) and they depend on the stopping rule used. There are also other problems derived from automated selection techniques such as the best model is derived solely on statistical grounds, the regression coefficients are biased and standard errors and p-values are too small, especially if the sample sizes are very small and if there are little events (Bradburn, Clark, Love, Altman, 2003).

Other selection methods have emerged in order to improve the mentioned ones. However, these methods are generally computer intensive and are infrequently encountered in medical applications. Some approaches use resampling methods such as the bootstrap bagging and boosting and others use principles of Bayesian analysis, such as Bayesian model averaging (BMA) (Steyerberg, 2009).

3.2 - Sample size considerations

There are several aspects in the design of a model of survival data that must be considered. One crucial issue that must be considered is the number of patients that is required to make the study. A predictive model based on a small sample of individuals will be less reliable than one based on a larger number of individuals. However, it is unethical to waste resources in studies that are unnecessarily large. The automatic selection procedures described previously have some problems when dealing with small data sets, as well as with large data sets, as can lead to overoptimistic results.

Although it is necessary to give a very high relevance to the number of individuals in the study, the number of events per variable (EPV) is a bigger measure of power or even validity of a survival analysis study. Some work has been made to conclude about the effects of changing the EPV in multivariable analysis, as the effect of overfitting or underfitting. The overfitting may be caused by unimportant variables being predictively important. On the other hand the underfitting may be caused by the reverse effect, which is rejecting a variable that

have a significant impact in survival. Some simulation studies (Concato, Peduzzi, Holford, Feinstein, 1995), (Concato John, Peduzzi, Holford, Feinstein, 1995), (Concato, Peduzzi, Holford, Kemper, Feinstein, 1996) have been made to ensure the number of EPV that can lead to fewer problems while designing a predictive model. These studies suggested that an EPV or 10 is the minimum value that can obtain good predictions. For example, as the EPV value decreases the regression coefficients increases, producing overestimation as well as underestimation of the true effect. Also while the EPV decreases, the power to detect significant effects also decreases, resulting in problems of underfitting and significant effects could be identified in the wrong direction. Other problems such as the low coverage of confidence intervals, the lack of validity of the test statistics for the model and the increasing of the frequency paradoxical associations may result when the EPV is lower than 10.

3.3 - Missing Covariate Data

Missing data is a common problem when developing survival models. Unfortunately, it is often neglected or not properly handled during analytic procedures, and this may substantially bias the results of the study, reduce study power, and lead to invalid conclusions. While bias may be introduced into research through several other mechanisms (e.g., study design, patient sampling, data collection, and or other aspects of data analyses), naïve methods of handling missing data may substantially bias estimates while reducing their precision and overall study power, any of which may lead to invalid study conclusions.

The usually method to overcome this problem is excluding the individuals whose prognostic factors are missing from the study. This method not only wastes valuable data, but also can lead to invalid results, if the excluded group is a non-random sub-sample of the entire data. Here, the completely observed cases that remain will be unrepresentative of the population for which the inference is usually intended: the population for all cases, rather than the population of cases with no missing data. In addition, the statistical power of the analyses decreases (Greenland, Finkle, 1995) as well as the number of events per variable, which can result in less stable results (Bradburn, Clark, Love, Altman, 2003).

There are several reasons why the data may be missing. Depending on these reasons, missing data can be classified as MCAR (Missing completely at random), MAR (Missing at random) and MNAR (Missing not at random). It is important to consider these, since approaches to handle missing data in statistical analysis rely on the assumption on the

mechanism. Missing data is called to be missing completely at random when the probability of an observation X missing is unrelated to the value of X or to the value of any other variables. Equipment malfunctioned, the data were not correctly entered and spilling of material are some examples of MCAR data. The analysis of this kind of data remains unbiased, that is some power can be lost, but the estimated parameters are not biased in the absence of data.

If the data meet the requirement that missingness depends on values of variables that were actually measured, then the missing data is classified as MAR. In other words, in a given data set (Y) consisting of observed values (Y_{obs}) and missing values (Y_{mis}), MAR is present if the probability that a value is censored is dependent only on Y_{obs} and not on Y_{mis}. MAR examples include more missing values in older subjects, subjects from a certain region or from an earlier calendar time. This missing data is a problem, although there are ways of dealing with the issue in order to produce meaningful and relatively unbiased estimate. Assuming a MAR mechanism exists it is less restrictive and more tenable than assuming that an MCAR mechanism exists.

When the missing data is not at random, the only way to obtain an unbiased estimate of parameters is to model missingness. The MNAR mechanism is present when the pattern of censoring is related to variables that were not collected and are not related to Y_{obs}, or to Y_{mis} rather than to Y_{obs}. As such, it is impossible to estimate the missing values that are censored from other known values in the data set. This underlying mechanism of missing is often referred to as ‘‘nonignorable’’ because the probability that a value is missing depends on other unknown or missing values. Examples include selective non-response on certain questions (sexual orientation, income) or clinical condition (missing if a severe condition is present, which is not measured accurately). To identify MNAR as the existing mechanism, data must be available to fully explain the pattern of missingness. Unfortunately, this never occurs when censoring is beyond the investigator’s control and rarely occurs otherwise. Available methods of handling incomplete data with an MNAR mechanism may not produce valid results, and there is no universal method for handling incomplete data in this situation. However, some methods (e.g., Multiple Imputation) have been shown to produce less biased results than other methods, even when data are MNAR.

While certain naïve methods for handling incomplete data (e.g., complete-case analysis, available-case analysis, and the missing indicator method) are likely to generate biased results

under a MAR mechanism, because the data would have to be MCAR for these methods to work, the MAR assumption is necessary and sufficient to justify handling missing data using more sophisticated techniques (e.g., Multiple imputation or maximum likelihood estimation) to produce valid estimates. The major distinction between these two sophisticated approaches, imputation and likelihood-bases is that imputation methods substitute the missing values with plausible values so that the completed data can then be analysed with standard statistical techniques, while likelihood-based approaches do not require the missing data to be estimated explicitly. The later approach is computationally complex and the software is less readily available for survival data.

Imputation methods can be considered as single or multiple imputation methods. Both can be an alternative procedure that can be used; however this thesis will only be focused on multiple imputation (Clark, Altman, 2003), because when Multiple Imputation (MI) was compared with alternative methods of handling incomplete data it has been shown that MI generates less biased estimates with more statistical efficiency (Newgard, Haukoos, 2007). In this framework, missing data are imputed or replaced with a set of plausible values. Then, several data sets are constructed, each being analysed separately. After, their results are combined in order to diminish the uncertainty introduced by the imputation.

Multiple Imputation:

Data imputation, which is the practice of "filling in" missing data with plausible values, is an attractive approach to analysing incomplete data sets. This approach solves the missing data problem at the beginning of the analysis. However, a naïve or unprincipled imputation method can create more problems than it solves.

The general basis of multiple imputation it to use observed values to generate a range of plausible values (imputations), based on existing correlations between variables. In multiple imputation, the unknown missing data are replaced by *m* independent draws from an imputation model, and each of the *m* complete data sets is analyzed by standard complete data methods.

One question that arises with imputation models is that we may want to predict missing values for one predictor, using other predictors that also have missing values. This can be however solved by an iterative process of an imputation step, which imputes values for the missing data and a posterior step, which draws new estimates for the model parameters, based

on the previously imputed values. This process continues until convergence. The variation among the *m* imputations reflects the uncertainty with which the missing values can be predicted from the observed data. At the end of this step there will be *m* complete data sets that reflect uncertainty about the true values of the missing data.

After creating *m* completed data sets, analysis are performed by treating each completed data set as a real complete data set and finally the results from the *m* complete data analyses are combined using a set of rules that appropriately account for the variance in the MI process. Multiple imputation results in a three-step process: imputation, analysis and pooling (see Figure 3.1).

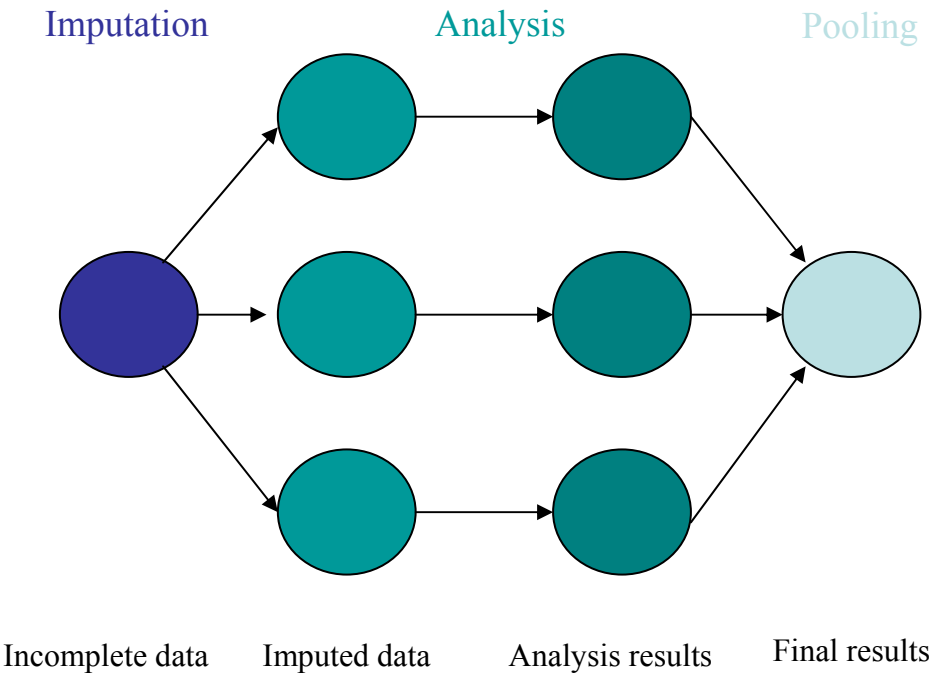

Figure 3.1 – This figure represents all the three phases of multiple imputation. Here the incomplete data set is imputed 3 times and after its imputation, the 3 complete data sets are analysed and finally pooled.

The imputation step imputes the missing entries of the incomplete data set, not once, but *m* times. The imputed values are drawn from a distribution, which can be different for each missing entry. The analysis phase is an important step of the imputation method because each of the *m* completed data sets must be analysed. After that the *m* analysis results into a final model. This phase is called the Pooling phase and consists of computing the mean over the m repeated analysis, its variance, and its confidence interval.

Some simulation studies have shown that in multiple imputation a *m* value of 3 is often

adequate for data with 20% of missingness (Van Buuren, Boshuizen, Knook, 1999). In his work, Rubin (Rubin, 1987) shows that the computational efficiency of an estimate is influenced by the number of imputations and by the rate of missing information.

This fact can be surprising as in other applications of Monte Carlo, hundreds or thousands of draws are often needed to achieve an acceptable level of accuracy. There are two fundamental reasons that in multiple imputation a small value of *m* will usually suffice. First, multiple imputation relies on simulation to solve only the missing data aspect of the problem. This means that one could effectively eliminate Monte Carlo error by choosing *m* to be large, but with multiple imputation the resulting gain in efficiency would typically be unimportant, because the Monte Carlo error is a relatively small portion of the overall inferential uncertainty. The relative efficiency of a point estimate based in m imputation to one based on an infinite number of imputations is approximately $\left(1+\frac{\lambda}{m}\right)^{-1}$, where λ is the rate of missing data. For example, for a λ=0.2 and m=3, the error estimate will be 1.033 times as large as the estimate with m=∞. On Table 3.1 it the relative efficiency of multiple imputation can be observed.

Number of imputations (m)	**Proportion of Missing data (λ)**				
	10%	30%	50%	70%	90%
3	0.97	0.91	0.86	0.81	0.77
5	0.98	0.94	0.91	0.88	0.85
10	0.99	0.97	0.95	0.93	0.92
20	1.00	0.99	0.98	0.97	0.97
30	1.00	0.99	0.98	0.98	0.97

Table 3.1 – Relative efficiency of multiple imputation.

So, the additional resources that would be required to create and store more than a few imputations would not be worthy. Second the rules for combining the *m* complete data analysis account for Monte Carlo error (Shafer, 1997).

The specification of the imputation model is the most complex step in multiple imputation, as its objective is to approximate the true distributional relationship between the unobserved data and the available information. There are two modelling choices that usually have to be made: the form of the model (e.g. linear, logistic, polytomous) and the set of variables that enter the model. As a general rule, using all available information yields multiple imputations that have minimal bias and maximal certainty.

The Schafer approach for imputed values (Schafer, 1999), assumes the distribution to be a multivariate Normal for continuous data, a log-linear model for categorical data or a general location model for a mixture of continuous and categorical data. Under certain circumstances, categorical variables can be quite reasonably when it is applied the same distribution. In other situations, however, it is desirable to use a model specifically designed for categorical data, as a log-linear model.

There are other multiple algorithms besides the Schafer approach, where it is not assumed a particular form for the multivariate distribution of the data, that is no explicit non-response model is needed, and only the posterior conditional distributions $p(Y_{missing}/X)$ needs to be specified. It is assumed that a multivariate distribution exists, and that draws from it can be generated by iteratively sampling (Gibbs sampling) from the conditional distributions. That is each incomplete entry is initialized by filling in a random draw from the marginal distribution of Y_{obs}. Then, Y_1 is imputed by the elementary procedure conditional on all other data (observed and imputed combined), then Y_2 conditional on all other data (using the most recent imputations for Y_1), and so on, until all incomplete variables in Y, Z, U and V have been imputed. Subsequently, start a second pass through the data, using all imputations created during the first pass, and so on. Therefore, the multivariate problem is split into a series of univariate problems.

First, each incomplete entry is initialized by filling in a random draw from the marginal distribution of Y_{obs}. After, Y_1 is imputed by the elementary procedure conditional on all other data (observed and imputed combined), then Y_2 conditional on all other data, using the most recent imputations for Y_1. The process continues until all incomplete variables have been imputed. The whole procedure is executed *m* times in parallel, thus producing *m* complete data sets.

The application of Gibbs sampling ensured that the imputation process was not deterministic because there was a random variation between the completed data sets, and this is the main reason for the approach to be considered as a form of Bayesian simulation.

The number of iterations needed is much lower than is common in modern Markov chain simulation techniques that often require thousands of iterations. In regression switching, the posterior distributions of the regression coefficients absorb the uncertainty in the predictors. The main question now is whether the number of steps is or not enough to stabilize these posteriors. Also, the elementary procedure creates imputations that are already statistically

independent, therefore no iterations need to be wasted for achieving independence between successive draws, as is typical for MCMC methods. Brand's simulation study successfully used just 5 iterations (Brand, 1998). In our implementation we have used 10 iterations to check convergence.

Multiple imputation requires a selection of specific types of variables, namely the target variables and the auxiliary variables. The inclusion of the last ones has been considered to reduce bias and variance and improve statistical efficiency (Newgard, Haukoos, 2007).

3.4 - Predictive accuracy and Validation of predictive models

The purpose of prognostic models is to provide valid outcome predictions for new patients. Essentially, the data set to develop a model is not of interest other than to learn for the future. To show that a prognostic model is valuable, however, it is not sufficient to show that it successfully predicts outcome in the training data. It is needed evidence that the model performs well for other group of patients, that is prognostic models need to be internally and externally validated (Altman, Vergouwe, Royston, Moons, 2009). The idea of validating a prognostic model generally means that it works satisfactorily for patients other than those from whose data was derived.

A key threat to validity is overfitting, i.e. that the data under study are well described, but that predictions are not valid for new subjects. Overfitting causes optimism about a model's performance in new subjects. Overfitted models will show both poor calibration and poor discrimination when validated in new patients. A drop in discriminative ability at external validation compared with the development setting can be explained by overfitting.

In this context, the assessment of model performance has to focus on the accuracy of the predictions, rather than merely on the covariate effects and their statistical significance.
Almost all models are developed in order to predict the outcome of future patients. The reasons for predict this outcome, are:

1. Inform treatment or other clinical decisions for individual patients
2. To inform patients and their families
3. To create clinical risk groups for informing treatment or for stratifying patients by disease severity in clinical trials.

A prognostic model in medicine only has clinical value if it has been shown that predict outcome with some success. For all types of validation we need performance criteria in line with the research question.

As the purpose of prognostic statistical models is to identify the combination of risk factors that might predict patient survival, a model must be able to:

1. Make unbiased predictions, that is, make the agreement between observed and predicted event rates for group of patients – called *calibration*
2. Ability to distinguish between patients who experience or not the event of interest, called *discrimination*

If a predictive model has poor discrimination, no adjustment or calibration can correct the model. On the other hand, if discrimination is good, the predictor can be calibrated without sacrificing the discrimination.

Measures of discrimination include the c-index and Negelkerkes's R^2 (Harrell, 2001). The c-index is a generalisation of the area under the ROC (receiver operating characteristic) curve to the case of censored survival data and is the concordance between the observed and predicted survival. C index considers the "concordance" between the ranking of the predicted failure times and that of the observed times, for pair of subjects. Calibration may be quantified using an estimate of slope shrinkage (Harrell, 2001).

For a model including covariates with time-dependent effects and/or time-dependent covariates, the original definition of C would require the prediction of individual failure time. The time-dependent discrimination index, C^{td} (Antolini, Boracchi, Biganzoli, 2005) uses the whole predicted survival function as outcome prediction, and the ability to discriminate among subjects having different outcomes is summarized over time. Therefore, the C^{td} index can be viewed as a novel definition of concordance, which is: a subject who developed the event should have less predicted probability of surviving beyond his/her survival time than any other subject who survived longer. C^{td} ranges from 0.5 (representing absence of discrimination) to 1 as maximum discrimination.

Calibration can be assessed by plotting the observed proportions of events against the predicted probabilities for groups defined by ranges of predicted risk. Perfect predictions should be on the 45º line. This plot can be accompanied by the Hosmer-Lemeshow goodness of fit test (D'Agostino, Nam, 2004). This test has limited statistical power to assess poor

calibration and is oversensitive for very large samples. In survival context the calibration of a model is usually studied at a fixed time point.

The calibration plot can be extended into a "validation plot" as a central tool to visualize model performance. Calibration is shown by observed outcomes being close to prediction, while discrimination aspects can be indicated with the distribution of the predicted probabilities.

An important aspect of prediction is to consider whether a model derived from an analysis of the original data set is transportable to similar patients in other locations. This concept is usually called generalizability or validity and depends on the quality of the prediction model as developed for the development setting and on characteristics of the population where the model is applied. A model that passes this test is considered to be validated. This concept can be considered at two levels: patients coming from the same population where the model was developed (reproducibility) and to patients coming from a different plausibly population (transportability). There are some features that the development of a successful model in medicine depends on:

1. The potential for accurate prognosis (which is presumably unknown);
2. The intrinsic prognosis information in the available factors;
3. The measurement process, which converts the intrinsic information into numbers;

A model might fail either because is statistically invalid or because the intrinsic prognostic information is weak. There is nothing to do about the last reason for fail. Regarding these two types of models fails, there can have two types of validating a clinical model:

1. A statistically validated model, which passes all appropriate statistical checks, including goodness-of-fit on the original data and unbiased prediction on a new data set.

2. A clinically validated model, which performs satisfactorily on a new data set according to context-dependent statistical criteria, laid down for it.

According with these both types of validating a model, it can happen that a statistically validated model is not a clinically validated model, or the contrary. However, a clinically validated model is more useful than a statistically validated one.

There are several reasons why prognostic models may not perform well. The first is related with deficiencies of standard modelling methods. These methods are used to derive prognostic models with data-dependent aspects, leading to an overoptimistic assessment of predictive

performance. One problem of the standard modelling method is how to choose the models variables. The stepwise selection algorithm for example, is a fully automated procedure that doesn't require intellectual input. It is desirable to use clinical criteria or statistical methods to reduce the number of candidate variables, reducing the risk of an overoptimistic model.

The second reason for a prognostic model not performs well is related with the deficiencies in the design of the prognostic studies, which can result in misleading findings, creating over optimism and/or bias. These include the absence of clear inclusion or exclusion criteria, as what to do with missing data, unclear rationale for the choice of treatments, an inadequate sample size and the number of events per variable in the data.

The third reason for a prognostic model not performs well is that models may not be transportable. The main problem for this reason is the degree of dissimilarity between the settings of patients in different centres, including differences in healthcare systems, methods of measurement and patient characteristics. If the model contains all the important prognostic variables, then the model should be transportable to a centre with different setting patients. However, if other important variables are not present in the model it can lead to different model performances in different centres.

Here arises the question: How could we validate a model?

Altman and Royston (Altman, Royston, 2000) considers the following considerations in validating a model:

1. Study design - The model validation should include internal validation temporal validation and external validation. The first validation can be made using the data splitting methods, cross-validation or bootstrapping. The second one can be processed by evaluating the performance of a model on subsequent patients within the same centre(s). The last validation is related with the generalizability of the model. The goal of this validation is to demonstrate satisfactory performance for patients from a different population from the original.
2. Measuring the intrinsic prognostic information - There are some studies about the measurement of the prognostic information of a model. The idea of greater or lesser separation between prognostic groups as a measure of prognostic information remains attractive, as is interpretable and pragmatic.

3. Comparing predictions with observations - It can be said that evaluation consists of comparing the appropriate observed and predicted measure, an aspect of model calibration.
4. Quantifying the performance of a model - The performance of a model can't be determined by a statistical criteria, it must be considered the clinical aim, as the comparison between predicted and observed probabilities for each patient or the difference between observed and predicted probabilities in group level.
5. Pre-specifying adequate performance - It is helpful to pre-specify adequate performance of a model. However it should be remembered that one feature of validation is to provide an unbiased estimate of the prediction error of the model. Consequently, the measures should be focused on quantifying the performance of a model, but ensuring that the final assessment requires clinical judgement and is context-dependent.

When evaluating a model with new data, usually only p-values are calculated and frequently it is concluded that validation is satisfactory if there is no significant difference between observed and predicted event rates. However, p-values do not provide a satisfactory answer. Even if the performance is less good, the model may still be clinically useful. The assessment of usefulness of a model requires clinical judgment and depends on context.

3.5 - Modelling strategy

Generally, prediction models may be inaccurate due to violation of assumptions, omission of important predictors, high frequency of missing data and/or improper methods, and especially with small data sets, overfitting.

Therefore, it is necessary to define a modelling strategy in order to consider the model the more accurate as possible. The modelling strategy followed in this thesis is based on a previous published methodology (Harrell Jr., Lee, Mark, 1996) and has the following steps:

1. The databases used, for training and validating the model were analysed and considered to have accurate and pertinent data, as large as possible. There were enough events captured, which are consistent with the minimum considered in (Concato John, Peduzzi, Holford, Feinstein, 1995).

2. The relevant predictors were found and to enhance the accuracy of the model, the number of variables was reduced. Some previously studies demonstrated (Harrell Jr., Lee, Mark, 1996) that the number of predictors' degrees of freedom to fit a prediction model must be less than *m/10*, where *m* represents the number of event times in the training sample. The number of predictors chosen was consistent with this number. It was used backward, forward and stepwise variable selection, using also bootstrapping techniques.
3. The relevant predictors were compared with the clinical relevant published predictors, in order to verify its consistency.
4. The missing data in both databases, one to train the model and another one to validate the model, was analysed and it was considered that imputation was the best methodology to apply. The imputed databases were also studied to verify the variables coherency.
5. The entire sample was used in the model development, using Cox Proportional Hazards and PLANN-ARD. Methods such as bootstrap and cross-validation were used to test the data set.
6. The final model developed for both methodologies (Cox Proportional Hazards and PLANA-ARD), was validated for discrimination and calibration ability.
7. An external data set was used were the model was applied and it was also studied its ability for discrimination and calibration, in order to validate it.
8. To both models Cox Proportional Hazards and PLANN-ARD, it was applied different stratification methodologies with the objective to separate the risk scores into distinct prognostic risk groups. The model was also validated plotting the observed survival for each risk group, using KM estimated survival.
9. The stratification methodologies were applied to the external data set and it was validated its ability to separate the risk scores into distinct prognostic risk groups.
10. The obtained risk groups were compared with the known breast cancer prognostic risk groups (NPI, St. Gallen and TNM) with the aim of verifying which stratification methodologies are more coherent.

Chapter 4 - Results

This chapter reports all results obtained while developing the study of this thesis. First this chapter gives a description of the two datasets used for training and external validation, respectively, including also the analysis of the missing data and the imputation results. After it describes the semi-parametric linear and neural network prognostic modelling methodologies and how multiple imputation of missing data is integrated into the non-linear modelling methodologies. It is also presented the evaluation of the predictive performance for the two alternative models, applied to an external validation data set and both methodologies are compared.

The results for the different proposed stratification methodologies are reported and compared in this chapter. In addition the two alternative models followed by the chosen stratification methodology results are compared with the existent breast cancer classification schemes. An evaluation of the results is also presented, followed by a methodology to provide confidence intervals to the individual prognostic predictions.

4.1 - Databases

There were used three data sets for this study. They comprise a retrospective longitudinal cohort study of post-operative breast cancer female patients. Two of the data sets were recruited at Christie Hospital in Wilmslow, near Manchester, where there are patients across the range from requiring no adjuvant therapy to receiving aggressive treatment. The first

cohort was recruited between 1983-1989 with a total of 917 patients and the second one between 1990-1994 with a total of 931 patients. The third data set consists of 4083 females from the British Columbia Cancer Agency (BCCA), Vancouver, during 1989-1993. It is important to mention that the last mentioned cohort is the same which was used to validate the very widely used prognostic model *Adjuvant!*. The existing patients have early, or operable breast cancer, and were selected using the standard TNM (Tumour, Nodes, Metastasis) staging system as tumour with maximum diameter less than 5 cm, node stage less than 2 (no nodes affected in the axilla or mobile nodes) and without clinical symptoms of metastatic spread. The remaining patients were therefore excluded from the study.

For Christie Hospital data sets there are 16 explanatory variables in each patient record, in addition to codes for therapy received and outcome, as we are only interested in the variables available after surgery and those that are not related with the treatment. Those related to the treatment are of no interest as they contain implicit information about the mortality risk analysed by the doctor. Consequently, these variables were not included on the model: radiotherapy, adjuvant treatment and surgery. The clinical stage variable derives from variables recorded in the dataset, so there would be some collinearity problems if we use it. Therefore, this variable has to be also excluded from the study.

The BCCA data set contains 10 explanatory variables in each patient record, as well as the outcome variables. Among the two sets of covariates there are 9 that match with identical categorization in both data sets.

The Christie Hospital data set has one record with a value of *Nodes involved* recorded in category 4. An analysis was carried out to found the sensitivity of the model selection to this record and it was found that aggregating this record into category 3 impacted on the statistical significance of more than one covariate following stepwise forward model selection with Cox regression. It was considered that this might be due to excessive leverage of this case, which were therefore treated as outlier and removed from the study.

The data from Christie Hospital contains a catch-all category labelled ‘others’ for the variable *Histological Type*, which includes also patients with “in situ” tumour. These records were excluded from the study because this category of patients has a different disease dynamic (Silverstein, Buchanan, 2003), focusing only on the histological types lobular and ductal carcinoma. The sample sizes for training and validation data sets were 743 and 4016.

Time-to-event (death by any cause) was measured in months from surgery for the Christie data set and from date of diagnosis for BCCA. There is an assumption that surgery for BBCA data set took place soon after diagnosis.

The timescale for the study is 5 years of follow-up for the Christie data set and 10 years of follow up at the end of which all surviving patients are administratively censored. The proportion of cases who were lost to follow-up before the 5 year period elapsed is around 9.8% for the Christie Hospital data set and 0.1% for the BCCA data set.

The marginal distributions are represented in tables Table 4.1, Table 4.2 and Table 4.3.

Variable description		Categories	Frequency Christie Hospital 83-89	%	Freq. BCCA	%	Frequency Christie Hospital 90-94	%
Age category	1	20-39	81	9	278	7	62	7
	2	40-59	430	47	1676	42	345	48
	3	60+	406	44	2062	51	336	45
Histological type	1	Invasive ductal	724	79	3688	92	632	85
	2	Invasive lobular/lobular in situ	95	10	328	8	111	15
	3	In situ/ mixed/ medullary/ ucoid/ papillary/ tubular/ other mixed in situ	98	11	-	-	-	-
Menopausal Status	1	Pre-Menopausal	289	31	1141	28	203	27
	2	Peri-Menopausal	47	5	5	0	27	4
	3	Post-menopausal	581	63	2758	69	513	69
	9	Missing	0	0	112	3	0	0
Histological Grade	1	Well differentiated	33	4	388	10	106	14
	2	Moderately differentiated	118	13	1772	44	298	40
	3	Poorly differentiated	89	10	1464	36	194	26
	9	Missing	677	74	392	10	145	20
Ivn	1	Positive status of lymphatics, veins and nerves	-	-	1461	36	-	-
	2	Negative status of lymphatics, veins and nerves			2323	58		
	8	Not applicable			21	1		
	9	Missing			211	5		
Nodes involved	1	0	194	21	2622	65	377	51
	2	1-3	167	18	999	25	184	25
	3	4+	50	6	389	10	97	13
	4	98 (too many to count)	1	0	-	-	-	-
	9	Missing	505	55	6	0	85	11

Table 4.1 – Variables' description and marginal distributions. These are for Christie Hospital and BCCA data set.

Variable description		Categories	Frequency Christie Hospital 83-89	%	Freq. BCCA	%	Frequency Christie Hospital 90-94	%
Nodes removed	1	0-9	725	79	1719	43	233	31
	2	10-19	111	12	1891	47	356	48
	3	20+	59	6	292	7	139	19
	4	98(too many to count)	14	2	0	0	12	2
	9	Missing	8	1	114	3	3	0
Nodes Ratio	1	0-20%	256	28	3219	80	540	73
	2	20-40%	18	2	320	8	28	4
	3	40-60%	40	4	163	4	50	7
	4	60+%	98	11	200	5	39	5
	9	Missing	505	55	114	3	86	11
Pathological Size	1	<2 cm	383	42	2170	54	413	56
	2	2-5 cm	534	58	1846	46	330	44
Oestrogen	1	0-10	242	26	892	22	145	20
	2	10+	434	47	2352	59	262	35
	9	Missing	241	26	772	19	336	45
Clinical stage	1	0	734	80	-		634	85
	2	1	183	20			109	15
Predominant site	1	Upper outer	442	48	-		376	51
	2	Lower outer	102	11			90	12
	3	Upper inner	231	25			104	14
	4	Lower inner	77	8			59	8
	5	Subareolar	56	6			43	6
	9	Missing	0	1			71	9
Side	1	Right	427	47	-		361	49
	2	Left	490	43			382	51
Maximum diameter of tumour	1	<2 cm	206	23	-		207	28
	2	2-5 cm	683	75			381	51
	3	5+ cm	2	0			1	0
	9	Missing	26	3			154	21
Clinical stage tumour	1	T1(tumour <2 cm)	213	23	-		357	48
	2	T2(2-5 cm)	704	77			386	52
Clinical stage nodes	1	N_0 (no nodes found clinically, or node negative by histological type)	734	80	-		627	84
	2	N_1(ipsilateral and mobile axillary nodes)	183	20			116	16
Local treatment	1	BCS + RT	-		2086	52	-	-
	2	BCS, no RT			0	0		
	3	Total mastectomy+RT			380	9		
	4	Total mastectomy,no RT			1550	39		

Table 4.2 – Continuation of variables' description and marginal distributions. These are for Christie Hospital and BCCA data sets.

Variable description		Categories	Frequency Christie Hospital 83-89	%	Freq. BCCA	%	Frequency Christie Hospital 90-94	%
Sys 2 (Adjuvant Systemic treatment)	0	None	736	80	1812	45	306	42
	1	Hormone alone	137	15	1232	31	345	46
	2	Chemo alone	44	5	607	15	91	12
	3	Combined hormone and Chemo treatment	0	0	365	9	1	0
Surgery	1	None	0	0	-		1	0
	2	Incision Biopsy	0	0			0	0
	3	Excision Biopsy	349	38			98	13
	4	Simple Mastectomy	383	42			18	3
	5	Radical Mastectomy	157	17			9	1
	6	Wide local excision + axillary clearance	4	0			269	36
	7	Radical Mastectomy + axillary clearance	24	3			304	41
	8	Surgery after neo adjuvant chemotherapy	0	0			43	6
	9	Missing	0	0			1	0
Adjuvant Radiation	1	No	572	62	-		361	49
	2	Yes	345	38			379	51
	3	Missing	0	0			3	0
Events at 5 years follow up (Overall mortality)			377	41	555	14	115	16
Total of records			917		4016		743	

Table 4.3 – Continuation of variables' description and marginal distributions. These are for Christie Hospital and BCCA data set.

The missing values in the data sets are an important issue to consider and should be analysed and explained, specially the changes from Christie Hospital 1983-1989 data set to Christie Hospital 1990-1994 data set.

Looking at the marginal distributions of Christie Hospital data set, acquired from 1983-1989, it can be analysed that there is a high number of missingness for *Histological grade* and *Nodes involved* variables, because it wasn't usual to acquire these variables. The proportion of Christie Hospital missing data increased in time for *Oestrogen* (19%), *Predominant Site* (8%) and *Maximum tumour diameter* (18%). On the other hand, the proportion of Christie Hospital missing data of *Histological grade* (54%), *Nodes involved* (44%) and *Nodes ratio* (44%) has decreased in time (comparing both Christie Hospital data bases).

Nodes ratio missingness have a similarity of 99,4% with *Nodes involved* missingness. This missingness is related with the category 1 of *Nodes removed* variable. *Nodes involved* missing can potentially be explained by the fact that there were so few nodes to measure for each

patient that it wasn't really measured. The category 1 of nodes removed can also represent that no nodes were actually removed.

As one of the main objectives of this study is to make a predictive model using the Christie Hospital 1990-94 data set and validate it using the BCCA data set, the marginal distribution for each variable and for both data sets should be analysed and compared. Considering that there is not a similarity on variables' marginal distribution for differences higher than 10%, the variables that can be considered less similar are *Histological grade*, *Nodes involved*, *Nodes removed* and *Oestrogen*. The higher difference is observed for *Oestrogen* category 2 and 9, with differences of 24 and 26% respectively. Analysing the *Oestrogen* survival curves for Christie 1990-94 data set, it can be observed that the missing values are related with a good outcome, and with a similar survival curve to category 2. With that assumption, the *Oestrogen* marginal distribution in both data sets would be very similar.

Analysing the missing values for both data sets it can be observed that for *Nodes involved* and *Histological grade* variables, there is respectively 11% and 10% more missing in Christie Hospital 1990-94 data set than on the BCCA data set. *Nodes involved* and *Nodes ratio* missingness is related with a good outcome (as category 1). Therefore, these records could be considered on category 1. Substituting the missing records as proposed, the marginal distribution between both data sets for these variables will be more similar.

Histological grade has 10% more of missing in Christie Hospital 1990-94 data set than for the BCCA data set. Looking at survival curves, this missingness is related with categories 2 and 3 of *histological grade*, which means that if these values were distributed between both *histological grade* categories, there is a higher frequency similarity between both data sets.

Concluding, the marginal distributions for each covariate in both data sets, Christie Hospital 1990-94 and BCCA are remarkably consistent, except for the occurrence of missing values. Therefore, the patients in both data sets have potentially the same profile, which is good for the purpose of this study.

Moreover, the KM curves from both data sets, Christie Hospital 1990-1994 and BCCA were analysed in order to verify the consistency of clinical profiles. It was observed that the *Nodes ratio, Oestrogen, Histological Grade* KM curves are very similar for both data sets, but almost all the categories have a lower survival in Christie Hospital data set. For *Nodes involved* and *Nodes removed* the main difference between both KM curves lies in the missing category. Analysing both data sets' *Histological type* and *Pathological size* KM curves, it can

be concluded that they are very similar. Looking at *Age* and *Menopausal status* KM curves it can be observed that there are some differences on their shape, for both data sets. With all these evidences it can be suggested that patients survival from both data sets, over the 5 years follow-up are highly related.

4.2 - Analysis of variables' missingness

Frequently, data on prognostic factors are missing for some patients. There are three main strategies for managing incomplete data: a new attribute may be created to denote missing, the missing values can be imputed, or the cases with missing values can be removed from the study. The latest strategy is the most widely used when modeling clinical data.

The BCCA data has a total of 4016 subjects. Excluding all missing records it will remain 2685 (411 deaths) of them, which is 67% of the entire data set. Excluding all missing records from Christie Hospital data set, from 1990-1994, which has 743 patients it would remain 265 (44 deaths) records, which is 36% of the whole data set and. If all of these missing records were excluded, valuable data would be wasted and modelling could lead to invalid results if the excluded group is a selective subset from the entire sample in respect to prognosis.

It exists several approaches for dealing with incomplete covariates in survival analysis. However, these methods all rely on the assumption that non-response probabilities do not depend on any unobserved information, that is, the data are missing at random (MAR). This property was looked for evidence in both data sets. Assessing associations between missing data and observed variables within our data sets made this evidence. While the relation between missing data and the outcome was explored using the Kaplan-Meier curves for each explanatory variable, the relationship between missing and other variables was measured using univariate and multivariate logistic regressions. KM curves were analysed, using the event of interest death attributed to any cause other than breast cancer and a follow up of 5 years.

4.2.1. Christie Hospital data set Missingness

For Christie Hospital 1990-94 data set, the variables that have missing values are *Maximum Tumour Diameter, Nodes ratio, Nodes involved and Oestrogen, Histological grade, Nodes removed* and *Predominant site*, and its KM curves are represented in Figure 4.1. These variables were analysed and it was concluded that the missing values for *Maximum Tumour*

Diameter, Nodes ratio, Nodes involved and Oestrogen are related with a better prognosis, suggesting that eliminating the patients with these missing values would lead to an underestimate of the true survival of the cohort. The opposite effect was seen for *Predominant site*.

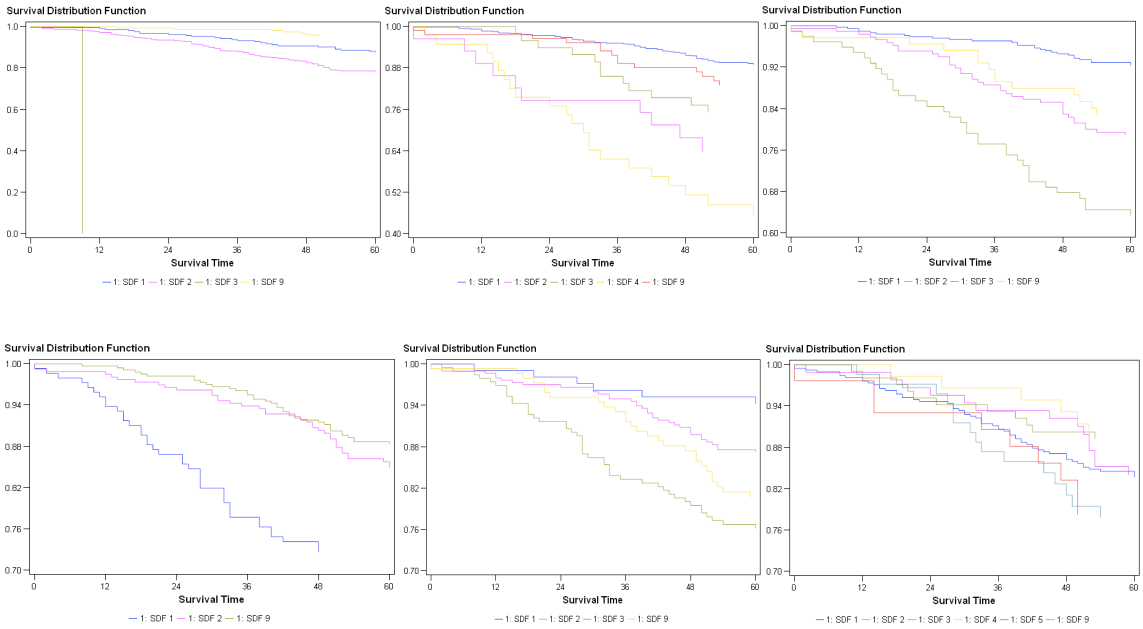

Figure 4.1 – KM curves for Christie Hospital 1990-94 data set variables'.
The top pictures represent the KM curves for Maximum Tumour Diameter, Nodes ratio and Nodes involved variables, respectively from left to right. The bottom pictures represent the KM curves for Oestrogen, Histological grade and Predominant site variables, respectively from left to right. Missing is labeled as number 9 for all variables.

In addition, the values of the potential prognostic factors for Christie Hospital data set can be related with the missingness of those factors. Looking at the cross-tabulations between the prognostic variables it can be seen that the missingness of *Oestrogen* is highly correlated with *Nodestage* categorized as 1, and the missingness of *Maximum Tumour Diameter* is highly correlated with *Pathological Size* equals caterorized as 1. It is known by the KM curves that the category 1 of *Nodestage* and *Maximum Tumour Diameter* variables' are related with a good survival. Consequently, once again it is concluded that the missingness of *Oestrogen* and *Maximum Tumour Diameter* variables' are both related with a good survival.

		Nodestage		
		1	2	Total
Oestrogen	1	112	33	145
	2	212	50	262
	Missing	303	33	336
	Total	627	116	743

		Pathological size		
		1	2	Total
Maximum Tumour Diameter	1	168	39	207
	2	115	266	381
	3	0	1	1
	Missing	130	24	154
	Total	413	330	743

Table 4.4 – Cross tabulations between some Christie Hospital variables. The left table represents Nodestage and Oestrogen variables and the right table Maximum Tumour Diameter and Pathological Size.

The same analysis was made, for comparing *Nodes ratio* with *Oestrogen*, *Node ratio* with *Histological grade* and *Node ratio* with *Maximum Tumour Diameter*. It was concluded that the missingness of *Oestrogen*, *Histological grade* and *Maximum Tumour Diameter* are all related with both category 1 and missing of *Node ratio*. Despite this fact, it was concluded that these patients are not the same, as it can be observed in the following figure, where the missing values for *Oestrogen* and *Histological Grade* are split into the diverse categories of *Maximum Tumour Diameter* variable (Figure 4.2).

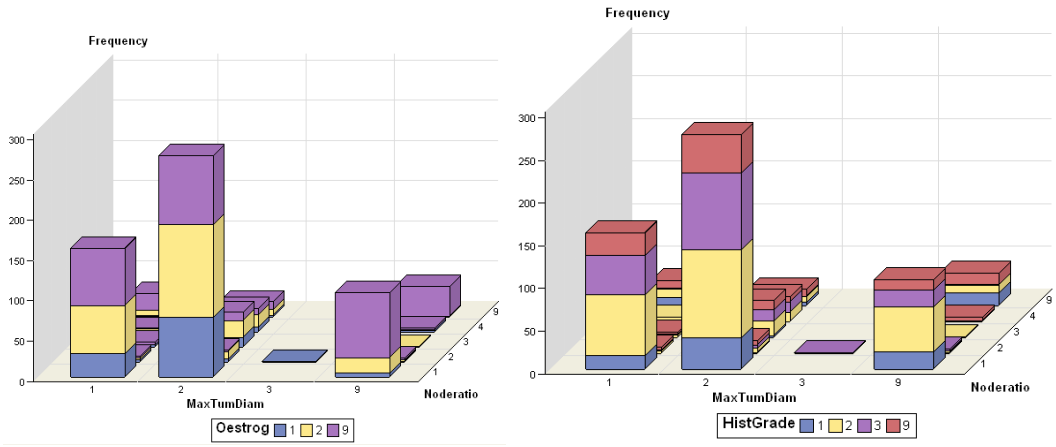

Figure 4.2 – Bar chart comparing the frequency of the categories for different variables. These variables are Oestrogen, Histological grade, Maximum tumour diameter and Nodesratio from the Chrisite Hospital data set.

Univariate and multivariate logistic models were performed for Christie Hospital 1990-1994 data set, considering the outcome missing or non-missing of a variable, with the purpose of demonstrate if the missingness of potential prognostic factors are associated with other potential prognostic factors and auxiliary variables.

There are a number of additional variables that were not considered as factors in prognostic studies, but were potentially associated with missing data, such as *surgery*, *adjuvant radiation* and *adjuvant treatment*. Table 4.5 reflects the associations between missingness and other potential prognostic factors and auxiliary variable. In Table 4.5 it can be seen that *tumour stage* and *surgery* are associated with the missingness of all but one prognostic variables, looking at the p-values (p-value < 0.05) and to the regression coefficients of each univariate logistic regression. The missingness of O*estrogen* is associated with almost all variables.

With this it can be concluded that there is evidence that missingness in Christie Hospital data set is missing at random. The associations found with the multivariate logistic regressions were expected, and there is no evidence to exclude the MAR assumption.

		Missingness of					
		Histological Grade	**Oestrogen**	**Nodes ratio**	**Maximum Tumour Diam**	**Predominant Site**	**Nodes Involved**
Association with	**Adjuvant Treatment**	p-value: 0.9944	2.068;0.957;1.533 ;1.347;-12.36; 16.080.88;-12.36 p-value: 0.0012	p-value: 0.3636	p-value: 0.0737	3.408;5.339;4.797; -6.209;-6.728; -6.728; 5.815; -6.49 p-value: 0.0003	p-value: 0.3911
	Adjuvant Radiation	4.462; 3.755 p-value: 0.0010	-0.161;0.488 p-value: <0.0001	p-value: 0.0769	3.793;4.435 p-value: 0.0027	p-value: 0.0546	p-value: 0.0930
	Surgery	11.22;-3.44;-2.977; -3.218; -4.425; -3.287;-4.553 p-value: 0.0072	-13.21;1.5; -0.22;0.23;-0.17; -0.634;-0.723 p-value: <0.0001	-4.844;10.791; 8.673;-5.086;-5.14;3.653;-5.13 p-value:<0.0001	-9.288;4.347; 2.349;3.176; 2.952;2.652; 3.099 p-value:<0.0001	p-value: 0.7389	-4.841;10.741; 8.687;-5.093; -5.143;3.666;-5.132 p-value:<0.0001
	Histological Grade	-	0.489;-0.042 p-value: 0.0007	1.498;0.132 p-value:<0.0001	0.669;-0.0039 p-value:<0.0001	p-value: 0.1601	1.466;0.149 p-value:<0.0001
	Age	p-value: 0.1843	p-value: 0.1769	p-value: 0.1179	p-value: 0.2561	p-value: 0.0751	p-value: 0.1247
	Histological Type	-3.0179 p-value: <0.0001	p-value: 0.3211	p-value: 0.1839	p-value: 0.4477	0.5772 p-value: 0.0280	p-value: 0.1648
	Max. Tum. Diameter	p-value: 0.9992	4.308;3.588 p-value: 0.0003	3.959;2.836 p-value: 0.0012	-	p-value: 0.0782	3.96;2.825 p-value: 0.0012
	Menopausal status	p-value: 0.5264	-0.3702;0.345 p-value: 0.0368	-0.542;0.146 p-value: 0.0148	-1.0875;0.4271 p-value:<0.0001	p-value: 0.4265	-0.54;0.153 p-value: 0.0165
	Manchester stage	p-value: 0.0797	0.4494 p-value: <0.0001	0.836 p-value: 0.0051	1.4128 p-value:<0.0001	p-value: 0.1258	0.8337 p-value: 0.0052
	Nodes Involved	p-value: 0.0678	0.645;-0.126 <0.0001	p-value: 0.9962	0.7096;-0.4155 p-value:<0.0001	-0.1026;0.6694 p-value: 0.0048	-
	Nodes Removed	p-value: 0.1470	0.722;-0.228;-0.399 p-value: <0.0001	p-value: 0.9989	0.379;-0.434;-0.118 p-value: 0.0012	p-value: 0.1618	p-value: 0.9989
	Node ratio	-0.512;0.2634;0.1337 p-value: 0.0488	0.641;-0.419;-0.166 p-value: 0.0041	-	0.0775	p-value: 0.8871	p-value: 1
	Node stage	p-value: 0.0617	0.423 p-value: 0.0001	0.7191 p-value: 0.0059	1.0873 p-value:<0.0001	p-value: 0.2919	0.7165 p-value: 0.0061
	Oestrogen	p-value: 0.1430	-	p-value: 0.4797	p-value: 0.3585	p-value: 0.3405	p-value: 0.4797
	Side	p-value: 0.0514	p-value: 0.285	p-value: 0.2732	p-value: 0.1393	p-value: 0.5323	p-value: 0.3286
	Site	p-value: 0.9202	p-value: 0.3189	p-value: 0.7555	p-value: 0.1051	-	p-value: 0.7052
	Pathological Size	p-value: 0.9111	0.6464 p-value: <0.0001	0.8277 p-value:<0.0001	0.8838 p-value:<0.0001	p-value: 0.8934	0.8209 p-value:<0.0001
	Tumour stage	p-value: 0.8053	0.6352 p-value: <0.0001	0.7854 p-value:<0.0001	1.8174 p-value:<0.0001	0.3127 p-value: 0.0147	0.7776 p-value:<0.0001

Table 4.5 – *Missingness associations for Christie Hospital data set.*
These associations are between missingness and other potential prognostic factors and auxiliary variables for Christie Hospital 1990-94 data set. The values represent the regression coefficients and the p-values based on a log-rank test for survival distributions in missing and non-missing groups.

4.2.2. BCCA data set Missingness

It is also essential to conclude about the type of missing mechanism in BCCA data set. In BCCA data set, an analysis of the KM curves of each prognostic variable (Figure 4.3) with missing data has shown that those patients missing *Menopausal status*, *Nodes removed*, *Node ratio* and *Oestrogen* had a better prognosis, suggesting that eliminating the patients with missing values would lead to an underestimate of the true survival of the cohort. The opposite effect was seen for *Nodes involved*. An analysis of the survival distributions of non-missing and missing data within *Ivn* and *Histological grade* factors, showed no visual of significant differences.

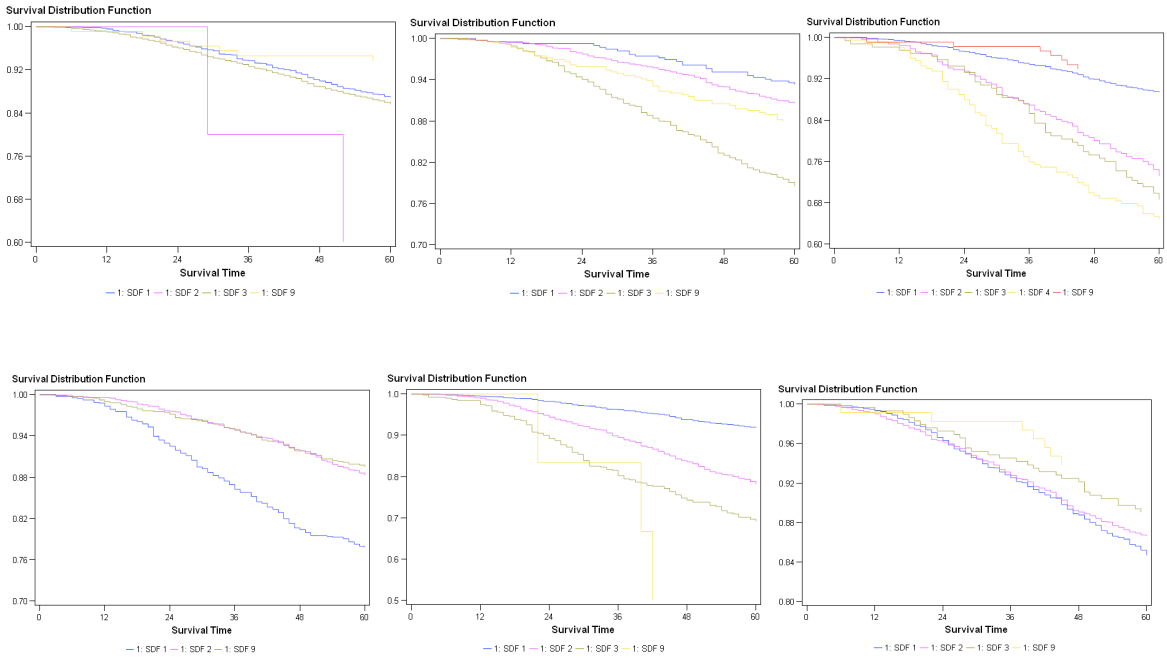

Figure 4.3 – KM curves BCCA data set variables'.
The top pictures represent the KM curves for Menopausal status, Histological grade and Nodes ratio, respectively from left to right. The bottom pictures represent the KM curves for Oestrogen, Nodes involved and Nodes removed, respectively from left to right. Missing is labeled as number 9 for all variables.

Cross-tabulations were used to conclude about the relation between the existing missing of the variables with the other variables. The patients, who have missing values on *Nodes Involved* and *Nodes Removed* variables, have also on *Nodes ratio* variables, as this was computed from the previous two referred variables. The missing values at *Oestrogen*, *Ivn* and *Menopausal Status* are related with categories 1 and 2 of *Nodes Removed* variable. However, as can be observed on the following figures (Figure 4.4), these patients are not the same.

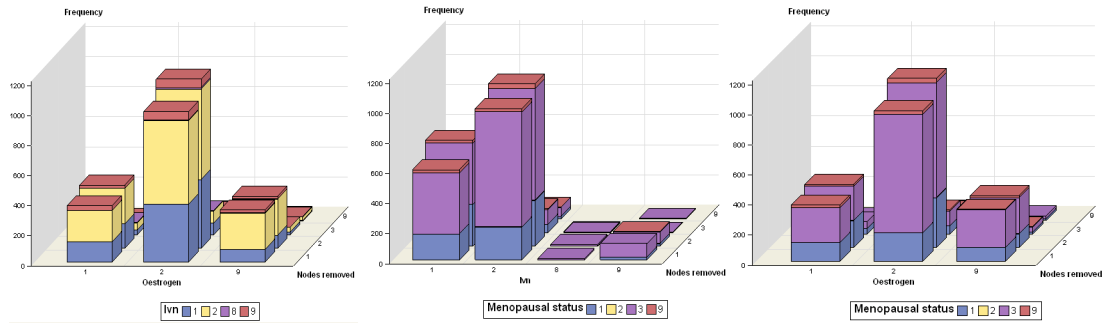

Figure 4.4 – Bar chart comparing the frequency of the categories for different variables. These variables are Oestrogen, Menopausal status, Ivn and Nodes removed from BCCA data set.

		Missingness of						
		Menopausal status	**Nodes ratio**	**Histological Grade**	**Ivn**	**Nodes involved**	**Nodes removed**	**Oestrogen**
Association with	**Sys2**	-0.2839; -0.898;0.545 p-value:<0.0001	0.478;0.129; -0.5621 p-value: 0.0222	0.104;0.484; -0.3562 p-value:<0.0001	-0.058;0.474;-0.388 p-value:0.0004	p-value:0.1795	0.478;0.129; -0.562 p-value:0.0222	0.748;-0.185; -0.22 p-value:<0.0001
	Local	p-value:0.9732	0.399;-0.274 p-value: 0.0202	-0.387;0.119 p-value:<0.0001	-0.55;0.36 p-value:<0.0001	-1.266;1.542 p-value: 0.0230	0.399;-0.234 p-value: 0.0202	0.28;-0.386 p-value: 0.0001
	Age category	0.2745;1.1188 p-value:<0.0001	p-value: 0.1273	-0.714;0.124 p-value:<0.0001	-0.317;-0.114 0.0005	p-value: 0.5814	p-value: 0.1273	-0.431;0.288 p-value: 0.0008
	Histological type	p-value: 0.7655	p-value: 0.3514	-0.3474 p-value:<0.0001	p-value: 0.9523	p-value: 0.9747	p-value: 0.3514	p-value: 0.2456
	Menopausal status	-	p-value: 0.8424	-0.986;1.3365 p-value:<0.0001	-0.909;1.129 0.0005	p-value: 0.8021	p-value: 0.8424	p-value: 0.7501
	Histological Grade	p-value: 0.3209	p-value: 0.2020	-	p-value: 0.0905	p-value: 0.3686	0.2020	0.445;-0.0224 <0.0001
	Ivn	p-value: 0.7123	p-value: 0.0779	-0.3824;-0.523 p-value: 0.016	-	p-value: 0.6350	p-value: 0.0779	-0.378;0.314 p-value:<0.0001
	Nodes involved	p-value: 0.1123	0.902;-0.518 p-value:<0.0001	p-value: 0.9987	-0.421;0.116 0.0001	-	0.902;-0.518 p-value:<0.0001	0.549;-0.169 p-value:<0.0001
	Nodes removed	p-value: 0.5268	p-value: 1	p-value: 0.2028	1.058;0.854 0.0061	p-value: 1	-	p-value: 0.3123
	Nodes Ratio	p-value: 0.5226	-	p-value: 0.4626	-0.534;0.207; 0.067 p-value:<0.0001	p-value: 1	p-value: 1	0.503;-0.089; -0.23 <0.0001
	Pathological size	p-value: 0.4990	p-value: 0.2233	p-value: 0.3825	p-value: 0.6715	p-value: 0.1056	p-value: 0.2233	0.5322 p-value:<0.0001
	Oestrogen	p-value: 0.3442	p-value: 0.1337	0.2882 p-value:<0.0001	p-value: 0.2123	p-value: 0.5363	p-value: 0.1337	-

Table 4.6 – Missingness associations for the BCCA data set. These associations are between missingness and other potential prognostic factors and auxiliary variables for BCCA data set. The values represent the regression coefficients and the p-values based on a log-rank test for survival distributions in missing and non-missing groups.

Table 4.6 reflects the associations between the missingness of some variables with other potential prognostic factors and auxiliary variable for BCCA data set. The dependent variable is, in this analysis, the presence or absence of missing and for each prognostic factor, it can be observed if it is associated with other potential prognostic factors and auxiliary variables.

Table 4.6 shows that *Sys2* (*Adjuvant systemic treatment*) and Local (local treatment) are associated with missingness of all but one prognostic variable, looking at the p-values (p-value < 0.05) and to the regression coefficients of each univariate logistic regression. This relation is expected, because these are auxillary variables related with patients' treatment, which must be related with patients' covariates. It can be seen in this table that there is at least one variable that is related with the missing of a prognostic factor. As a conclusion, there is evidence that missing mechanism in BCCA data set is missing at random (MAR). The associations found with the multivariate logistic regressions were expected and there is no evidence to exclude the MAR assumption.

4.3 - Imputation results

Whereas creating an attribute for 'missing' can be effective and has been successfully used in earlier studies with the Christie data set e.g. (Lisboa, Wong, Harris, Swindell, 2003), the structure of 'missingness' may be different in the external validation data from that in the modelling data. This is partly indicated by the markedly different prevalence of missing in the two data sets. Moreover, inferences made for individual patient cases in the future cannot make assumptions relating to the distribution of 'missingness' in the original modelling data. For these reasons, as well as the assumption of MAR evidence in the presented databases, missing data were imputed following the practice indicated in Chapter 3. Simulations studies have shown that the required number of repeated imputations *m* can be as low as three for data with 20% percent of missing data (Van Buuren, Boshuizen, Knook, 1999).

In this study, the variables which have the higher values of missing, are *Oestrogen* (with 19% for BCCA data set and 45% for Christie Hospital data set) and *Histological Grade* (with 10% for BCCA data set and 20% for Christie Hospital data set). Therefore, 10 different imputations were applied, which is a conservative choice, because unless rates of missing information are unusually high, there is little or no practical benefit to use more than 10 imputations (Schafer, 1999). It has been found that in the presence of large amounts of missing data, convergence can be obtained in as few as 10 iterations, which was the number of iterations used in this study (Van Buuren, Boshuizen, Knook, 1999). In fact, the marginal distributions following imputation with 10 and 20 iterations matched very well, with a difference by category generally less than 1% and never greater than 2% of the population for Christie Hospital data set, as it can be observed in Table 4.7. The complete summary statistics

for the datasets with imputed data for Christie Hospital 1990-94 and BCCA data sets are shown in Table 4.8.

For both data sets, the prevalence (%) of prognostic factors in the original data sets (ignoring missing data) were consistent with those from the 10 imputations. The BCCA ranges of imputation values for each potential prognostic variable are very narrow, which suggests that the distributions for each of the potential prognostic factors in the 10 imputed data sets are very similar. The Christie Hospital, has higher ranges of imputation values, which is the result of a very high missing values for some covariates.

		Christie Hospital 1990-94 completed data set (10 iterations)				**Christie Hospital 1990-94 completed data set (20 iterations)**				**Differences**
Variables		Mean #	Median	Range	%	Mean #	Median	Range	%	%
Hist. Grade	1	151	144	119-219	20	147	139	114-200	20	0
	2	387	393	318-416	52	390	397	341-422	52	0
	3	205	206	197-208	28	206	205	202-210	28	0
	9	-	-	-	-	-	-	-	-	-
Nodes ratio	1	572	579	540-616	77	560	557	540-593	75	2
	2	60	55	28-110	8	61	45	28-110	9	1
	3	59	53	50-89	8	67	54	50-116	9	1
	4	52	43	39-98	7	56	42	39-123	7	0
	9	-	-	-	-	-	-	-	-	-
Nodes involved	1	428	433	377-462	58	429	450	379-459	58	0
	2	210	227	184-237	28	212	212	186-266	28	0
	3	105	99	97-130	14	102	99	97-137	14	0
	9	-	-	-	-	-	-	-	-	-
Nodes removed	1	233	233	233-234	31	233	233	233-234	31	0
	2	357	357	356-358	48	357	357	356-358	48	0
	3	140	140	139-142	19	140	140	139-142	19	0
	4	12	12	12-13	2	13	13	12-14	2	0
	9	-	-	-	-	-	-	-	-	-
Oestrog.	1	240	247	226-275	33	253	255	234-274	34	1
	2	495	496	467-517	67	490	489	469-509	66	1
	9	-	-	-	-	-	-	-	-	
Site	1	423	425	411-430	57	417	423	396-429	56	1
	2	99	97	93-114	13	99	97	91-112	13	0
	3	111	110	106-125	15	114	111	105-132	15	0
	4	65	66	60-68	9	65	67	59-69	9	0
	5	45	45	43-47	6	48	47	43-65	6	0
	9	-	-	-	-	-	-	-	-	-
Max. Tumour Diam.	1	271	265	222-345	36	261	260	235-287	35	1
	2	428	424	388-470	58	429	424	407-479	58	0
	3	44	40	2-112	6	53	60	10-82	7	1
	9	-	-	-	-	-	-	-	-	-

Table 4.7 – Results of the missing data imputation with 10 and 20 iterations.

		BCCA Original data set		BCCA Completed data set				Christie Hospital 1990-94 original data set		Christie Hospital 1990-94 completed data set			
Variables		#	%	Mean #	Median	Range	%	#	%	Mean #	Median	Range	%
Menop. status	1	1141	29	1204	1205	1197-1209	30	-	-	-	-	-	-
	2	5	0	5	5	5-5	0	-	-	-	-	-	-
	3	2758	71	2808	2807	2802-2814	70	-	-	-	-	-	-
	9	112	0	-	-	-	-	-	-	-	-	-	-
Ivn	1	1461	38	1547	1546	1541-1555	38	-	-	-	-	-	-
	2	2323	61	2443	2445	2437-2451	61	-	-	-	-	-	-
	8	21	1	36	26	23-29	1	-	-	-	-	-	-
	9	211	0	-	-	-	-	-	-	-	-	-	-
Hist. Grade	1	388	11	461	460	445-485	12	106	18	151	144	119-219	20
	2	1772	49	1944	1945	1926-1958	48	298	50	387	393	318-416	52
	3	1464	40	1611	1613	1596-1624	40	194	32	205	206	197-208	28
	9	392	0	-	-	-	-	145	0	-	-	-	-
Nodes ratio	1	3219	83	3248	3239	3219-3323	81	540	82	572	579	540-616	77
	2	320	8	369	366	320-430	9	28	4	60	55	28-110	8
	3	163	4	168	165	163-195	4	50	8	59	53	50-89	8
	4	200	5	231	203	200-306	6	39	6	52	43	39-98	7
	9	114	0	-	-	-	-	86	0	-	-	-	-
Nodes involved	1	2622	65	2623	2623	2622-2624	65	377	57	428	433	377-462	58
	2	999	25	1004	1004	1003-1005	25	184	28	210	227	184-237	28
	3	389	10	389	389	389-389	10	97	15	105	99	97-130	14
	9	6	0	-	-	-	-	85	0	-	-	-	-
Nodes removed	1	1719	44	1803	1807	1771-1814	45	233	39	233	233	233-234	31
	2	1891	49	1907	1904	1895-1946	48	356	43	357	357	356-358	48
	3	292	7	306	307	299-310	7	139	16	140	140	139-142	19
	4	0	0	0	-	0	0	12	2	12	12	12-13	2
	9	114	0	-	-	-	-	3	0	-	-	-	-
Oestrog.	1	892	27	1098	1093	1080-1128	27	145	36	240	247	226-275	33
	2	2352	73	2918	2923	2888-2936	73	262	64	495	496	467-517	67
	9	772	0	-	-	-	-	336	0	-	-	-	-
Site	1	-	-	-	-	-	-	376	56	423	425	411-430	57
	2	-	-	-	-	-	-	90	13	99	97	93-114	13
	3	-	-	-	-	-	-	104	16	111	110	106-125	15
	4	-	-	-	-	-	-	59	9	65	66	60-68	9
	5	-	-	-	-	-	-	43	6	45	45	43-47	6
	9	-	-	-	-	-	-	71	0	-	-	-	-
Max. Tumour Diam.	1	-	-	-	-	-	-	207	35	271	265	222-345	36
	2	-	-	-	-	-	-	381	65	428	424	388-470	58
	3	-	-	-	-	-	-	1	0	44	40	2-112	6
	9	-	-	-	-	-	-	154	0	-	-	-	-

Table 4.8 – Missing data imputated compared with original data for both data sets.

Modelling using Cox proportional hazards and its validation

An analysis was carried out to verify if the prognostic factors survival distributions were consistent for each imputed data set. It was concluded that all the distributions were very similar between them and with the original data set (ignoring the missing data), which means that the imputation methodology was successfully applied. It can therefore be concluded that the survival curves for imputed data are very well defined for all 10 imputations.

The purpose of this modelling is to identify a model, which predicts the overall survival for the Christie Hospital 1990-94 data set and validate this model using the BCCA data set. As previously identified there are 9 explanatory variables with identical categorization that match in both data sets. Consequently, there is a subset of 9 variables that can be used to predict survival. These variables are *Age*, *Histological type*, *Menopausal status*, *Histological Grade*, *Nodes Involved*, *Nodes Removed*, *Nodes Ratio*, *Pathological Size* and *Oestrogen*.

All the analysis were performed and implemented in SAS software (SAS software).

4.3.1. Modelling breast cancer overall mortality using Cox proportional hazards

The event of interest used in this model was death attributed to any cause, with 5 years of follow-up. It is firstly necessary to identify the variables that predict better the breast cancer overall mortality for Christie Hospital, in order to simplify the prognostic model and enhance its accuracy.

Firstly the analysis was developed using the missing values as a different attribute. A preliminary model selection with a proportional hazards model was fitted, following a forward selection stepwise procedure, applying Akaike's information criterion (using a 95% significance level to enter in the model and 99% significance level to leave the model). This model has selected 6 explanatory variables: *Age category, Histological type, Histological grade, Nodes ratio, Oestrogen* and *Pathological size*.

The same model selection methodology was applied to the 10 imputed data sets, and the above six variable model were selected 6 times out of 10. This means that, not only the imputation is consistent, but also these variables predict survival well.

The event of interest, death attributed to any cause, can be divided in death caused by breast cancer and caused by other causes, where two models can be fitted, using the same parameters as before, in order to identify which variables are related with which events of interest. These two models were fitted using the same procedure as before and considering missing as a different attribute. It was found that for the model which event of interest is death due to breast cancer specificity, 5 explanatory variables were selected: *Nodes ratio*, *Oestrogen*, *Histological grade*, *Histological type* and *Pathological size*. The model with the events related with other causes of dead it was selected *Age category* variable as the only variable that predicts survival. Therefore, it can be concluded that the variables selected with stepwise procedure for overall mortality are the same if the two previous models are joined.

This result can suggest that the other causes of death are related with very high age and with the treatments as well. As it can be seen in Table 4.9 these events are related with the higher age category; with adjuvant radiation equals to 1 and adjuvant treatment equals to 1 and 3. With these results it can be concluded that other causes of death is related with elderly people, as the treatments are not existent or are the less aggressive ones. This means that these patients are not dying due to the applied treatment.

Variables		Categories	Number of deaths attributed to other than breast cancer
Age	1	20-39	1
	2	40-59	4
	3	60+	17
Adjuvant radiation	1	No	19
	2	Yes	3
	3	Missing	0
Adjuvant treatment	0	None	9
	1	Hormone alone	11
	2	Chemo alone	2
	3	Combined hormone and Chemo treatment	0

Table 4.9 – Relation between the event other causes of death and some variables.

The automatic routines have a number of disadvantages as they lead to the identification of a one particular subset and they depend on the stopping rule to determine whether a term should be included in or excluded from a model. The chosen variables must be carefully analysed in order to substitute, remove or introduce any variable. Therefore, there are some methods that help to obtain nearly unbiased models, such as data-splitting, cross-validation and bootstrapping. The last method used the entire dataset for model development. 1000 bootstrap re-samples were applied to the Christie Hospital 1990-94 data set with missing coded as a separate attribute and a 100 bootstraps were applied to each of the 10 imputed missing data set. Variable selection was repeated for each sample using the same stopping rule as previously, following a forward selection stepwise procedure and applying Akaike's information criterion (using a 95% significance level to enter in the model and 99% significance level to leave the model). The 10 most chosen models examples are shown in Table 4.10

Most chosen model	Nm. of times chosen	Missing coded as a different attribute	Nm. of times chosen	Imputed data sets
1	95	1,2,3,4,7,8,9	62	1,2,3,4,7,8,9
2	85	1,2,3,4,6,8,9	60	1,2,3,4,7,8
3	71	1,2,3,4,6,7,8,9	42	1,2,3,4,6,7,8,9
4	65	1,2,3,4,7,8	39	1,3,4,6,7,8,9
5	61	1,2,3,4,6,8	31	2,3,4,6,7,8,9
6	55	1,2,3,4,5,6,8,9	30	2,3,4,7,8
7	47	1,2,3,4,6,7,8	30	2,3,4,6,7,8
8	46	1,2,3,4,8,9	29	1,2,3,4,5,6,7,8 ,9
9	40	1,2,3,4,8	28	1,3,4,7,8
10	31	1,2,3,4,5,6,7,8,9	28	1,3,4,6,7,8

Table 4.10 – Models chosen using the bootstrapping ordered by their frequency.
The model characterized in green represents the model chosen with missing as a different attribute with the stepwise procedure.
Legend: 1-Oestrogen; 2- Histological grade; 3- Histological type; 4 - Node ratio; 5- Nodes removed; 6 – Nodes involved; 7 - Pathological size; 8 – Age category; 9- Menopausal status.

In this bootstrap analysis there are five variables that are almost always selected, namely *Nodes ratio, Histological grade, Histological type, Oestrogen* and *Age*. There are other 3 variables, *Pathological size*, *Nodes involved* and *Menopausal status*, that there is some doubt if they should enter in the model. However, when the variables *Nodes Involved* and *Nodes Ratio* are both in the same model, considering missing as a different attribute and the baseline model, the beta values and the standard deviation, applying Cox proportional model, are very high, as it can be observed in Table 4.11. These values mean that these two variables are very correlated, which suggests that they shouldn't be together in the model.

Categories	Nodes ratio		Nodes involved	
	Beta values	Standard error	Beta values	Standard error
1	10.48401	569.444	-11.55989	569.444
2	11.25590	569.444	-10.84243	569.444
3	10.49255	569.444	-10.52854	569.444
4	11.52662	569.444	-	-

Table 4.11 – Beta values for Cox proportional modelling
Cox proportional hazards model was fitted with variables Oestrogen, Pathological size, Histological type, Age, Histological grade, Nodes ratio and Nodes involved.

Moreover, in order to better identify which variables are really significant in predicting survival, the -2LogL statistic was compared for different predictive models, using missing as a different attribute. The results are in the following table:

			-2logL statistic	d.f.	p-value differences from 5 variables model	AIC value
Models	1	Without variables	1495,275	-	-	1495,275
	2	1,2,3,4,5	1376,203	12	-	1400,203
	3	1,2,3,4,5,6	1370,867	13	0.05	1396,867
	4	1,2,3,4,5,7	1370,483	14	<0.05	1398,483
	5	1,2,3,5,6,8	1373,907	12	-	1397,907
	6	1,2,3,4,5,6,7	1364,597	15	<0.05	1394,597

Table 4.12 – -2LogL statistic for different fitted models.
Legend: 1-Oestrogen; 2- Histological grade; 3- Histological type; 4 - Node ratio; 5-Age category ; 6 – Pathological size; 7 – Menopausal status; 8-Nodes involved.

Analysing both AIC values and the -2LogL values, there is evidence that adding the *Pathological size* variable to the model with the five variables, there is a significant reduction on their values and it can be concluded that this is also a significant model. On another hand, with the bootstrapping method, the most chosen model was the one with the five most predictive variables, adding the *Pathological size* and *Menopausal status* variables. However, from the model number 3 to model number 6 there is not a significant reduction on the AIC values that justifies adding another variable (diminution of 2,2 on 2 degrees of freedom). Even model 5 that has never been selected on the bootstrapping analysis also shows good prediction values, looking at the AIC values. Consequently, some analysis should be done for the models when missing imputation is applied.

There are therefore 4 different models that can be considered to predict this data set. Its beta parameters are listed in Table 4.13 for the imputed data set and in Table 4.14 for the non-imputed data set, where the baseline population was chosen as the one with the higher frequencies, excluding the missing categories. It can be concluded firstly that the effect of removing *menopausal status* and after *pathological size*, all betas retain the same sign. Secondly, there is not a significantly difference in their values, with the exception of *Age category,* for both imputed and non-imputed data set. Moreover, the prognostic index ranges for the four models are also very similar.

Concluding, the influence of more 1 or more 2 variables in the 5 variables model is the same for imputed and not imputed missing data. Comparing the models, one with Nodes ratio variable and the other with *Nodes involved* variable, the beta values are very similar, and there is not an evidence that one model predicts better than the other.

Furthermore, a comparison of the beta values for the model with missing as a different attribute and the model with missing imputed showed comparable values for the observed covariates, which increases our expectations that the imputed data is consistent.

Variables	Category	Betas for model: Age + Histological grade + Histological type + Oestrogen + Nodes ratio	Betas for model: Age + Histological grade + Histological type + Oestrogen + Nodes involved + Pathological size	Betas for model: Age + Histological grade + Histological type + Oestrogen + Nodes ratio + Pathological size model	Betas for model: Age + Histological grade + Histological type + Oestrogen + Nodes ratio + Pathological size model + Menopausal status
Node ratio	1	*Baseline*	*Baseline*	*Baseline*	*Baseline*
	2	1.254700391	-	1.221011	1.25291764
	3	0.831657469	-	0.685276	0.738188991
	4	1.776765282	-	1.652103	1.714117099
Histological type	1	*Baseline*	*Baseline*	*Baseline*	*Baseline*
	2	-0.759140605	-0.784716	-0.779706	-0.741725142
Histological grade	1	-0.630550611	-0.368312	-0.494438	-0.530857271
	2	*Baseline*	*Baseline*	*Baseline*	*Baseline*
	3	0.562734991	0.402875	0.516832	0.535875904
Oestrogen	1	0.610158793	0.636483	0.607958	0.627812087
	2	*Baseline*	*Baseline*	*Baseline*	*Baseline*
Age category	1	0.700142796	0.789706	0.617202	0.94465774
	2	*Baseline*	*Baseline*	*Baseline*	*Baseline*
	3	0.479431844	0.567988	0.461044	0.215964939
Path size	1	-	*Baseline*	*Baseline*	*Baseline*
	2	-	0.441709	0.613669	0.650141205
Menopausal status	1	-	-	-	-0.599008159
	2	-	-	-	0.252118407
	3	-	-	-	*Baseline*
Nodes involved	1	-	*Baseline*	-	-
	2	-	0.755296	-	-
	3	-	1.493828	-	-

Table 4.13 – Models parameters for the imputed data sets.

Variables	Category	Betas for model: Age + Histological grade + Histological type + Oestrogen + Nodes ratio	Betas for model: Age + Histological grade + Histological type + Oestrogen + Nodes involved + Pathological size	Betas for model: Age + Histological grade + Histological type + Oestrogen + Nodes ratio + Pathological size	Betas for model: Age + Histological grade + Histological type + Oestrogen + Nodes ratio + Pathological size + Menopausal status
Nodes ratio	1	*Baseline*	-	*Baseline*	*Baseline*
	2	1.41611	-	1.28206	1.34598
	3	0.73434	-	0.63693	0.74416
	4	1.83346	-	1.72628	1.82309
	9	0.72183	-	0.82950	0.88682
Histological type	1	*Baseline*	*Baseline*	*Baseline*	*Baseline*
	2	-2.03423	-2.03095	-1.99489	-2.00017
Histological grade	1	-0.63418	-0.50111	-0.55602	-0.60004
	2	*Baseline*	*Baseline*	*Baseline*	*Baseline*
	3	0.59909	0.53407	0.58127	0.61641
	9	1.26132	1.32421	1.22102	1.26599
Oestrogen	1	0.82096	0.83302	0.83004	0.87434
	2	*Baseline*	*Baseline*	*Baseline*	*Baseline*
	9	-0.01113	0.17255	0.09281	0.09722
Age category	1	0.71009	0.85100	0.64269	1.01620
	2	*Baseline*	*Baseline*	*Baseline*	*Baseline*
	3	0.45329	0.57738	0.44559	0.12114
Path size	1	-	*Baseline*	*Baseline*	*Baseline*
	2	-	0.43317	0.49708	0.52278
Menopausal status	1	-	-	-	-0.73444
	2	-	-	-	0.21092
	3	-	-	-	*Baseline*
Nodes involved	1	-	*Baseline*	-	-
	2	-	0.82883	-	-
	3	-	1.58428	-	-
	9	-	1.05793	-	-

Table 4.14 – Models parameters for the not imputed data sets.

Comparing the imputed and not imputed models betas, for the different used variables, it can be observed that they are very similar, as the sign for all of them is the same and there is no significant difference between them, with the exception of *Histological Type* variable. With this fact it can be concluded that the "missingness" influences the significance of the *Histological Type* in the model. The narrow ranges of beta values for each prognostic variable suggests that distributions for each of the potential prognostic factors in the 10 imputed data sets were similar, as it can be observed in Table 4.15. There is a very high correlation between prognostic indices for the imputed and not imputed data, as it is shown in Figure 4.5 , which suggests that the imputations make sense and are not completely at random. This analysis was developed for all previous models and the conclusions obtained were the same.

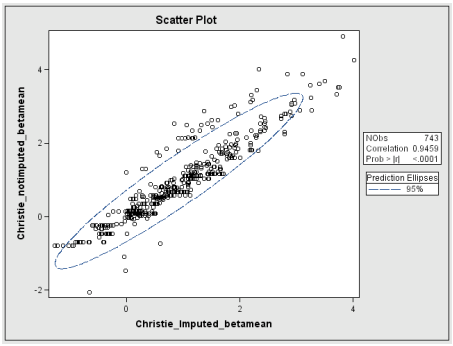

Figure 4.5 – Scatter plot between the imputed and not imputed model.
It uses Histological Grade, Histological type, Oestrogen, Age category, Pathological size and Nodes ratio as predictor variables.

To conclude about the model to use to predict breast cancer overall mortality it is necessary to verify if an addition of a variable makes a significantly better model and it is not over-fitted to the data set. Therefore, it is necessary a more carefully analysis of the identified models with an internal and external validation. However ,with the results obtained so far there is an evidence that the model that predicts better is the one with the following six variables: *Pathological size*, *Histological grade*, *Histological type*, *Oestrogen* and *Nodes ratio*. Considering this model as the most predictive one, an important requirement must be assessed, that is the ratio of events per variable. Any model based on a small number of individuals will be less reliable than one based on a larger number. However, the power of a survival analysis model is related to the number of events rather than the number of participants and simulation work has suggested that at least 10 events need to be observed for each covariate considered (Bradburn M.J., Clark, Love, Altman, 2003), (Concato John, Peduzzi, Holford, Feinstein, 1995). In our analysis there are on the training data set 115 events and 10 covariates for the 6 prognostic variables, implying 11 events per covariate, which means that the regression coefficients are possibly not biased. The confidence intervals may have the proper coverage and the test statistics might be valid for the model. Moreover, Harrell et al. (Harrel, Lee, Califf, Pryor, Rosati, 1984) suggests a rough rule of thumb that in order to have predictive discrimination that validates on a new sample, the number of covariates or degrees of freedom shouldn't be more than the number of events/10. In our analysis this rule is very approximate, as the number is 11, therefore it can be considered to have predictive discrimination.

		Missing as a different attribute				Imputed missing data					
Variables	**Category**	**Betas**	**Std error**	**Wald Chi-square**	**Significance**	**Betas**	**95% confidence limits**	**Std error**	**Min beta**	**Max beta**	**Beta Differences**
Nodes ratio				46.6	<0.0001						
	1	*Baseline*	-	-	-	*Baseline*	-	-	-	-	
	2	1.28206	0.35	13.3	0.0003	1.221011	(0.59488; 1.847143)	0.32	1.096674	1.388680	0.06105
	3	0.63693	0.33	3.8	0.0513	0.685276	(0.03403; 1.336522)	0.33	0.490181	0.835719	0.04835
	4	1.72628	0.27	40.8	<0.0001	1.652103	(1.07537; 2.228833)	0.29	1.384110	1.827760	0.07418
	9	0.82950	0.34	6.1	0.0239	-	-	-	-	-	-
Hist. type				21.2	<0.0001						
	1	*Baseline*		-	-	*Baseline*	-	-	-	-	-
	2	-1.99489	0.43	21.2	<0.0001	-0.779706	(-1.63886; 0.079448)	0.44	-0.981595	-0.333998	1.2152
Hist. grade				23.5	<0.0001						
	1	-0.55602	0.45	1.5	0.2186	-0.494438	(-1.43704; 0.448162)	0.47	-0.835856	-0.112374	0.06158
	2	*Baseline*	-	-	-	*Baseline*	-	-	-	-	-
	3	0.58127	0.24	6.1	0.0135	0.516832	(0.04618; 0.987483)	0.24	0.324459	0.651287	0.06444
	9	1.22102	0.30	16.7	<0.0001	-	-	-	-	-	-
Oest.				13.6	0.0011						
	1	0.83004	0.25	11.4	0.0007	0.607958	(0.09949; 1.116421)	0.26	0.394441	0.917643	0.22208
	2	*Baseline*	-	-	-	*Baseline*	-	-	-	-	-
	9	0.09281	0.25	0.14	0.7067	-	-	-	-	-	-
Age				6.7	0.0346						
	1	0.64269	0.29	5.1	0.0246	0.617202	(0.05059; 1.183818)	0.29	0.544848	0.693451	0.02549
	2	*Baseline*	-	-	-	*Baseline*	-	-	-	-	-
	3	0.44559	0.21	4.4	0.0366	0.461044	(0.03031;0.891774)	0.22	0.366594	0.531312	0.01545
Path. size				5.2	0.0231						
	1	*Baseline*	-	-	-	*Baseline*	-	-	-	-	-
	2	0.49708	0.22	5.2	0.0231	0.613669	(0.17458;1.052756)	0.22	0.504409	0.784853	0.11659

Table 4.15 – Beta values comparison for imputed and not imputed model.
The model was fitted with variables Histological Grade, Histological type, Oestrogen, Age category, Pathological size and Nodes ratio.

Furthermore, the three chosen variables *Pathological size, Nodes ratio* and *Histological grade* reflect the known Nottingham Prognostic Index (NPI), where the node status is replaced by *Nodes ratio. Histological type* and *Oestrogen* relate to a coarse sub-typing of the disease and response to therapy, respectively. *Age category* was selected because it is related with the event of interest, death attributed to any cause other than breast cancer.

4.3.2. Sensitivity analysis of Nodes involved variable

The original data set, removing category 3 from histological type, comprised for 1 record for category 4 of *Nodes involved* variable, for Christie Hospital data set and therefore it was carried out a study in order to analyse the sensitivity of this category to the data set and to the model. The strategy to deal with this record must be one of the following: remove this record and category from the data set; recode this record as category 3 or leave this record and category as it is in the original data set.

The imputation methodology was also performed leaving this record on category 4, and it was concluded that the imputation methodology was successfully applied, with the exception of *Nodes involved* variable. This is a consequence of the fact that there exists only 1 record for category 4 of this variable. When the missing is imputed, this category can have from 1 to 51 subjects which compromises the survival curves, being very different from each other and from the survival curves without imputing. These results suggest that the category 4 of *Nodes Involved* variable should be recoded or eliminated, because it can compromise the analysis. The survival curves for the 10 imputations were compared in order to verify the convergence of imputation and the impact of category 4 of *Nodes involved*. Removing the record form the data set, it was concluded that there were significant differences on the survival curves for *Nodes involved* and *Nodes ratio* variable. However, if this record is removed, these survival curves become more consistent for all the 10 different imputations. Generally it can be concluded that the survival curves for imputed data are very well defined for all imputations, when the 2 records are removed from the data set.

Ahead there are the analysis of the model without these two subjects and coding them as *Nodes involved* equals to 3 in order to conclude about what to use in this modelling selection.

The bootstrap methodology applied previously in order to detect the most predictive model, was also applied coding the record being analysed as category 3 and leaving it as category 4, for both imputed and not imputed data sets.

1000 bootstrap re-samples were applied to the Christie Hospital 1990-94 data set, coding missing as a separate attribute and a 100 bootstraps were applied to each of the 10 imputed missing data set. Variable selection was repeated for each sample, using the same stopping rule, with a proportional hazards model, following a forward selection stepwise procedure and applying Akaike's information criterion (using a 95% significance level to enter in the model

and 99% significance level to leave the model). The first 10 most chosen models are shown in Table 4.17 and Table 4.18.

		Christie Hospital 1990-94 removing category 4 from Nodes involved variable				Christie Hospital 1990-94 original data set		Christie Hospital 1990-94 completed data set			
Variable		Mean #	Median	Range	%	#	%	Mean #	Median	Range	%
Nodes involved	1	428	433	377-462	58	377	57	415	421	378-453	56
	2	210	227	184-237	28	184	28	204	206	184-248	27
	3	105	99	97-130	14	97	15	100	98	97-112	13
	4	-	-	-	-	1	0	24	15	1-51	4
	9	-	-	-	-	85	0	-	-	-	-

Table 4.16 – Imputation results comparison.
This analysis was performed removing category 4 from Nodes involved variable and leaving the variable as it is in the original data set.

Most chosen model	Nm. of times chosen	Nodes involved coded as 4	Nm. of times chosen	Nodes involved coded as 3	Nm. of times chosen	Nodes involved removed
1	101	1,2,3,4,6,8,9	83	1,2,3,4,7,8,9	95	1,2,3,4,7,8,9
2	89	1,2,3,4,6,7,8,9	65	1,2,3,4,6,8,9	85	1,2,3,4,6,8,9
3	67	1,2,3,4,6,8	62	1,2,3,4,5,6,8,9	71	1,2,3,4,6,7,8,9
4	65	1,2,3,4,7,8,9	61	1,2,3,4,6,7,8,9	65	1,2,3,4,7,8
5	58	1,2,3,4,5,6,8,9	53	1,2,3,4,6,8	61	1,2,3,4,6,8
6	46	1,2,3,4,6,7,8	52	1,2,3,4,7,8	55	1,2,3,4,5,6,8,9
7	45	1,2,3,4,8	42	1,2,3,4,8,9	47	1,2,3,4,6,7,8
8	34	1,2,3,4,5,6,7,8,9	39	1,2,3,4,5,6,8	46	1,2,3,4,8,9
9	33	1,2,3,4,8,9	39	1,2,3,4,5,6,7,8,9	40	1,2,3,4,8
10	29	1,2,3,4,7,8	35	1,2,3,4,8	31	1,2,3,4,5,6,7,8,9

Table 4.17 – Models chosen using bootstrapping, coding missing as a different attribute.
The model characterized in green represents the first model chosen with missing as a different attribute. Legend: 1-Oestrogen; 2- Histological grade; 3- Histological type; 4 - Node ratio; 5- Nodes removed; 6 – Nodes involved; 7 - Pathological size; 8 – Age category; 9- Menopausal status.

Although it appears that the coding of *Nodes involved* variable does not interfere in model selection, the obtained results show a different perspective. Actually, with the obtained results it can be concluded that the model selection is more consistent when the category 4 of *Nodes involved* variable is deleted from the data set, because there is less noise in the data set and in model selection. Furthermore, using the imputed data sets in model selection, there is more consistency in variables selection when this category is deleted, as the variables selected without the bootstrapping are in the second place when there is a ranking of all models. This

indicates a high sensitivity of the models to this record and category; therefore the record was removed from the modelling data, as outlier.

Most chosen model	Nm. of times chosen	Nodes involved coded as 4 - Imputed	Nm. of times chosen	Nodes involved coded as 3 - Imputed	Nm. of times chosen	Nodes involved removed - Imputed
1	57	1,2,3,4,6,8,9	67	1,3,4,6,7,8	62	1,2,3,4,7,8,9
2	51	1,2,3,4,5,6,8,9	56	1,3,4,5,6,7,8	60	1,2,3,4,7,8
3	47	1,2,3,4,6,8	53	1,3,4,5,6,7,8,9	42	1,2,3,4,6,7,8,9
4	47	1,2,3,4,6,7,8,9	50	1,2,3,4,5,6,7,8,9	39	1,3,4,6,7,8,9
5	44	1,2,3,4,5,6,7,8,9	44	1,2,3,4,6,7,8,9	31	2,3,4,6,7,8,9
6	36	1,2,3,4,6,7,8	40	1,2,3,4,7,8,9	30	2,3,4,7,8
7	32	1,3,4,6,7,8,9	39	1,2,3,4,7,8	30	2,3,4,6,7,8
8	31	1,2,3,4,7,8	36	1,2,3,4,5,6,8,9	29	1,2,3,4,5,6,7,8,9
9	30	2,3,4,6,7,8	35	1,2,3,4,6,8,9	28	1,3,4,7,8
10	29	2,3,4,6,7,8,9	34	1,2,3,4,6,8	28	1,3,4,6,7,8

Table 4.18 – Models chosen using bootstrapping applied to the imputed data sets. The model characterized in green represents the first model chosen with missing as a different attribute. Legend: 1-Oestrogen; 2- Histological grade; 3- Histological type; 4 - Node ratio; 5- Nodes removed; 6 – Nodes involved; 7 - Pathological size; 8 – Age category; 9- Menopausal status.

4.3.3. Cox Proportional hazards model validation

There are a few ways to assess or validate the performance of a prognostic model, that can be divided in internal and external validation. Firstly, the Cox proportional hazards underwent an internal validation, that is model was developed in a portion of a data set and it was applied to the other portion of the data set. The model was developed with missing as a different attribute, and it a 10 fold cross validation of the data was used. A random 74 subjects was left out each time and 10 models were developed, with the variables already found as significant, *Histological type, Histological grade, Pathological size, Node ratio, Oestrogen* and *Age category,* on the 669 subjects. The prognostic index (βx) was calculated for the leaving out random subjects, with the beta values modelled with the remaining subjects. Figure 4.6 represents a scatterplot between the prognostic indices calculated with the model developed for the entire subjects and the prognostic indices calculated using cross-validation.

This figure shows that both prognostic indices for the same records have a high correlation, concluding that the model is very well fitted. It can be observed that there are some outliers, mainly for the Prognostic indices upper 4, for the cross validation. However, these outliers are

due to the fact that some beta values of the 10 produced models are very high or low, resulting on the few records for each training model. It can also be observed that the higher correlation is for the 6 variables model trained using *Nodes Ratio* variable.

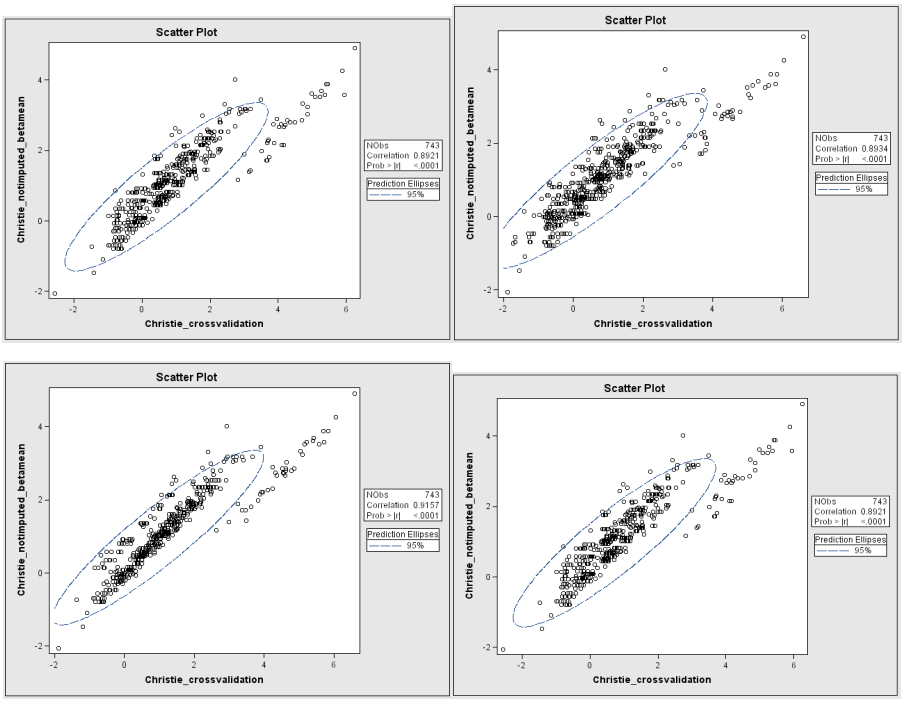

Figure 4.6 – Scatter plot between the cross-validated PI and the PI not cross-validated. The upper left picture represents the 5 variables model and the upper right picture represents the 7 variables model. The bottom left picture represents the 6 variable model developed with Nodes ratio and the bottom right picture represents the 6 variables model developed with Nodes involved.

Other analysis was produced using the imputed training data sets and cross-validation methodology. For each imputed data set a different model was built. However, instead of building a model for the 10 imputed data sets, it was built one for each 9 imputed data sets. This model was then applied to the remaining imputed data set. At the end, all the imputed data sets were applied to these different produced models. A prognostic index was obtained for each patient and compared with the prognostic index obtained with the model constructed with all the imputed data sets, as it can be observed on Figure 4.7. This figure shows that both prognostic indices for the same records have a high correlation for the different 4 models, concluding that all models are very well fitted.

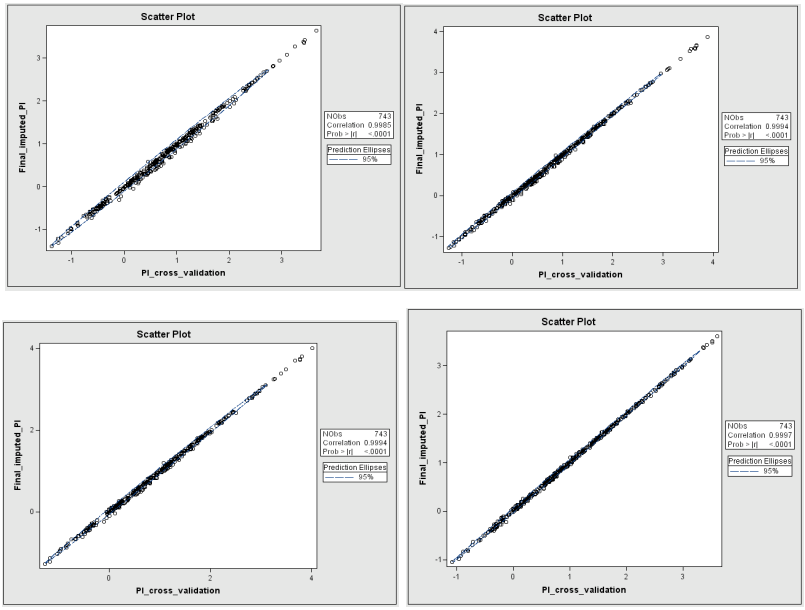

Figure 4.7 – Scatter plot between different prognostic indexes.
These are the final prognostic index developed with the 10 imputed data sets and the prognostic index obtained by cross-validating the imputed data sets. The upper left picture represents the 5 variables model and the upper right picture represents the 7 variables model. The bottom left picture represents the 6 variables model developed with Nodes ratio and the bottom right picture represents the 6 variables model developed with Nodes involved.

The consistency of the imputed data has already been demonstrated. Therefore, it is important to validate the performance of the prognostic model with the imputed data, assessing its calibration and discrimination. Calibration compares the observed and predicted event rates for group of patients and discrimination, quantifies the model's ability to distinguish between patients who do or do not experience the event of interest. The C^{td} values and the Hosmer-Lemeshow statistic were obtained for the 4 models in discussion, in order to assess the discrimination and calibration of the models, respectively.

C^{td} values as well as the standard deviation in square brackets, for each discrete time, are represented in the first line of Table 4.19. The Hosmer-Lemeshow χ^2 statistic is also represented in Table 4.19 , next to the C^{td} index, with associated p-values in square brackets. For the calculation of Hosmer-Lemeshow statistic, the patients were grouped in 10 classes. It must be regarded that large values of χ^2 statistic and small p-values indicate a lack of fit of the model. The calibration plots evaluated by grouping subjects according to the predicted survival, S(t), at t=1,2,3,4 and 5 years are shown in Figure 4.8 .

		Years				
		1	**2**	**3**	**4**	**5**
Models	**Age + Histological grade + Histological type + Oestrogen + Nodes ratio**	0.75(0.144); 0.11[1]	0.78(0.085); 0.49[1]	0.81(0.057); 1.68[0.996]	0.78(0.049); 2.67[0.976]	0.75(0.046); 2.50[0.981]
	Betas for model: Age + Histological grade + Histological type + Oestrogen + Nodes involved + Pathological size	0.83(0.117); 0.1[1]	0.82(0.059); 0.2[1]	0.83(0.044); 0.9[1]	0.83(0.037); 1.21[0.999]	0.80(0.039); 1.5 [0.997]
	Betas for model: Age + Histological grade + Histological type + Oestrogen + Nodes ratio + Pathological size	0.86(0.091); 0.06[1]	0.85(0.06); 0.26[1]	0.85(0.046); 0.95 [1]	0.83(0.041); 1.31[0.998]	0.81(0.041); 2.81 [0.973]
	Betas for model: Age + Histological grade + Histological type + Oestrogen + Nodes ratio + Pathological size model + Menopausal status	0.80(0.125); 0.01[1]	0.82(0.076); 0.49[1]	0.82(0.055); 2.04[0.991]	0.79(0.048); 1.05[0.999]	0.77(0.045); 0.94[1]

Table 4.19 – Models Calibration and discrimination assessment.
It was performed for the 4 different models, using the training data set and imputation methodology.

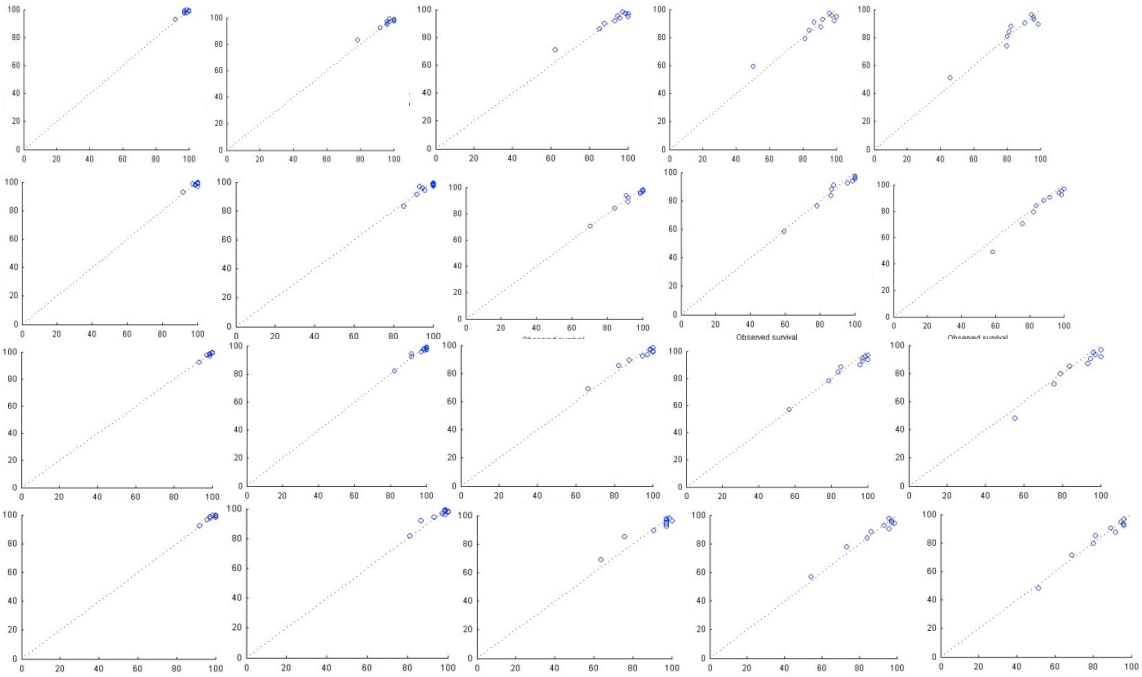

Figure 4.8 – Calibration plots for the 4 different models for the training data set.
From left to right calibration is at t=1,2,3,4,5 years. The x axis is the observed survival and the y axis is the predicted survival. The upper pictures represent the calibration plots for the 5 variables model, the second upper pictures represent the calibration plots for the 6 variables model developed using Nodes involved variable, the third upper pictures represent the calibration plots for the 6 variables model developed using Nodes ratio variable and the bottom pictures represent the calibration plots for the 7 variables model.

From the C^{td} values for all models, it can be seen that generally, a satisfactory performance with values of about 0.8 was reached. However, the discrimination of both 6 variables model is much better than the other two models. In adittion, the model which includes the *Node Ratio* variable has better discrimination than the one that includes the *Nodes Involved* variable, that increases our expectation that the model comprised by *Histological Grade*, *Age category*, *Histological type*, *Oestrogen*, *Pathological size* and *Nodes Tatio* is the one that predicts better. From the graphical inspection of the calibration plots, none of the models showed any major tendency to systematic over/under estimation of the observed survival in groups. Moreover, calibration is better in groups with high-predicted survival, for all models. It can be inspected that all models have a very good calibration performance.

Even internal validation is essential and produces a conclusion about the model validation; external validation a more efficient approach. In order to demonstrate the prognostic model is valuable, it is not sufficient to show that it successfully predicts outcome in the initial development data. It is necessary to prove that the model performs well for other group of patients. The external validation examines the generalisability of the model, for which it is required a new data collected from an appropriate (similar) patient population in a different centre. Therefore, the BCCA data set was used to validate externally all the models developed and considered previously. All the 4 different models were subjected to validation, with the purpose of verifying the conclusions obtained with the training data set. Once obtained the betas and the prognostic indexes for the derivation data set, the 5-year estimated prognostic index was calculated for each patient in the imputed validation data sets. The final computed beta values were applied to each imputed data set (from the validation cohort), obtaining 10 prognostic indexes for each patient, which were then averaged in order to obtain a single prognostic risk. The discrimination and calibration were assessed for the validated model, using the C^{td} values and the Hosmer-Lemeshow statistic, respectively for the different 4 models in discussion. The C^{td} values, as well as the standard deviation in brackets, for each discrete time, are represented in the first line of Table 4.20, as it was represented for the training data set. The Hosmer-Lemesow statistic is presented on Table 4.20 , next to the C^{td} index, with the χ^2 value and the p-value in square brackets. The calibration plots evaluated by grouping subjects according to the predicted survival, S(t), at t=1,2,3,4 and 5 years are shown in Figure 4.9 .

		Years				
		1	**2**	**3**	**4**	**5**
Models	**Age + Histological grade + Histological type + Oestrogen + Nodes ratio**	0.72(0.09); 0.35[1]	0.73(0.039); 0.83[1]	0.74(0.028); 1.98[0.992]	0.71(0.024); 4.91[0.842]	0.70(0.022); 8[0.534]
	Betas for model: Age + Histological grade + Histological type + Oestrogen + Nodes involved + Pathological size	0.74(0.090); 0.04[1]	0.74(0.040); 0.36[1]	0.73(0.028); 1.18[0.999]	0.71(0.024); 2.09[0.990]	0.69(0.022); 2.84[0.970]
	Betas for model: Age + Histological grade + Histological type + Oestrogen + Nodes ratio + Pathological size	0.74(0.090); 0.38 [1]	0.75(0.039); 0.79[1]	0.75(0.027); 2.09 [0.99]	0.74(0.023); 5.28[0.808]	0.72(0.021); 8.39[0.495]
	Betas for model: Age + Histological grade + Histological type + Oestrogen + Nodes ratio + Pathological size model + Menopausal status	0.75(0.087); 0.42[1]	0.76(0.038); 0.99[0.999]	0.75(0.027); 2.32[0.985]	0.74(0.023); 5.66[0.773]	0.72(0.021); 8.65[0.471]

Table 4.20 – Models Calibration and discrimination assessment.
It was performed for the 4 different models, using the validation data set and imputation methodology.

From the C^{td}, it was reached a satisfactory performance of about 0.7, for all models, which slightly lower than the one obtained with the training data set (0.8). However, this lowering was already been expected. The 6 variables model, fitted with *Nodes Ratio* and the 7 variables model have a slightly better discrimination performance than both the 5 variables model and the 5 variables model fitted with *Nodes Involved*. Analysing the Chi-square values, as well as the p-values, it can be verified that all models have a very good calibration performance and, as it was expected, the calibration performance is lower than the trained models. From the graphical inspection of the calibration plots it can be concluded that for all models, with the exception of the 6 variables models fitted with *Nodes Involved*, there is an underestimation of the observed survival in the groups, more accentuated in groups with low predicted survival and for time equals to 3, 4 and 5 years. Moreover, calibration is better in groups with high-predicted survival, for all models.

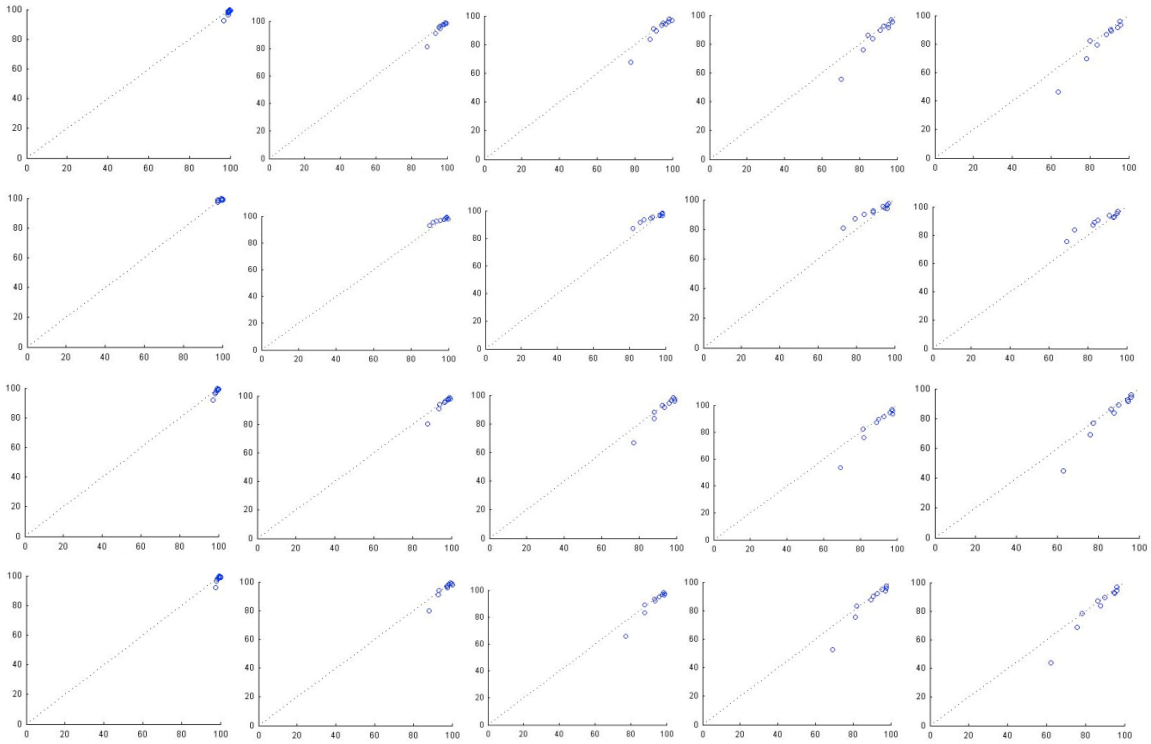

Figure 4.9 – Calibration plots for the 4 different models, for the validation data set. From left to right calibration is at t=1,2,3,4,5 years. The x axis is the observed survival and the y axis is the predicted survival. The upper pictures represent the calibration plots for the 5 variables model, the second upper pictures represent the calibration plots for the 6 variables model developed using Nodes involved variable, the third upper pictures represent the calibration plots for the 6 variables model developed using Nodes ratio variable and the bottom pictures represent the calibration plots for the 7 variables model.

4.4 - PLANN-ARD Modelling and its validation

Partial Logistic Artificial Neural network with Automatic Relevance Determination (PLANN-ARD) model accounts implicitly for non-linear and non proportional covariate effects. The neural network does not seek merely to explain the observed variation in survival, as a function of covariate effects. Instead, it fits the hazard function directly, without resort to proportionality assumptions about the covariate effects. In this way, it is suited to making individual predictions of the event rate. The event of interest and the follow up time is the same as it was used in Cox Proportional hazards modelling.

As a consequence of imputation, PLANN-ARD was run 10 times, one for each imputed data set, resulting in 10 different trained networks. PLANN-ARD was trained using the variables previously selected. Using the PLANN-ARD, the mean of the hazard of the 10 imputed data sets can be computed as in equation 53.

According to the discrete time perspective, a prognostic index can be defined to enable

each patient to be allocated into a risk group, using the 5-year survival as the stratification index. This was defined using the identity identified in equation 52.

Similarly to Cox proportional hazards modelling, a careful validation was carried out for the PLANN-ARD modelling, also divided in internal and external validation. Calibration and discrimination performance was assessed for all the 4 different models. Both training and validation data sets were used to carry out this validation. Table 4.21, Table 4.22, Figure 4.10 and Figure 4.11 represent all the obtained values for the training and validation data set , respectively.

PLANN-ARD discrimination was measured through the C^{td} values. A satisfactory performance was reached, for all models, with values of about 0.8 for the training data set, with the exception of the 5 variables model, reaching that performance only at 3 years. The validation data set reached a value approximately of 0.7. For the training data set, it can be analysed, not only that the discrimination for both 6 variables models is much better than for the other 2 models, but also that the model which includes the *Node Ratio* variable has better discrimination than the one that includes the *Nodes Involved* variable. This increases our expectation that the model comprised by *Histological Grade*, *Age category*, *Histological type*, *Oestrogen*, *Pathological size* and *Nodes Ratio* is the one that predicts better.

	Years				
	1	**2**	**3**	**4**	**5**
Age + Histological grade + Histological type + Oestrogen + Nodes ratio	0.77(0.138); 0.14[1]	0.79(0.081); 0.33[1]	0.81(0.055); 2.56[0.979]	0.79(0.048); 2.65[0.977]	0.76(0.046); 2.15[0.989]
Betas for model: Age + Histological grade + Histological type + Oestrogen + Nodes ratio + Pathological size	0.86 (0.093); 0.05[1]	0.85 (0.062); 0.26[1]	0.85(0.046); 1.65[0.996]	0.83(0.041); 1.28[0.998]	0.81(0.041); 1.49[0.997]
Betas for model: Age + Histological grade + Histological type + Oestrogen + Nodes involved + Pathological size	0.83(0.117); 0.05[1]	0.84(0.055); 0.28[1]	0.84(0.042); 0.67[1]	0.83(0.037); 1.68[0.996]	0.80(0.038); 1.94[0.992]
Betas for model: Age + Histological grade + Histological type + Oestrogen + Nodes ratio + Pathological size model + Menopausal status	0.81(0.114); 0.14[1]	0.83(0.067); 0.74[1]	0.83(0.050); 2.69[0.975]	0.80(0.045); 2.78[0.972]	0.77(0.043); 1.88[0.993]

Table 4.21– Models Calibration and discrimination assessment.
It was performed for the 4 different models, using the training data set and imputation methodology.

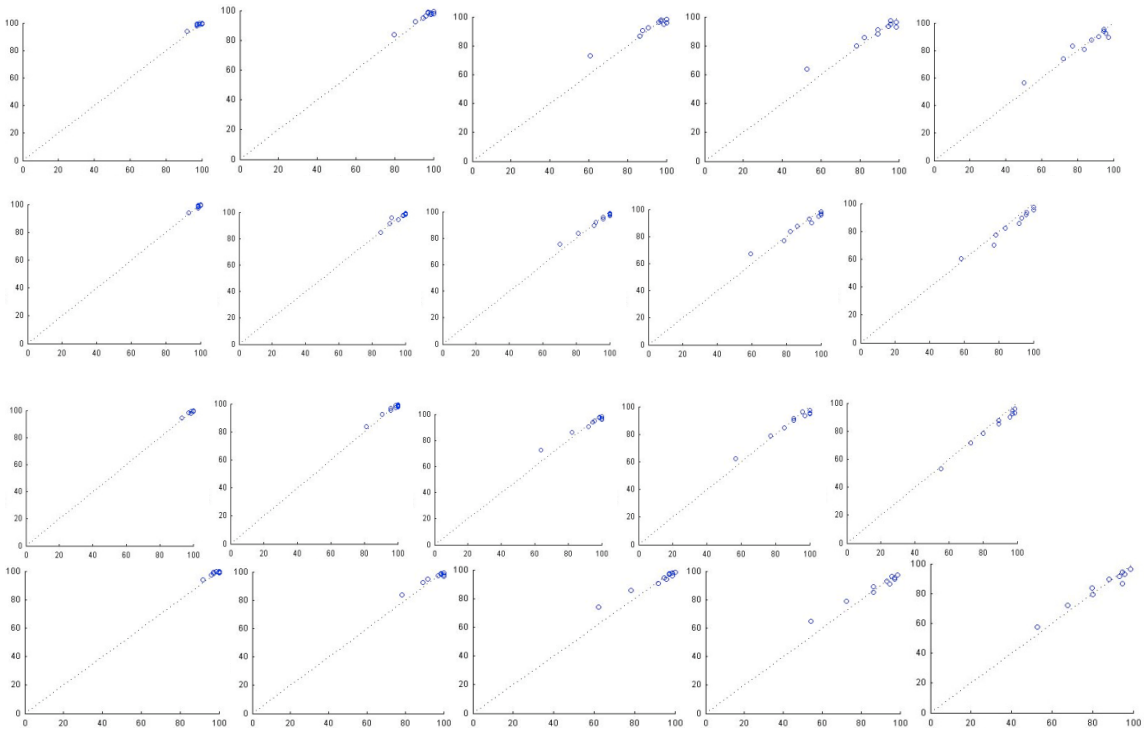

Figure 4.10 – Calibration plots for the 4 different models, for the training data set. From left to right calibration is at t=1,2,3,4,5 years. The x axis is the observed survival and the y axis is the predicted survival. The upper pictures represent the calibration plots for the 5 variables model, the second upper pictures represent the calibration plots for the 6 variables model developed using Nodes involved variable, the third upper pictures represent the calibration plots for the 6 variables model developed using Nodes ratio variable and the bottom pictures represent the calibration plots for the 7 variables model.

		Years				
		1	**2**	**3**	**4**	**5**
Models	**Age + Histological grade + Histological type + Oestrogen + Nodes ratio**	0.73 (0.089); 0.2 [1]	0.73 (0.040); 0.68 [1]	0.73 (0.028); 1.29 [0.998]	0.72 (0.024); 1.96 [0.992]	0.70 (0.022); 3.56 [0.938]
	Betas for model: Age + Histological grade + Histological type + Oestrogen + Nodes ratio + Pathological size	0.75 (0.086); 0.08 [1]	0.75(0.039); 0.22 [1]	0.75(0.028); 0.53 [1]	0.74(0.023); 0.73 [1]	0.72(0.021); 1.23 [0.999]
	Betas for model: Age + Histological grade + Histological type + Oestrogen + Nodes involved + Pathological size	0.77 (0.091); 0.05 [1]	0.78 (0.038); 0.12 [1]	0.77 (0.027); 0.21 [1]	0.75 (0.022); 0.31[1]	0.74 (0.020); 0.36[1]
	Betas for model: Age + Histological grade + Histological type + Oestrogen + Nodes ratio + Pathological size model + Menopausal status	0.77(0.080); 0.20 [1]	0.76 (0.038); 0.36 [1]	0.76 (0.027); 0.63[1]	0.74 (0.023); 1.27[0.999]	0.73 (0.021); 1.96 [0.992]

Table 4.22 – Models Calibration and discrimination assessment. It was performed for the 4 different models, using the validation data set and imputation methodology.

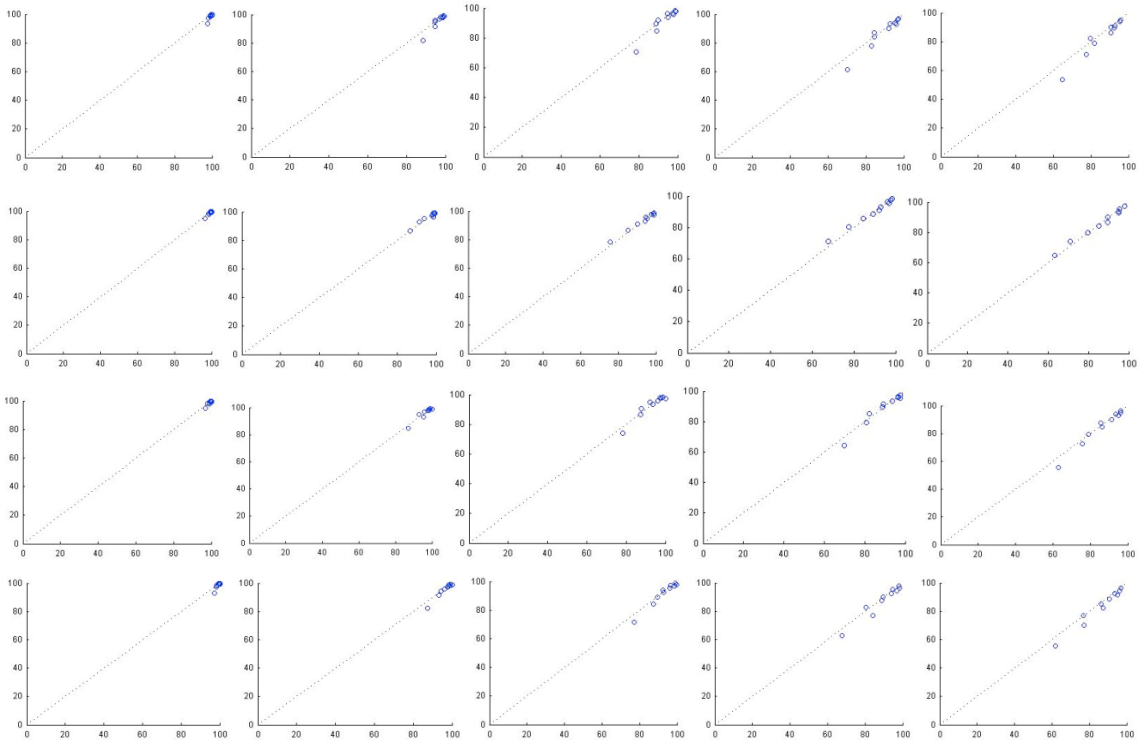

Figure 4.11 – Calibration plots for the 4 different models, for the validation data set. From left to right calibration is at t=1,2,3,4,5 years. The x axis is the observed survival and the y axis is the predicted survival. The upper pictures represent the calibration plots for the 5 variables model, the second upper pictures represent the calibration plots for the 6 variables model developed using Nodes involved variable, the third upper pictures represent the calibration plots for the 6 variables model developed using Nodes ratio variable and the bottom pictures represent the calibration plots for the 7 variables model.

PLANN-ARD calibration was achieved using the Hosmer-Lemeshow statistic and the calibration plots at t=1,2,3,4 and 5 years. From the Hosmer-Lemeshow statistic all models show an excellent calibration, for the training and validation data set. From the graphical inspection of the calibration plots, none of the models showed any major tendency to systematic over/under estimation of the observed survival in groups, for the training data set. Observing the calibration plots for the validation data set it can be concluded that for all models there is an underestimation of the observed survival in the groups, with the exception of the 6 variables model with *Nodes Involved*. Moreover, calibration is better in groups with high-predicted survival, for all models, and for the training and validation data set. Therefore, it can be concluded that all models have a very good calibration performance.

4.5 - Comparison between Cox and PLANN-ARD modelling

The performance of the models developed with Cox proportional hazards can be compared with the performance of the models developed with PLANN-ARD, for both training and validation data set.

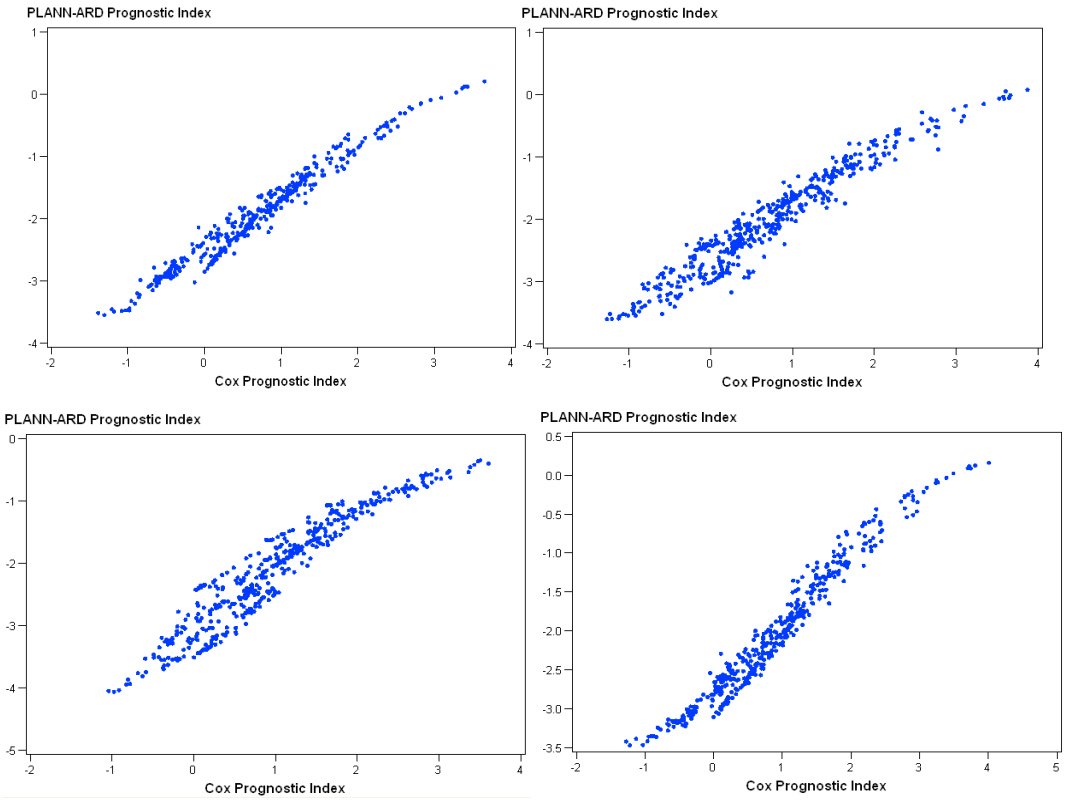

Figure 4.12 – Comparison between Cox and PLANN-ARD PI for the training data set. The top figures represent the models developed with 5 and 7 variables, from left to right respectively. The bottom pictures represent the model developed for the 6 variables models, with Nodes involved and Nodes ratio, from left to right respectively.

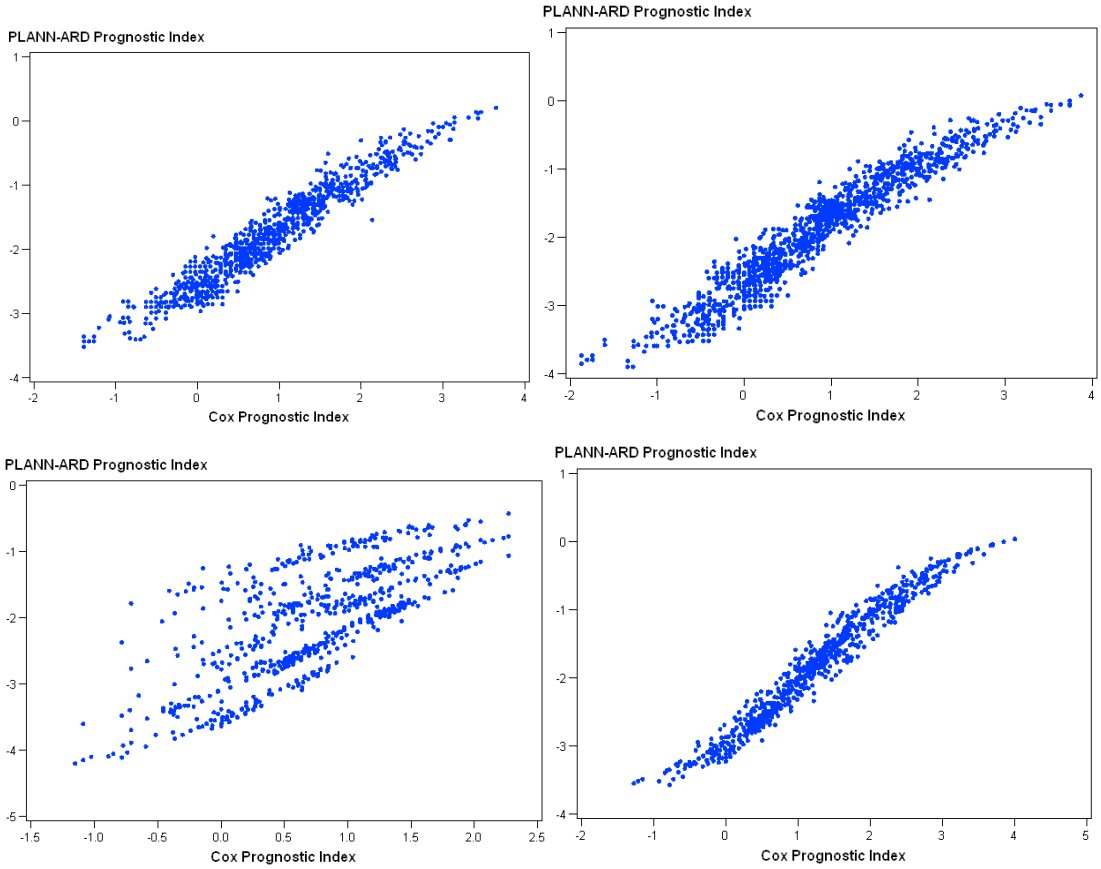

Figure 4.13 – Comparison between Cox and PLANN-ARD PI for the validation data set. The top figures represent the models developed with 5 and 7 variables, from left to right respectively. The bottom pictures represent the model developed for the 6 variables models, with Nodes involved and Nodes ratio, from left to right respectively.

For the training data set, comparing the Cox regression models values with PLANN-ARD values, overlapping performances are evident. Analysing the Chi-square values, as well as the p-values, it can be concluded that all models have a very good calibration performance, with the PLANN-ARD model slightly better than the Cox model.

For the validation data set the discrimination values are better for the PLANN-ARD model than for the Cox proportional hazards for all the models and for all the time intervals. Analysing the Chi-square values, as well as the p-values, it can be concluded that the PLANN-ARD models are better calibrated than the Cox models. Observing the calibration plots for the validation data set it can be noticed that for all models with the exception of the 6 variables model built with *Nodes Involved*, there is an underestimation of the observed survival in the groups, more accentuated on Cox model. The prognostic index obtained for each patient with Cox and PLANN-ARD can also be compared (Figure 4.12 and Figure

4.13), for training and validation data set respectively. Comparing the obtained Cox prognostic index with the PLANN-ARD prognostic index it can be verified that they are non-linear related, for all models and for both training and validation. Nevertheless there is a very high correlation between them, from 0.95551(6 variables model with *Nodes Involved*) to 0.97877(5 variables model) for the training data set, and 0.79994 (6 variables model with *Nodes Involved*) to 0.97905 (6 variables model with *Nodes Ratio*), measured by the Pearson correlation. Particularly it seems that the PLANN-ARD model compresses the PI values in the extreme sectors but extends the dynamic range for the middle sector for all models with the exception of the 6 variables model developed with *Nodes Involved*. This compression and extension can be caused by the non-linear algorithm, which implicitly models interactions between covariates.

4.6 - Stratification methodologies

Stratification of patients by risk of adverse outcome is central to clinical practice. In clinical environment, stratification of patients by risk, based on survival models is frequently used in the evaluation of treatments or on the impact of prognostic factors on survival. This begins with modelling empirical data either with a classifier or a failure time model, depending on whether the data represents a snapshot in time of the patient's condition at diagnosis, or evolution of the disease over time in a longitudinal cohort study. Either way, the equivalent of the linear argument $\beta.x$ in a Generalised Linear Model defines a prognostic index that ranks patient data by severity of the illness. In the case of breast cancer, typically a piecewise linear model is used (Cox, 1972) from which the prognostic index can be derived. After defined a prognostic index, this may be used to stratify patients into risk groups with statisticy significant grouped outcomes. A good example of this is the Nottingham Prognostic Index (NPI) which is widely used in clinical practice (Haybittle, Blamey, Elston, Johnson, Doyle, Campbell, Nicholson, Griffiths, 1982). The same principles apply when flexible models are used, such as generic non-linear algorithms including artificial neural networks.

First, the results from two stratification methodologies are presented and compared, namely the log-rank bootstrapping methodology and the minimum p-value methodology. Secondly, the results obtained with the regression tree methodology, the k-means clustering methodology and the clustering methodology based on learning metrics are presented followed by a comparison between all the presented methodologies of stratification.

4.6.1. Log-rank bootstrapping methodology and minimum p-value methodologies

The log-rank bootstrapping stratification methodology, as well as the minimum p-value methodology, were applied to both prognostic indices, one obtained through the Generalised Linear model, Cox regression and other obtained with the PLANN-ARD model.

The log-rank bootstrap methodology and the minimum p-value methodology stratify the prognostic index directly, obtained for each patient. Both methodologies were applied to a previous published model where a new attribute was created to denote the missing values of the model (Jarman et al, 2008), using the Christie data set 1983-89 to train the model. The robustness of the bootstrapping log-rank approach to risk group identification is illustrated in Figure 4.14. The bootstrap log-rank methodology has found 4 different risk groups for both prognostic indexes whereas the minimum p-value methodology has found 6 different risk groups for both prognostic indexes. The small size sample causing the unexpected outcome profiles in the solution with 6 risk groups may be an indication that this methodology is over-fitted to the training data, for both prognostic index calculations.

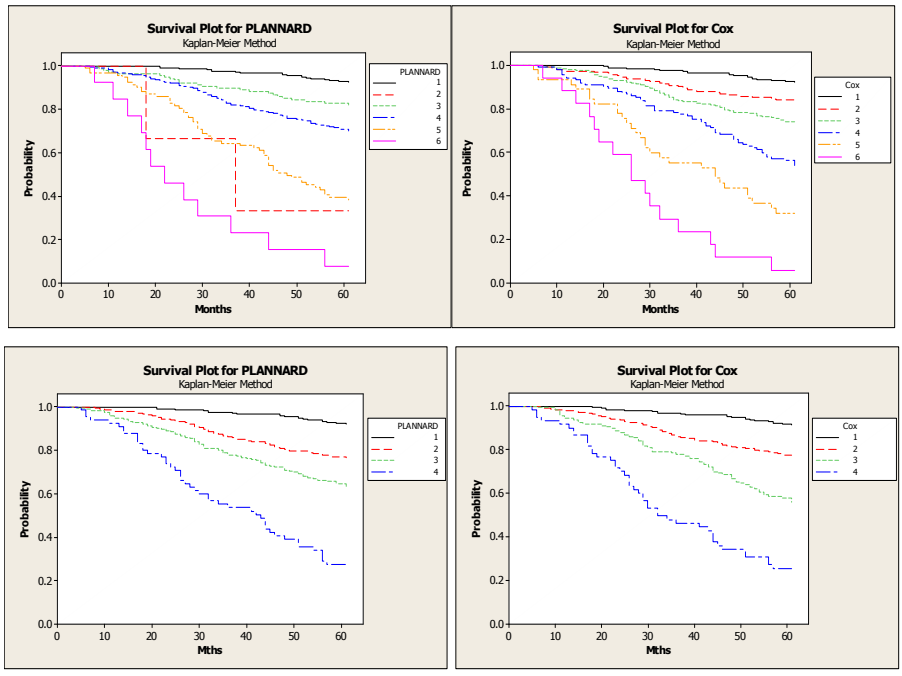

Figure 4.14 – Actuarial estimates of survival obtained with KM for the training data set. The number of cases is 917, stratified using the log-rank test over a 60 month period. The top row uses the minimum p-value method and the bottom row uses the proposed method for increasing robustness in the risk stratification. The left column uses Cox regression modelling and the right column the PLANN-ARD neural network.

An out-of sample or temporal validation for the log-rank bootstrapping methodology and minimum p-value methodology was developed using a second data set from the same centre, which was collected between 1990 and 1993. Analysing the "Kaplan Meier" curves, as well as the log-rank pairwise comparisons for the validation data set, the performance of the new methodology for stratification of illness indices can be evaluated. The robustness of the new approach to risk group identification applied to an out-of-sample data set is illustrated in Figure 4.15 and Table 4.23. Although the standard methodology applied to the training data set stratifies the patients with a significant difference in survival, the same does not happen when it is applied to the validation data set.

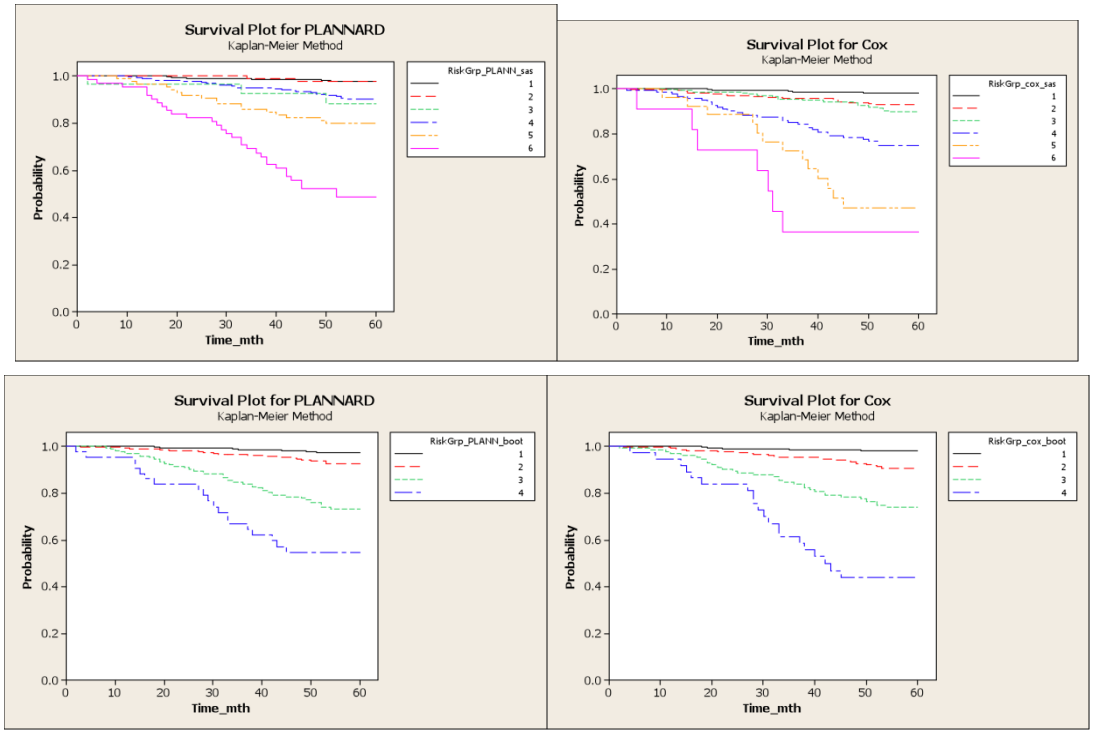

Figure 4.15 – Actuarial estimates of survival obtained with KM for the validation data set. There are 931 cases, stratified using the cut-off points found in the training data set, over a 60 month period. The top row uses the minimum p-value method and the bottom row uses the proposed method for increasing robustness in risk stratification. The left column uses Cox regression modelling and the right column the PLANN-ARD neural network.

The minimum p-value methodology's lack of robustness can be observed in the Kaplan Meier curves using the PLANN-ARD for modelling (Figure 4.15), where there is no significant survival difference between risk groups 1 and 2 and risk groups 3 and 4. The same evidence can be observed in the log-rank pairwise comparisons (Table 4.23), where there is

no significant difference between the referred groups. On the other hand, analysing both Kaplan Meier curves and log-rank pairwise comparisons, it can be observed that there is a significant survival difference in patient stratification using the methodology proposed.

While analyzing the out-of-sample validation of both stratification methodologies, it can be concluded, as it was suggested before, that the minimum p-value methodology is overfited to the training data set, resulting in not so good patient stratification when it is applied to a different data set. The new proposed methodology has however a very good performance when it is applied to an out-of-sample data set, as the difference in group-risk survival is highly significant.

	Risk Groups	**1**	**2**	**3**	**4**	**5**
		X^2 (sig.)	*X*^2 (sig.)	*X*^2 (sig.)	*X*^2 (sig.)	*X*^2 (sig.)
Log Rank (Mantel-Cox)	2	0.0006 (0.9803)	-	-	-	-
	3	6.31 (0.0120)	3.84 (0.05)	-	-	-
	4	12.29 (0.0050)	4.37 (0.0365)	0.1010 (0.7507)	-	-
	5	32.78 (0.0000)	12.97 (0.0003)	1.0443 (0.3068)	8.4239 (0.0037)	-
	6	134.66 (0.0000)	50.14 (0.0000)	10.9272 (0.0009)	91.59 (0.0000)	15.36 (0.0000)

	Risk Groups	**1**	**2**	**3**
		X^2 (sig.)	*X*^2 (sig.)	*X*^2 (sig.)
Log Rank (Mantel-Cox)	2	7.9833 (0.0047)	-	-
	3	69.59 (0.000)	39.71 (0.0000)	-
	4	125.99 (0.000)	73.50 (0.000)	7.23 (0.0072)

Table 4.23 – Log rank pairwise comparisons for the validation data set.
Both tables represent the validation modelling with PLANN-ARD. The top table represents the minimum p-value methodology and the bottom table represents the log-rank bootstrap methodology for increasing robustness in the risk stratification.

The log-rank bootstrapping methodology was applied to both prognostic models, Cox proportional hazards and PLANN, developed in chapter 2 for overall mortality using the Christie data set 1990-93, the six most predictive variables (*Histological grade*, *Histological type*, *Age*, *Oestrogen*, *Pathological size* and *Nodes Ratio*) and the missing imputation methodology. The log-rank bootstrapping methodology was validated on the BCCA data set

also using the missing imputation methodology previously explained, also applied to both model Cox proportional hazards and PLANN. This stratification methodology obtained 4 different risk survival groups for both prognostic models, where the risk group allocation retains very good separation between the observed survival in each group, measured by the actuarial estimates (Kaplan-Meier) for both training and validation data set, see Figure 4.16. This separation is quantified using the log-rank test, which gives strong statistical significance for all pairwise tests, for both training and validation data sets, reflecting the separation between the confidence intervals shown in Table 4.24 and Table 4.25.

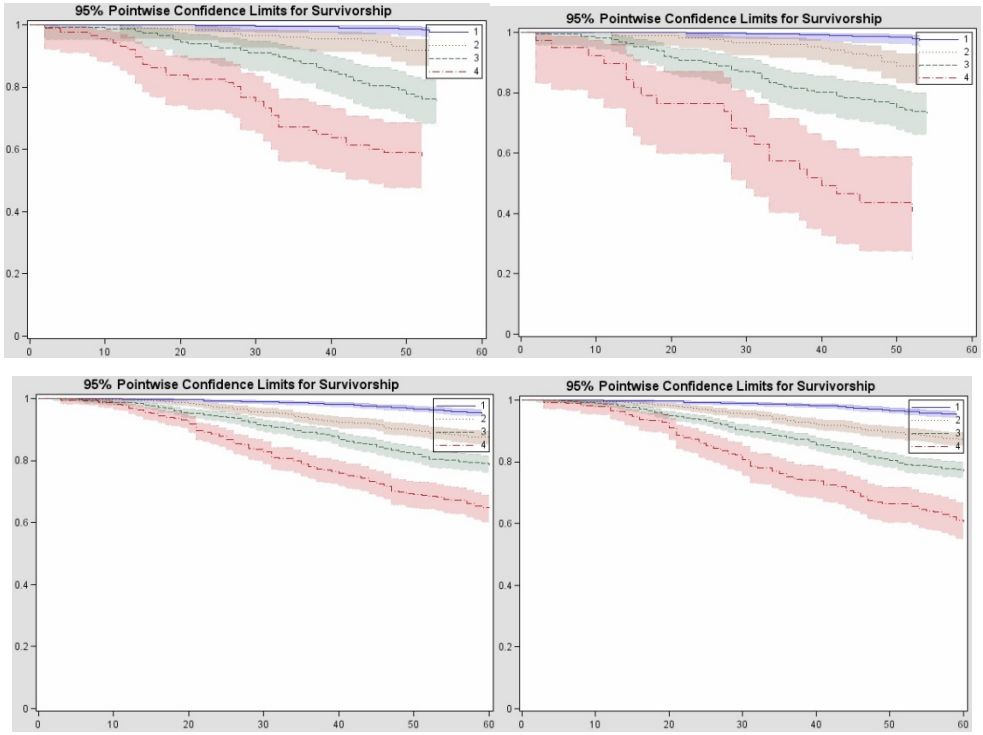

Figure 4.16 – KM curves using the log-rank bootstrapping methodology for both PI. The left pictures represents the KM curves using the PLANN-ARD index and the right pictures represents the KM curves using the Cox model. The top pictures were obtained for the training data set while the bottom pictures were obtained for the validation data set.

A concern of many clinicians is the 'black box' nature of artificial neural networks (ANN) (Lisboa, 2002) which raises the important issue of explaining individual inferences by the network. This is a key stage in evaluating the clinical plausibility of inferences made by analytical models to enable clinicians to apply these inferences with confidence. A previously published methodology designed to extract low-order Boolean rules from data will be used. The orthogonal search rule extraction (OSRE) algorithm (Etchells, Lisboa, 2006) provides a practical and efficient tool to explain the otherwise black box predictions from artificial neural

models. The OSRE methodology searches for rules using multivariate descriptions of data sub-sets and therefore can be applied to the bootstrap log-rank stratification methodology in order to define each risk group in terms of clinical rules. Table 4.26 represents the rules for the prognostic index obtained with Cox proportional hazards modelling and Table 4.27 represents the rules for the prognostic index obtained with PLANN-ARD.

	Risk Groups	1	2	3
		X^2 (sig.)	X^2 (sig.)	X^2 (sig.)
Log Rank (Mantel-Cox)	2	10.90 (0.0010)	-	-
	3	62.27 (0.000)	15.11 (0.0001)	-
	4	135.58 (0.000)	61.02 (0.000)	47.79 (0.000)

	Risk Groups	1	2	3
		X^2 (sig.)	X^2 (sig.)	X^2 (sig.)
Log Rank (Mantel-Cox)	2	18.26 (0.000)		
	3	77.47 (0.000)	13.31 (0.0003)	
	4	218.31 (0.000)	61.64 (0.000)	18.71 (0.000)

Table 4.24 – Log-rank pairwise values for the different risk groups. These were obtained using the log-rank bootstrapping methodology for both prognostic indices obtained. The left table represents the log-rank pairwise values using the PLANN-ARD index and the right table represents the log-rank values pairwise using the Cox model, for the training data set.

	Risk Groups	1	2	3
		X^2 (sig.)	X^2 (sig.)	X^2 (sig.)
Log Rank (Mantel-Cox)	2	57.20 (0.000)	-	-
	3	193.80 (0.000)	25.85 (0.000)	
	4	403.70 (0.000)	104.63 (0.000)	31.93 (0.0000)

	Risk Groups	1	2	3
		X^2 (sig.)	X^2 (sig.)	X^2 (sig.)
Log Rank (Mantel-Cox)	2	57.21 (0.000)	-	-
	3	230.24 (0.000)	37.96 (0.000)	-
	4	406.73 (0.000)	113.53 (0.000)	32.43 (0.000)

Table 4.25 – Log-rank pairwise values for the different risk groups. These were obtained using the log-rank bootstrapping methodology for both prognostic indices previously obtained. The left table represents the log-rank pairwise values using the PLANN-ARD index and the right table represents the log-rank values pairwise using the Cox model, for the validation data set.

It is interesting to verify that the differences from the training to the validation data set on the mean survival KM curves are very similar for both Cox and PLANN modelling. For both modelling there is a decreasing on the mean survival value from training and validation, for risk group 1 and 2, with a difference of approximately 3 % and 2%, respectively. However, there is an increasing on the mean survival values for the 4th risk group from training and validation. The training and validation survival values for risk group 3 are very similar, for both modelling.

Risk Group 1	Rule 1	Pathsize=1 and Noderatio=1 and Histgrade =1 or 2
	Rule 2	Age=2 and Noderatio=1 and Oestrogen=2 and Histgrade =1 or 2
	Rule 3	Noderatio=1 and Histype=2
	Rule 4	Pathsize=1 and Noderatio=1 or 2 or 3 and Histgrade = 1
	Rule 5	Pathsize=1 and Histype=2
	Rule 6	Pathsize=1 and Noderatio=1 and Oestrogen=2 and Age= 2
	Rule 7	Oestrogen=2 and Noderatio=1 and Histgrade=1
Risk Group 2	Rule 1	Histgrade=2 and Histype=1 and Noderatio=1 and Oestrogen=2 and Age= 1 or 3 and Pathsize=2
	Rule 2	Histgrade=3 and Noderatio=1 and Oestrogen=1 and Age= 2 and Pathsize=1
	Rule 3	Histgrade=3 and Noderatio=1 and Oestrogen=2 and Age= 1 or 3 and Pathsize=1
	Rule 4	Histgrade=2 and Histype=1 and Noderatio=2 and Age= 2 or 3 and Pathsize=1
	Rule 5	Histgrade=2 and Histype=1 and Noderatio=1 and Oestrogen=1 and Age=2 and Pathsize=2
	Rule 6	Histgrade=3 and Noderatio=1 and Oestrogen=2 and Age= 2 and Pathsize=2
	Rule 7	Histgrade=2 and Histype=1 and Noderatio=3 and Oestrogen=2 and Age=2 and Pathsize=2
	Rule 8	Histgrade=2 and Histype=1 and Noderatio=2 or 3 and Age=2 and Pathsize=1
Risk Grup 3	Rule 1	Histgrade=3 and Noderatio=1 or 3 and Oestrogen=1 and Pathsize=2
	Rule 2	Histgrade=1 or 2 and Noderatio=2 or 3 and Oestrogen=2 and Age= 2 or 3 and Pathsize=2
	Rule 3	Histgrade=3 and Noderatio=1 or 3 and Age= 1 or 3 and Pathsize=2
	Rule 4	Histype=1 and Noderatio=1 or 2 and Oestrogen=1 and Age= 1 or 3 and Pathsize=2
	Rule 5	Histgrade=3 and Noderatio=1 or 2 and Oestrogen=1 and Age= 1
	Rule 6	Histype=2 and Noderatio=2 or 4 and Pathsize=2
	Rule 7	Histgrade=2 and Histype=1 and Noderatio=4 and Pathsize=1
	Rule 8	Histgrade=3 and Noderatio=2 or 3 and Oestrogen=1 and Pathsize=1
Risk Group 4	Rule 1	Histgrade=2 or 3 and Histype=1 and Noderatio=2 or 4 and Pathsize=2
	Rule 2	Histype=1 and Noderatio=4 and Oestrogen=1
	Rule 3	Histgrade=3 and Noderatio=2 or 3 or 4 and Age= 1
	Rule 4	Noderatio=2 or 3 or 4 and Oestrogen=1 and Age= 1 and Pathsize=2
	Rule 5	Histgrade=3 and Noderatio=2 or 3 or 4 and Oestrogen=1 and Age= 1 or 3 and Pathsize=2

Table 4.26 – Rules obtained with OSRE.
OSRE was applied to bootstrap log-rank stratification methodology using the Cox proportional hazards prognostic index.

Risk Group 1	Rule 1	Pathsize=1 and Noderatio=1 and Histgrade =1 or 2 and Oestrogen=2
	Rule 2	Histgrade=1 or 2 and Noderatio=1 and Oestrogen=2 and Age= 2
	Rule 3	Histype=2 and Pathsize=1
	Rule 4	Histype=2 and Noderatio=1 and Oestrogen=2 and Age= 2 or 3
	Rule 5	Histgrade=1 and Noderatio=1 or 2 or 3 and Age= 2 and Pathsize=1
	Rule 6	Histgrade=1 or 2 and Noderatio=1 and Age= 2 and Pathsize=1
	Rule 7	Noderatio=1 and Oestrogen=2 and Age= 2 and Pathsize=1
Risk Group 2	Rule 1	Histgrade=1 or 2 and Histype=1 and Noderatio=1 and Oestrogen=2 and Age= 3 and Pathsize=2
	Rule 2	Histgrade=3 and Noderatio=1 and Oestrogen=1 and Age=2 and Pathsize=1
	Rule 3	Histgrade=1 or 2 and Histype=1 and Noderatio=2 or 3 and Age= 3 and Pathsize=1
	Rule 4	Histgrade=3 and Noderatio=1 and Oestrogen=2 and Age= 3 and Pathsize=1
	Rule 5	Histgrade=3 and Noderatio=1 and Oestrogen=2 and Age= 2 and Pathsize=2
	Rule 6	Histgrade=2 or 3 and Histype=1 and Noderatio=2 and Oestrogen=2 and Age= 2 or 3 and Pathsize=1
	Rule 7	Histgrade=1 or 2 and Histype=1 and Noderatio=3 and Oestrogen=2 and Age=2 and Pathsize=2
	Rule 8	Histgrade=2 and Noderatio=1 and Oestrogen=2 and Age= 1 and Pathsize=1
	Rule 9	Histgrade=2 and Histype=1 and Noderatio=3 and Oestrogen=1 and Pathsize=1
	Rule 10	Histype=2 and Noderatio=1 and Oestrogen=1 and Age= 3 and Pathsize=2
Risk Grup 3	Rule 1	Histgrade=2 or 3 and Histype=1 and Noderatio=1 and Oestrogen=1 and Age=2 and Pathsize=2
	Rule 2	Histgrade=2 and Noderatio=2 or 3 and Oestrogen=2 and Age= 2 or 3 and Pathsize=2
	Rule 3	Histgrade=3 and Noderatio=1 or 3 and Oestrogen=1 and Age= 1 or 3 and Pathsize=1
	Rule 4	Histgrade=3 and Noderatio=1 and Oestrogen=2 and Age= 1 or 3 and Pathsize=2
	Rule 5	Histgrade=1 or 2 and Histype=1 and Noderatio=1 and Oestrogen=1 and Age= 2 or 3 and Pathsize=2
	Rule 6	Histype=1 and Noderatio=1 and Oestrogen=2 and Age= 1 and Pathsize=2
	Rule 7	Noderatio=4 and Age=2 and Pathsize=1
	Rule 8	Histgrade=1 or 2 and Histype=1 and Noderatio=2 or 3 and Oestrogen=1 and Age=2 and Pathsize=2
	Rule 9	Histgrade=3 and Noderatio=2 or 3 and Oestrogen=1 and Age= 2 or 3 and Pathsize=1
	Rule 10	Histype=1 and Noderatio=1 and Oestrogen=1 and Age= 1 and Pathsize=1
Risk Group 4	Rule 1	Histgrade=3 and Oestrogen=1 and Age= 1 or 3 and Pathsize=2
	Rule 2	Histype=1 and Noderatio=4 and Pathsize=2
	Rule 3	Histype=1 and Noderatio=2 or 3 or 4 and Oestrogen=1 and Pathsize=2
	Rule 4	Histgrade=3 and Noderatio=2 or 3 or 4 and Age= 1 or 3 and Pathsize=2
	Rule 5	Histype=1 and Noderatio=4 and Age= 1 or 3
	Rule 6	Noderatio=4 and Age= 1 or 3 and Pathsize=2
	Rule 7	Oestrogen=1 and Age=1 and Pathsize=2
	Rule 8	Noderatio=2 or 3 or 4 and Oestrogen=1 and Age= 1
	Rule 9	Histype=1 and Noderatio=2 or 4 and Age= 1 and Pathsize=2
	Rule 10	Noderatio=4 and Oestrogen=1

Table 4.27 – Rules obtained with OSRE.
OSRE was applied to bootstrap log-rank stratification methodology using PLANN-ARD prognostic index.

		Mean KM survival Cox Training	Mean KM survival for Cox Validation	Differences	Mean KM survival PLANN –ARD Training	Mean KM survival PLANN-ARD Validation	Differences
Risk Groups	1	97,53	95,26	2,27	97,85	95,26	2,59
	2	89,02	87,33	1,69	91,71	87,23	4,48
	3	74,10	76,68	-2,58	76,51	78,21	-1,7
	4	43,59	60,37	-16,78	59,09	64,55	-5,46

Table 4.28 – Mean KM survival values at the end of follow up (5 years).
This was computed for the Cox Proportional hazards modelling and PLANN-ARD modelling, for the training and validation data set.

4.6.2. Regression tree stratification methodology

Regression tree decision methodology was performed using CART algorithm (Breiman, Friedman, Olsen, Stone, 1984) where there were 6 predictor categorical variables (*Histological grade*, *Histological type*, *Nodes Ratio*, *Oestrogen*, *Pathological size* and *Age*) and one continuous target variable, which was the prognostic index already obtained. One tree was developed using the prognostic index obtained with Cox Proportional Hazards and another tree was developed using the PLANN-ARD Modeling. Missing values were incorporated in the modelling, using the multiple imputation methodology previously mentioned. It is important to mention that the predictor variables used in CART are the mode of the 10 imputed data sets obtained previously, both for training and validation data set. It is important to mention, that a full tree is complex and can yield an overly optimistic goodness of fit. Thus, methods to reduce the tree size have been developed so that the model is predictive in other cohorts. Moreover, it is essential that the leaves contain statistically significant patients, as they need to be compared to each other in order to identify the belonging to a risk group. Therefore, it was defined that the minimum number of records for each node tree was 20, and the minimum number of records for each leaf was 10, in order to define a significance level of populations to be compared. At the maximum grow of the tree, both trees, one with the prognostic index calculated with Cox regression and another with the prognostic index calculated with PLANN-ARD, finalized with 41(N) and 47 (N) leaves, respectively. The "Cox regression tree" has an error of 0,079 and the "PLANN-ARD" regression tree has an error of 0,089, both calculated using a 10 fold cross-validation. After the development of the "pruning method" both trees finalized with 4 different groups, which means the regression tree, after the "pruning method" ended as a classification tree, as it can be observed in the following figures.

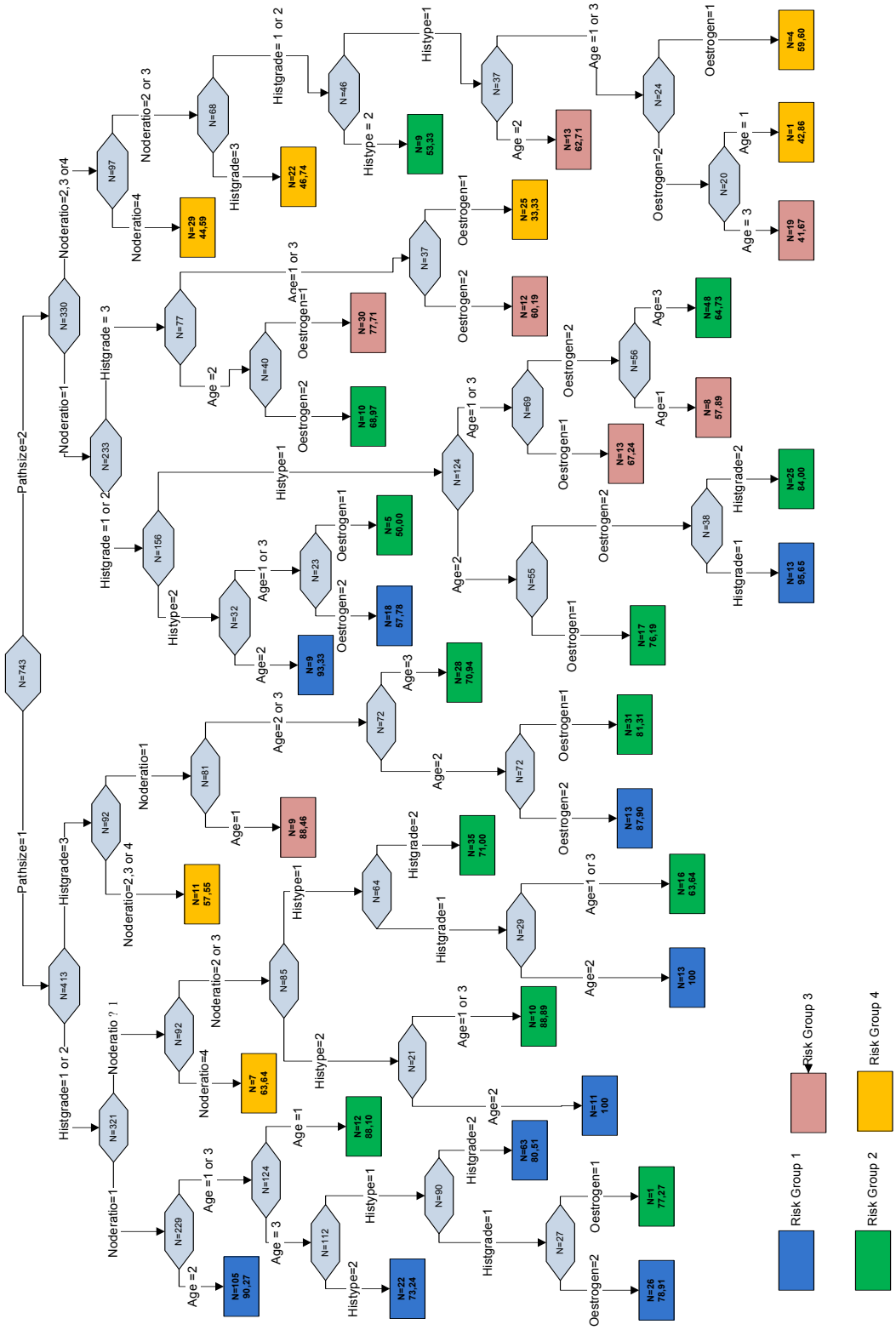

Figure 4.17 – Final classification tree using the PI obtained with PLANN-ARD.

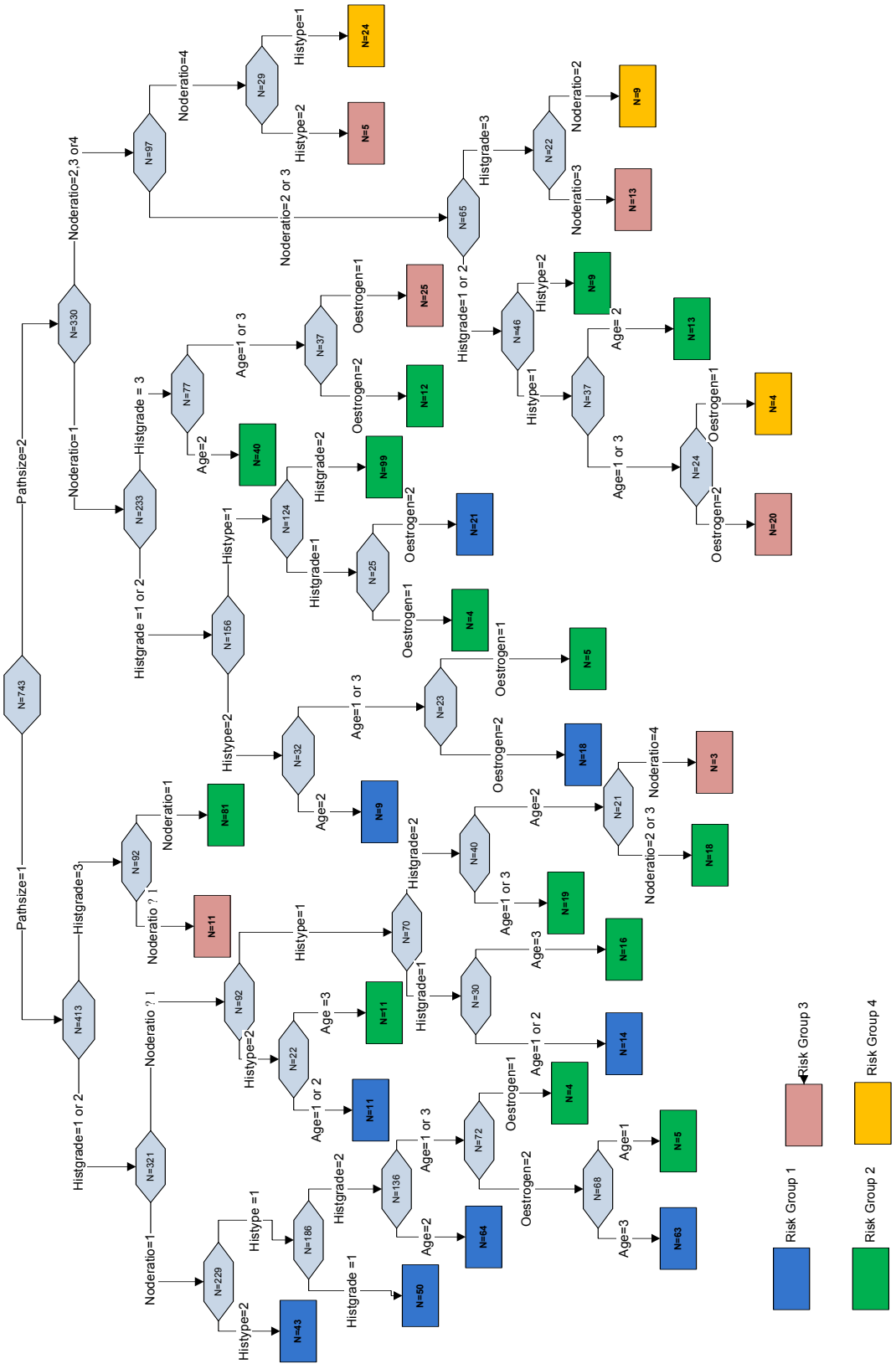

Figure 4.18 – Final classification tree using the PI obtained with Cox regression.

Figure 4.19 represents the box plots identified for the different risk groups as well as for both prognostic indices. It can be observed that the confidence intervals for the different risk groups are very low and as long as the risk group goes higher the prognostic index also goes higher, showing that the risk groups are consistent with the prognostic index.

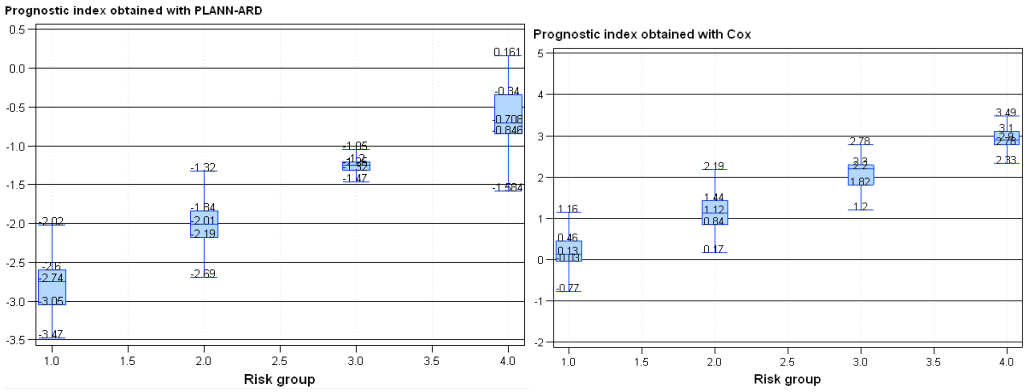

Figure 4.19 – Box-plots for the different risk groups.
The risk groups were obtained using the regression tree stratification method for both prognostic indices previously obtained, for the training data set. The left picture represents the box plots using the PLANN-ARD prognostic index and the left picture represents the box plot using the Cox prognostic index.

Kaplan-Meier curves, as well as the log-rank pairwise comparisons were obtained, for both obtained trees, and for the training data set, where it can be sustained that group allocation retains very good separation between the observed survival in each group, measured by the actuarial estimates (Kaplan-Meier). This separation is quantified using the log-rank test, which gives strong statistical significance for all pairwise tests. The weakest separation is between groups 2 and 3 using PLANN-ARD, for which the p-value is 0.0048.

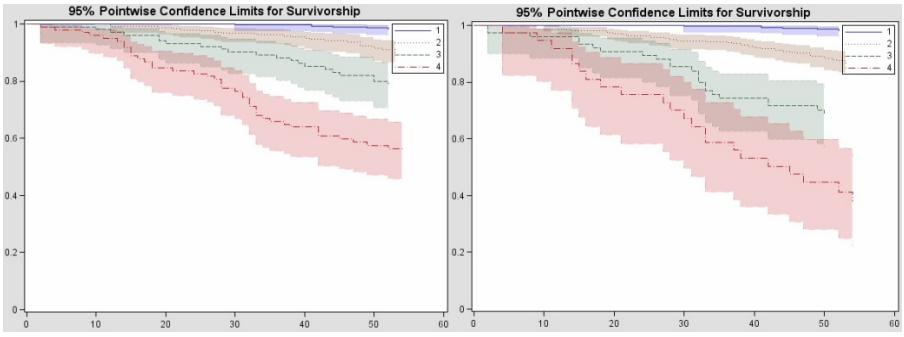

Figure 4.20 – KM curves using the regression tree stratification method for both PI.
The left picture represents the KM curves using the PLANN-ARD index and the right picture represents the KM curves using the Cox model, using the training data set.

	Risk Groups	1	2	3
		X ^2 (sig.)	X ^2 (sig.)	X ^2 (sig.)
Log Rank (Mantel-Cox)	2	17.50 (0.000)		
	3	46.86 (0.000)	7.96 (0.0048)	
	4	141.53 (0.000)	61.02 (0.000)	13.27 (0.0003)

	Risk Groups	1	2	3
		X ^2 (sig.)	X ^2 (sig.)	X ^2 (sig.)
Log Rank (Mantel-Cox)	2	27.96 (0.000)		
	3	80.14 (0.000)	16.32 (0.000)	
	4	207.01 (0.000)	73.09 (0.000)	9.44 (0.0021)

Table 4.29 – Log-rank pairwise values for the different risk groups. These were obtained using the regression tree stratification method for both prognostic indices previously obtained. The left table represents the log-rank pairwise values using the PLANN-ARD index and the right table represents the log-rank values pairwise using the Cox model, using the training data set.

One of the main advantages of this stratification methodology is its simplicity and transparency of group composition, using the rules that are obtained at the final built tree. The rules for the previously showed trees are on Table 4.30 and Table 4.31 .

Risk Group 1	Rule 1	Pathsize=1 and Noderatio=1 and Histype=2 and Histgrade =1 or 2
	Rule 2	Pathsize=1 and Noderatio=1 and Histype=1 and Histgrade =1
	Rule 3	Pathsize=1 and Noderatio=1 and Histype=1 and Histgrade = 2 and Age =2
	Rule 4	Pathsize=1 and Noderatio=1 and Histype=1 and Histgrade = 2 and Age =3 and Oestrogen =2
	Rule 5	Pathsize=1 and Noderatio≠1 and Histype=2 and Histgrade =1 or 2 and Age=1 or 2
	Rule 6	Pathsize=1 and Noderatio≠1 and Histype=1 and Histgrade =1 and Age=1 or 2
	Rule 7	Pathsize=2 and Noderatio=1 and Histgrade=1 or 2 and Histype=2 and Age =2
	Rule 8	Pathsize=2 and Noderatio=1 and Histgrade=1 or 2 and Histype=2 and Age =1 or 3 and Oestrogen=2
	Rule 9	Pathsize=2 and Noderatio=1 and Histgrade=1 and Histype=1 and Oestrogen=2
Risk Group 2	Rule 1	Pathsize=1 and Noderatio=1 and Histype=1 and Histgrade = 2 and Age =1 and Oestrogen =2
	Rule 2	Pathsize=1 and Noderatio=1 and Histype=1 and Histgrade=2 and Age =1 or 3 and Oestrogen =1
	Rule 3	Pathsize=1 and Noderatio≠1 and Histype=2 and Histgrade =1 or 2 and Age=3
	Rule 4	Pathsize=1 and Noderatio≠1 and Histype=1 and Histgrade =1 and Age=3
	Rule 5	Pathsize=1 and Histgrade=3 and Noderatio=1
	Rule 6	Pathsize=1 and Histgrade=2 and Noderatio≠1 and Histype=1 and Age=1 or 3
	Rule 7	Pathsize=1 and Histgrade=2 and Noderatio=2 or 3 and Histype=1 and Age=2
	Rule 8	Pathsize=2 and Noderatio=1 and Histgrade=1 or 2 and Histype=2 and Age =1 or 3 and Oestrogen=1
	Rule 9	Pathsize=2 and Noderatio=1 and Histgrade=1 and Histype=1 and Oestrogen=1
	Rule 10	Pathsize=2 and Noderatio=1 and Histgrade=2 and Histype=1
	Rule 11	Pathsize=2 and Noderatio =1 and Histgrade=3 and Age=2
	Rule 12	Pathsize=2 and Noderatio =1 and Histgrade=3 and Age=1 or 3 and Oestrogen=2
	Rule 13	Pathsize=2 and Noderatio = 2 or 3 and Histgrade=1 or 2 and Histype=2
	Rule 14	Pathsize=2 and Noderatio = 2 or 3 and Histgrade=1 or 2 and Histype=1 and Age=2
Risk Group 3	Rule 1	Pathsize=1 and Histgrade=3 and Noderatio≠1
	Rule 2	Pathsize=1 and Histgrade=2 and Noderatio=4 and Histype=1 and Age=2
	Rule 3	Pathsize=2 and Noderatio=1 and Histgrade=3 and Age=1 or 3 and Oestrogen=1
	Rule 4	Pathsize=2 and Noderatio=4 and Histype=2
	Rule 5	Pathsize=2 and Noderatio=3 and Histgrade=3
	Rule 6	Pathsize=2 and Noderatio=2 or 3 and Histgrade=1 or 2 and Histype=1 and Age=1 or 3 and Oestrogen=2
Risk Group 4	Rule 1	Pathsize=2 and Noderatio=2 or 3 and Histgrade=1 or 2 and Histype=1 and Age=1 or 3 and Oestrogen=1
	Rule 2	Pathsize=2 and Noderatio=2 and Histgrade=3
	Rule 3	Pathsize=2 and Noderatio=4 and Histype=1

Table 4.30 – Rules obtained with regression tree using the Cox proportional hazards PI.

Risk Group 1	Rule 1	Pathsize=1 and Noderatio=1 and Age=2 and Histgrade =1 or 2
	Rule 2	Pathsize=1 and Noderatio=1 and Histype=2 and Histgrade =1 or 2 and Age=3
	Rule 3	Pathsize=1 and Noderatio=1 and Histype=1 and Histgrade = 1 and Age =3 and Oestrogen =2
	Rule 4	Pathsize=1 and Noderatio=1 and Histype=1 and Histgrade = 2 and Age =3
	Rule 5	Pathsize=1 and Noderatio=2 or 3 and Histype=2 and Histgrade = 1 or 2 and Age =2
	Rule 6	Pathsize=1 and Noderatio=2 or 3 and Histype=1 and Histgrade =1 and Age=2
	Rule 7	Pathsize=1 and Noderatio=1 and Histgrade=3 and Oestrogen=2 and Age =2
	Rule 8	Pathsize=2 and Noderatio=1 and Histgrade=1 or 2 and Histype=2 and Age =2
	Rule 9	Pathsize=2 and Noderatio=1 and Histgrade=1 or 2 and Histype=2 and Age =1 or 3 and Oestrogen=2
	Rule 10	Pathsize=2 and Noderatio=1 and Histgrade=1 and Oestrogen=2 and Age=2 and Histype=1
Risk Group 2	Rule 1	Pathsize=1 and Noderatio=1 and Histgrade = 1 or 2 and Age =1
	Rule 2	Pathsize=1 and Noderatio=1 and Histype=1 and Histgrade = 1 and Age =3 and Oestrogen =1
	Rule 3	Pathsize=1 and Noderatio= 2 or 3 and Histype=2 and Histgrade =1 or 2 and Age=1 or 3
	Rule 4	Pathsize=1 and Noderatio=2 or 3 and Histype=1 and Histgrade =1 and Age=1 or 3
	Rule 5	Pathsize=1 and Noderatio=2 or 3 and Histype=1 and Histgrade =2
	Rule 6	Pathsize=1 and Histgrade=3 and Noderatio=1 and Age=2 and Oestrogen=1
	Rule 7	Pathsize=1 and Histgrade=3 and Noderatio=1 and Age=3
	Rule 8	Pathsize=2 and Noderatio=1 and Histgrade=1 or 2 and Histype=2 and Age =1 or 3 and Oestrogen=1
	Rule 9	Pathsize=2 and Noderatio=1 and Histgrade=1 or 2 and Histype=1 and Oestrogen=1 and Age=2
	Rule 10	Pathsize=2 and Noderatio=1 and Histgrade=2 and Histype=1 and Age=2 and Oestrogen=2
	Rule 11	Pathsize=2 and Noderatio =1 and Histgrade=3 and Age=2 and Oestrogen=2
	Rule 12	Pathsize=2 and Noderatio =1 and Histgrade=1 or 2 and Age=3 and Oestrogen=2 and Histype=1
	Rule 13	Pathsize=2 and Noderatio = 2 or 3 and Histgrade=1 or 2 and Histype=2
Risk Group 3	Rule 1	Pathsize=1 and Histgrade=3 and Noderatio=1 and Age=1
	Rule 2	Pathsize=2 and Histgrade=1 or 2 and Noderatio=1 and Histype=1 and Age=1 or 3 and Oestrogen=1
	Rule 3	Pathsize=2 and Histgrade=3 and Noderatio=1 and Age=1 or 3 and Oestrogen=2
	Rule 4	Pathsize=2 and Histgrade=1 or 2 and Noderatio=1 and Histype=1 and Age=1 and Oestrogen=2
	Rule 5	Pathsize=2 and Noderatio=1 and Histgrade=3 and Age=2 and Oestrogen=1
	Rule 6	Pathsize=2 and Noderatio=2 or 3 and Histgrade=1 or 2 and Age=2 and Histype=1
	Rule 7	Pathsize=2 and Noderatio=2 or 3 and Histgrade=1 or 2 and Age= 3 and Oestrogen=2 and Histype=1
Risk Group 4	Rule 1	Pathsize=1 and Histgrade=1 or 2 and Noderatio=4
	Rule 2	Pathsize=1 and Noderatio=2 or 3 or 4 and Histgrade=3
	Rule 3	Pathsize=2 and Noderatio=4
	Rule 4	Pathsize=2 and Noderatio=2 or 3 and Histgrade=3
	Rule 5	Pathsize=2 and Noderatio=2 or 3 and Histgrade=1 or 2 and Histtype=1 and Age=1 or 3 and Oestrogen=1
	Rule 6	Pathsize=2 and Noderatio=2 or 3 and Histgrade=1 or 2 and Histtype=1 and Age=1 and Oestrogen=2
	Rule 7	Pathsize=2 and Noderatio=1 and Histgrade=3 and Age=1 or 3 and Ostrogen=1

Table 4.31 – Rules obtained with regression tree using the PLANN-ARD PI.

An out-of sample or temporal validation for the regression tree methodology was performed using the BCCA data set. As in the development of both trees, it was also used the missing imputation methodology and it was used also the mode of the covariates, considering

the 10 imputed data sets. The obtained rules from the final trees were applied for each patient on the validation data set identifying which ones belong to each risk group.

Figure 4.21 represents the box plots identified for the different risk groups, for both prognostic indices obtained with PLANN-ARD and Cox proportional hazards. It can be observed that the confidence intervals for the different risk groups are very low as in the training data set. As long as the risk group goes higher the prognostic index also goes higher, showing that the risk groups are consistent with the prognostic index.

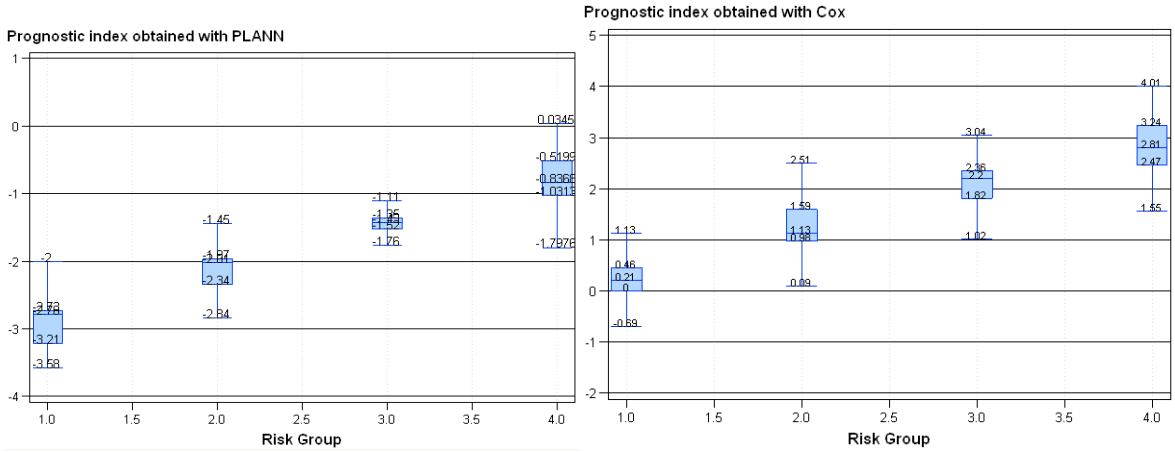

Figure 4.21 – Box-plots for the different groups.
These were obtained using the regression tree stratification method for both prognostic indices previously obtained, for the validation data set. The left picture represents the box plots using the PLANN-ARD prognostic index and the left picture represents the box plot using the Cox prognostic index.

Analysing the "Kaplan Meier" curves as well as log rank pairwise comparisons for the validation data set, the performance of the regression tree methodology for stratification of illness indices, can be evaluated. The robustness of this approach to risk group identification applied to an out-of-sample data set is illustrated in Figure 4.22 Table 4.33 .

		Mean KM survival Cox Training	Mean KM survival for Cox Validation	Differences	Mean KM survival PLANN –ARD Training	Mean KM survival PLANN-ARD Validation	Differences
Risk Groups	**1**	98.29	95.49	2.8	98.29	95.10	3.19
	2	87.20	86.14	1.06	90.28	88.73	1.55
	3	70.13	70.37	-0.24	79.81	79.27	0.54
	4	40.54	63.30	-22.76	56.57	67.63	-11.06

Table 4.32 – Mean KM survival values at the end of follow up (5 years).
These were obtained for Cox Proportional hazards modelling and PLANN-ARD modelling, for the training and validation data set.

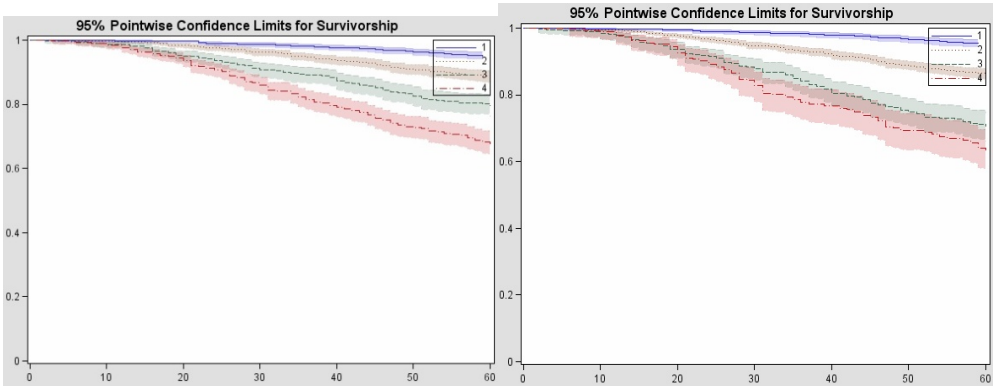

Figure 4.22 – KM curves using the regression tree stratification method for both PI. The left picture represents the KM curves using the PLANN-ARD index and the right picture represents the KM curves using the Cox model, using the validation data set.

Kaplan-Meier curves, as well as the log-rank pairwise comparisons were obtained, for both trees, where it can be confirmed that group allocation retains very good separation between the observed survival in each group, measured by the actuarial estimates (Kaplan-Meier). This separation is quantified using the log-rank test, which gives strong statistical significance for all pairwise tests. The weakest separation is for groups 3 and 4 using Cox prognostic index, the same obtained with the training data set, for which the p-value is 0.0526. The Kaplan Meier curves as well as the log-rank pairwise values obtained increase our expectation that this is a very good method to allocate the patients in different risk groups, even when there are some differences of the mean survival from the training to the validation data set as it can be observed on table Table 4.32 , which represents the KM survival values at the end of follow up. It is interesting to verify that the differences from the training to validation data set on the mean survival KM curves are very similar for both Cox and PLANN modelling. For both modelling there is a decreasing on the mean survival value from training and validation, for risk group 1 and 2, with a difference of approximately 3 % and 1%, respectively. However, there is an increasing on the mean survival values for the 4th risk group from training and validation. The training and validation survival values for risk group 3 are very similar for both modelling.

	Risk Groups	1	2	3
		X ^2 (sig.)	X ^2 (sig.)	X ^2 (sig.)
Log Rank (Mantel-Cox)	2	38.6166 (0.000)	-	-
	3	135.6330 (0.000)	31.2572 (0.000)	-
	4	327.8023 (0.000)	133.6450 (0.000)	22.5478 (0.0000)

	Risk Groups	1	2	3
		X ^2 (sig.)	X ^2 (sig.)	X ^2 (sig.)
Log Rank (Mantel-Cox)	2	79.0760 (0.000)	-	-
	3	244.9665 (0.000)	68.3584 (0.000)	-
	4	308.6466 (0.000)	100.0789 (0.000)	3.7561 (0.0526)

Table 4.33 – Log-rank pairwise values for the different risk groups. These were obtained using the regression tree stratification method for both prognostic indices previously obtained. The left table represents the log-rank pairwise values using the PLANN-ARD index and the right table represents the log-rank values pairwise using the Cox model, using the validation data set.

4.6.3. Unsupervised clustering stratification methodology

The clustering method is an orthogonal, unsupervised, approach where the clinical data is first clustered without reference to the PI, then organised in order of mean group survival. It is an iterative k-means algorithm, which uses Monte Carlo methods to overcome initialization problems. The algorithm was applied to the training data set and all variables were normalized using the metric *(value-min_variable)/ (max_variable- min_variable)*. As it was employed the imputation methodology, it was used the mode of variables obtained from the 10 imputed data sets. Moreover, only the variables previously reasoned as the most predictive ones to the prognostic model were considered (*Histological grade*, *Histological Type*, *Oestrogen*, *Age*, *Pathological size* and *Nodes Ratio*).

Two indices were used in order to verify the consistency of this method. The first one is the Fisher separation index for each cluster partition, invariant J value, and the second one is the Cramer V-index. For each individual partition, the Cramer V-index is measured for every pairing with the remaining cluster partitions of a given cluster number, and the median value is recorded. The Cramer V-index is a measure of concordance, which quantifies the extent of consistency between many cluster partitions with the same number of clusters but with random initializations. A value equal to one means that the mean of the patients are always assigned to the same cluster, even with different initializations. Thus, the more the values are close to 1, the clustering is more consistent for each initialization. For a given cluster number, the methodology starts with N initializations of the clustering algorithm. The results are not particularly sensitive to either the value of N or the proportion of clustering kept, as long as these numbers are large enough to show the structure of the cluster solution space.

Several cluster solutions were considered, from 2 to 10 cluster centres. Figure 4.23 presents the box plots considering the Cramer V values, showing the concordance for the different solutions. From this figure it can be concluded that the 4 cluster solution has the higher concordance values as it is the solution that the median and the mean Cramer V values are closer to 1. Mapping the space of cluster partitions, a Separation-Concordance plot, Figure 4.24 there will be visible an indication of the most suitable cluster number to use for a particular data set, since a match between the assumed and actual cluster number is likely to result in more stable solutions, i.e. those scoring highest under the Cramer V-index. Figure 4.25 represents the cumulative Cramer V Concordance values for the different cluster solutions and its area under the curve, where the smaller the area the better cluster solution. Therefore, the most likely cluster number is 4.

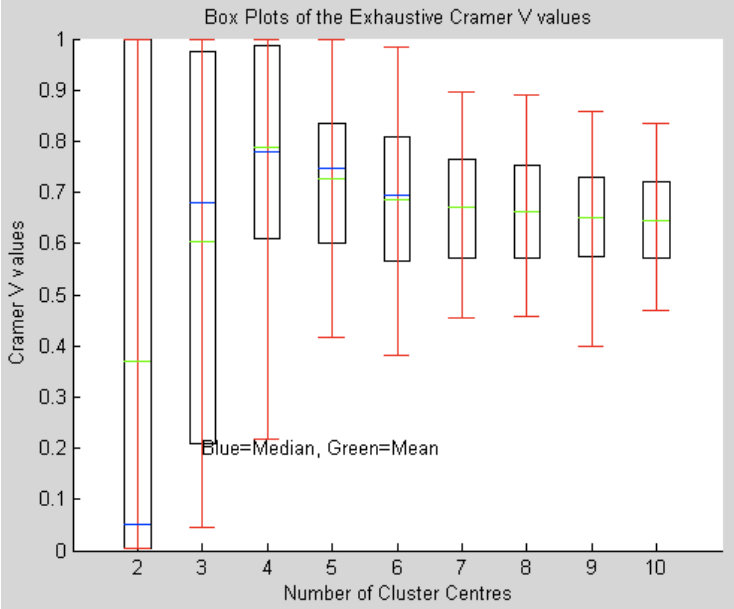

Figure 4.23 – Concordance plot for different number of clustering.

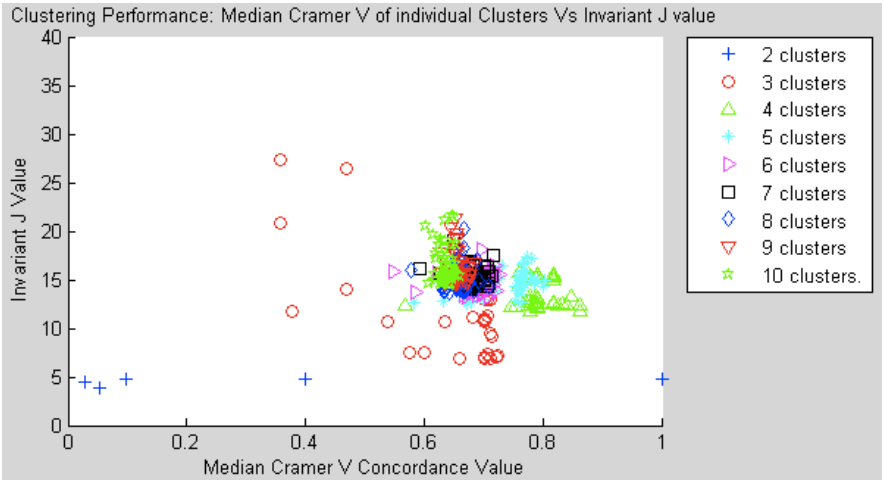

Figure 4.24 – Separation measure (y axis) versus concordance measure (x axis) plot.

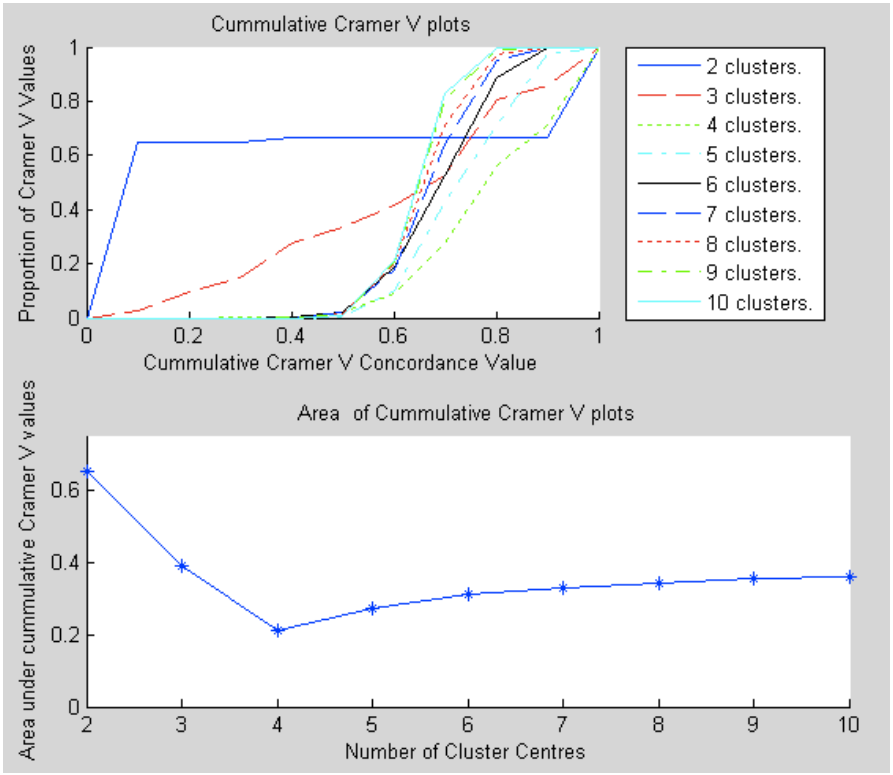

Figure 4.25 – Cramer Area plot.
The top figure represents the cumulative Cramer-V scores, for each cluster solution, where the smaller the area the better the solution. The bottom figure represents the area under the curve for each number of clusters.

The box plots related to the two prognostic indices, the Cox and the PLANN-ARD are presented on Figure 4.26. It can be observed that the mean of the prognostic indexes are well separated for each cluster, showing that the risk groups are consistent with the prognostic index, for both Cox and PLANN-ARD. It can also be concluded with that the patients profile is consistent with each patient prognostic index for the one obtained with PLANN-ARD and Cox proportional hazards modelling. Risk groups displayed distinct observed survival, measured by Kaplan-Meier actuarial estimates and log-rank pairwise values, observed in Figure 4.27 and Table 4.34. However, it can be observed that the separation between the 3nd and the 4th risk group, which are the ones with the higher survival, is not significantly different as the p-value is equal to 0.2309.

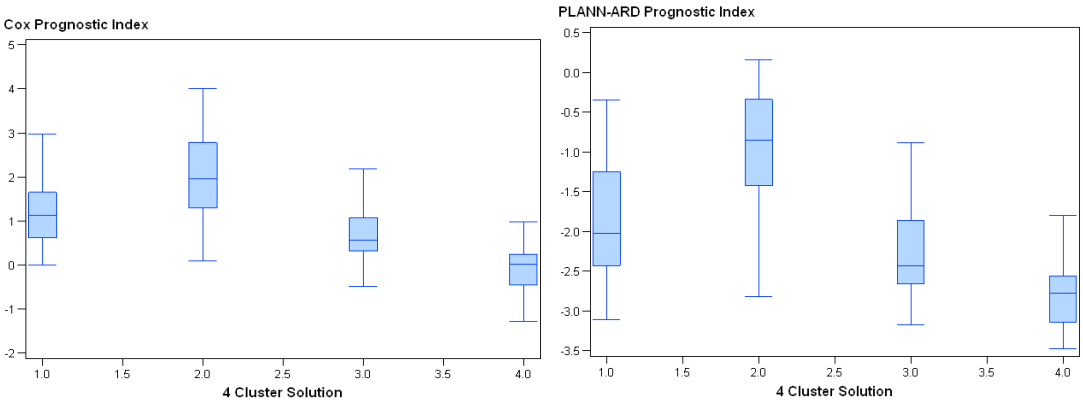

Figure 4.26 – Box plots for the 3 (top figures) and 4 (bottom figures) cluster solution. The left pictures represent the graphic using the prognostic index obtained with Cox while the right pictures represent the prognostic index obtained with PLANN-ARD.

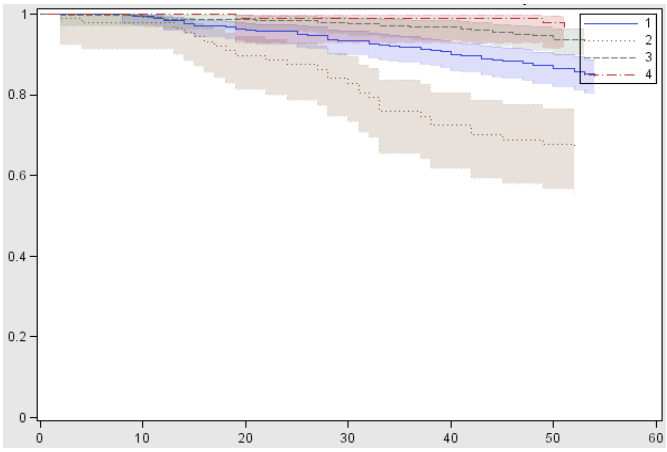

Figure 4.27 – KM curves the 4 cluster solution.

	Risk Groups	1	2	3
		X ^2 (sig.)	X ^2 (sig.)	X ^2 (sig.)
Log Rank (Mantel-Cox)	2	16.57 (0.0000)	-	-
	3	9.73 (0.0018)	43.81 (0.0000)	-
	4	9.09 (0.0026)	29.15 (0.0000)	1.43 (0.2309)

Table 4.34 – Log-rank pairwise values for 4 cluster solution.

At the end of the clustering algorithm, these can be converted into rules of variables, which determine the characterization of the patient cohorts using the OSRE algorithm, offering transparency of group composition, as it can be observed in Table 4.35 .

	Group 1	**Group 2**	**Group 3**	**Group 4**
Rule 1	Histgrade =2 or 3 and Histype = 1 and Noderatio =1 and Age =1 or 2	Noderatio = 3 or 4	Histgrade = 1 or 2 and Histtype = 1 and Noderatio = 1 or 2 and Age = 3	Histtype = 2 and Noderatio = 1 or 2
Rule 2	Histgrade = 3 and Noderatio=1 or 2	-	Histgrade = 1 and Histtype = 1 and Noderatio = 1 or 2	-
Rule 3	Histgrade =2 or 3 and Histype = 1 and Noderatio =1 or 2 and Oestrogen = 1 and Age =1 or 2	-	Histgrade = 1 or 2 and Histtype = 1 and Noderatio = 2 and Oestrogen = 2 and Age = 2 or 3	-
Rule 4	Histgrade = 2 or 3 and Histtype = 1 and Noderatio = 1 or 2 and Age = 1	-	-	-

Table 4.35 – Rule-based characterization of the patient cohorts.
These rules were found using the clustering method followed by the OSRE algorithm.

This stratification methodology was validated in an out-of sample data set, on the BCCA data set. The method for validate the clustering algorithm can be one of the three:

1. Using the centres obtained for each cluster
2. Using the rules obtained for each centre

3. Defining the k records closer to each record to validate. The class is represented as the mode of the class for all the k records. (it was used a k = 9)

The first method was not a good validation approach as it is demonstrated by applying the obtained centres to the training data set (the records which were first characterized as belonging to a risk group were now characterized as belonging to another risk group). There was an error of 79% using the centres of each cluster as a validation method.

Using the second method, there are some new records that may not be fitted to any of the rules obtained with the training data set and can be overlapped in different classes. If this happens, these records can be considered as outliers and should be flagged as special cases with an explanation of how they differ from the nearest rules.

The last method finds the nearest records between the training data set and the validation data set (euclidean distance) and the existing mode for the 9 nearest records is the class labelled for the new record in the validation data set.

This stratification methodology validation was carried out through the mentioned last two ways of doing it, and the patient risk group allocation was compared. Using the rules, 82 patients were classified as belonging to both risk group 1 and 4. Therefore, they were excluded as outliers. Using the k closer records, 47 of these 82 patients were considered to belong to risk group 4 and 35 to risk group 1. Analysing the cross-tabulation on Table 4.36 it can be observed that both validations are very similar as there were only 26 patients classified to different risk groups.

		Clustering rules				
		Group1	Group2	Group3	Group4	**Total**
Clustering 9 closer	Group1	2014	10	1	0	2025
	Group2	0	385	0	0	385
	Group3	0	0	1320	15	1335
	Group4	0	0	0	189	189
	Total	2014	395	1321	204	3934

Table 4.36 – Patients' cross tabulation using the application of nearest records and rules.

Using the k closer records methodology, different risk groups for the validation data set were obtained. They displayed distinct observed survival, measured by Kaplan-Meier actuarial estimates and log-rank pairwise values, with the exception of risk group 3 and 4 (the highest survival risk groups), as it can be observed in Figure 4.28 and Table 4.37 . It can be noticed that the survival curves obtained for the training data set are very similar to the ones obtained for the validation data set.

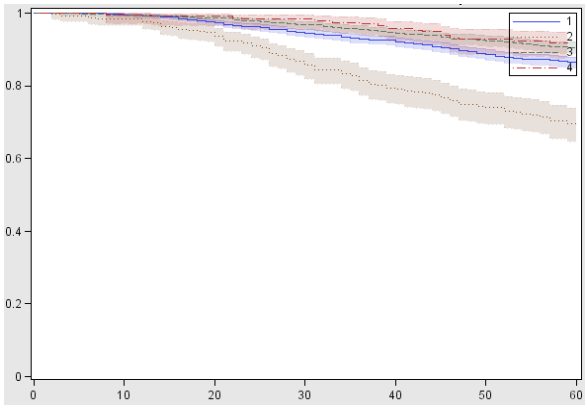

Figure 4.28 – KM curves for the 4 cluster solution, using the validation data set.

	Risk Groups	1	2	3
		X ^2 (sig.)	X ^2 (sig.)	X ^2 (sig.)
Log Rank (Mantel-Cox)	2	77.16 (0.0000)	-	-
	3	13.50 (0.0002)	122.89 (0.0000)	-
	4	5.19 (0.0162)	42.40 (0.0000)	0.32 (0.5711)

Table 4.37 – Log-rank pairwise values for the 4 cluster solution and validation data set.

4.6.4. Clustering methodology based on learning metrics

The clustering methodology based on learning metrics approach was applied to 3 distributions of auxiliary information, i.e. using PI_{Cox} alone, using $PI_{PLANNARD}$ alone and on the joint information from the two independent indices. The first two experiments predicted the cluster number to be either 6 or 7. By using the joint information, on the other hand, the algorithm stably identifies 5 clusters. Figure 4.29 shows the KM curves for the 5 patient risk groups for the training data set, obtained with the joint information, as well as a 2D plot of the clustered samples in the space of the prognostic indices. The KM plot clearly shows that

clusters 2 and 3 denote the same risk profile, while the remaining 3 clusters identify 3 markedly different survival behaviours.

On Table 4.38 are the log-rank pairwise values represented on the KM plot, where it can be observed that they do not display distinct observed survival, specially for risk groups 1 and 4; 2 and 3. Figure 4.30 shows the clusters projected onto the 3 principal components of the original data (leftmost) and of the dataset subject to the affine transformations induced by the Fisher metric (rightmost). In the original space, cluster 3 is fully contained in a separated sample group and such separation in the input space seems to be the cause of the generation of a separate cluster with the same survival profile of cluster 2, also in the Fisher space. Overall, the clustering analysis seems to confirm that there are 4 risk groups within the data. Clusters identify 4 markedly different survival behaviours.

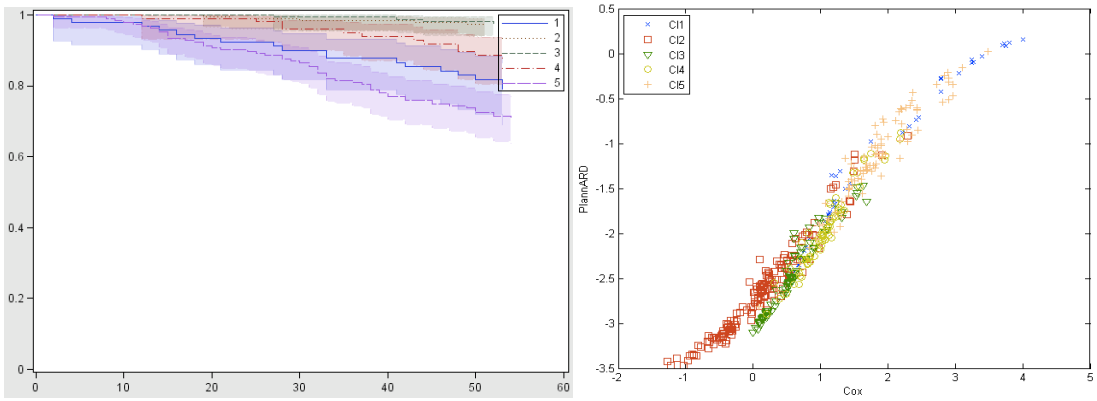

Figure 4.29 – KM curves and cluster in the space of two prognostic indices. The left-most picture represents the actuarial estimates of survival obtained with the Kaplan-Meier method, stratified over a 60 month period and the right-most picture depicts the clustered samples in the space of the two prognostic indices.

	Risk Groups	1	2	3	4
		X ^2 (sig.)	X ^2 (sig.)	X ^2 (sig.)	X ^2 (sig.)
Log Rank (Mantel-Cox)	2	25.59 (0.0000)	-	-	-
	3	22.95 (0.0000)	0.0684 (0.7970)	-	-
	4	2.69 (0.1085)	9.70 (0.0031)	9.49 (0.0027)	-
	5	2.11 (0.1316)	49.98 (0.0000)	42.15 (0.0000)	10.63 (0.0007)

Table 4.38 – Log-rank pairwise values for the 5 cluster solution

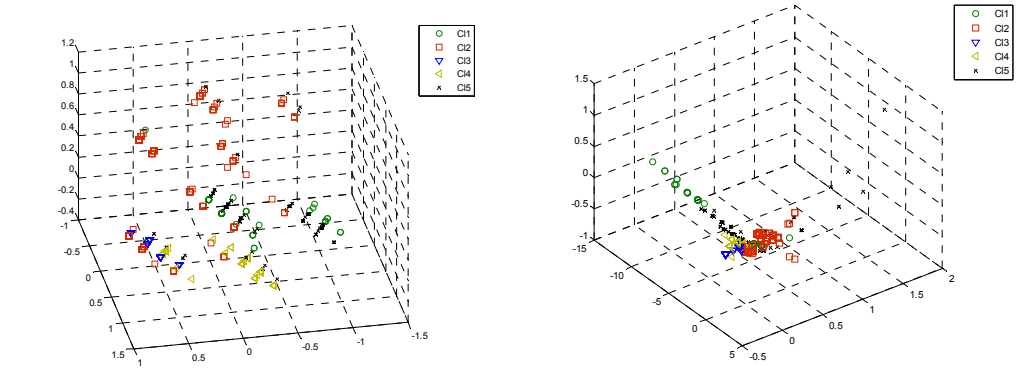

Figure 4.30 – Cluster projections on different components.
The left-most picture shows the clusters projected onto the 3 principal components of the original data and the right-most picture shows the affine transformed samples in the Fisher-induced space.

Accordingly, informed clustering with the Fisher information matrix as a metric, finds two different patient subgroups with similar disease progression and stratifies the patient population into distinct cohorts showing a progression in survival, also reflected by a localised distribution of risk scores estimated by either survival model. This approach has the merit of allowing a specific definition of the patient population, from which to forward the predict grouped survival, instead of inferring a threshold back from the log-rank separation index.

4.6.5. Comparison between the different stratification methodologies

Perhaps surprisingly, the four risk allocation methodologies broadly agree, although they are founded on very different principles. However, at the level of detail there are important differences. In particular, it is generally the case in breast cancer that the population of operable patients comprises a very well surviving group and another, thankfully a much smaller group, with especially poor survival. Nevertheless, it is the accurate discrimination and grouping of patients in the mid-surviving groups that is of most interest, since these two groups of patients are those likely to benefit most from better targeting of therapy.

The prognostic indices obtained with Cox proportional hazards, PI_{Cox} and with PLANN-ARD modelling, $PI_{PLANNARD}$, as well as the mode of the 6 variables found as the most predictive ones, for the 10 imputed data sets were used for the different stratification methodologies, for both the training and validation data set.

Using the bootstrap log-rank aggregation, the regression tree method and the unsupervised

clustering approach it were found 4 different risk groups. The clustering methodology based on learning metrics found 5 different risk groups. The two clustering methods did not display distinct observed survival measured by Kaplan-Meier actuarial estimates unlike the regression tree methodology and the log-rank bootstrap aggregation method, for both training and validation data sets. Therefore, it can be concluded that group separation is much better for regression tree and bootstrap log-rank methodology. These two methodologies have a very similar survival, for both prognostic indexes and for both training and validation data sets. However, looking at the log-rank pairwise values and KM curves, the bootstrap log-rank methodology has better separation between the risk groups. Although survival for both methods is similar, group membership is not the same, as it can be observed on Table 4.39 .

		Regression tree Cox				
		1	**2**	**3**	**4**	**Total**
Bootstrap log-rank Cox	**1**	291	73	1	0	365
	2	2	170	1	0	173
	3	0	92	69	5	166
	4	0	1	6	32	39
	Total	293	336	77	37	743

		Regression tree PLANN				
		1	**2**	**3**	**4**	**Total**
Bootstrap Log-rank PLANN	**1**	284	10	0	1	325
	2	9	166	5	1	181
	3	0	39	93	17	149
	4	0	2	6	80	88
	Total	293	247	104	99	743

		Regression tree Cox				
		1	**2**	**3**	**4**	**Total**
Bootstrap log-rank Cox	**1**	1371	381	0	0	1752
	2	20	900	18	1	939
	3	7	628	356	64	1055
	4	0	10	58	202	270
	Total	1398	1919	432	267	4016

		Regression tree PLANN				
		1	**2**	**3**	**4**	**Total**
Bootstrap Log-rank PLANN	**1**	1439	325	6	1	1771
	2	68	744	62	19	893
	3	3	146	554	192	895
	4	0	1	10	446	457
	Total	1510	1216	632	658	4016

Table 4.39 – Patients' cross tabulation between two different stratification methodologies. These methodologies are Regression tree and bootstrap log-rank aggregation. The left tables represent for the prognostic index obtained with Cox proportional hazards and the right tables represent for the prognostic index obtained with PLANN-ARD. The top tables are for the training data set and the bottom ones are for the validation data set.

For the training data set, with the exception of the 4th risk group, the bootstrap log-rank aggregation is generally more conservative in terms of patients' risk group allocations than the regression tree method. However, this analysis is found more for the prognostic index obtained with Cox proportional hazard, as the one obtained with PLANN-ARD, the risk group allocation is less sparse. This conclusion is also observed for the validation data set, when the

Cox proportional hazard is utilized as the prognostic index. As opposite, with the prognostic index obtained with PLANN-ARD, the regression tree is more conservative in terms of patient's allocations than the bootstrap log-rank aggregation method.

Consequently, the bootstrap log-rank method showed clearly the better discrimination in survival between the most and least surviving group, and is more conservative than the use of regression trees, compared to which it draws a substantial number of patients from group 2 into group 3. This effect is more pronounced when the linear survival estimator is used, in part reflecting the observation that the non-linear estimator, PLANN-ARD, is itself slightly more conservative than Cox regression with respect to these two risk groups.

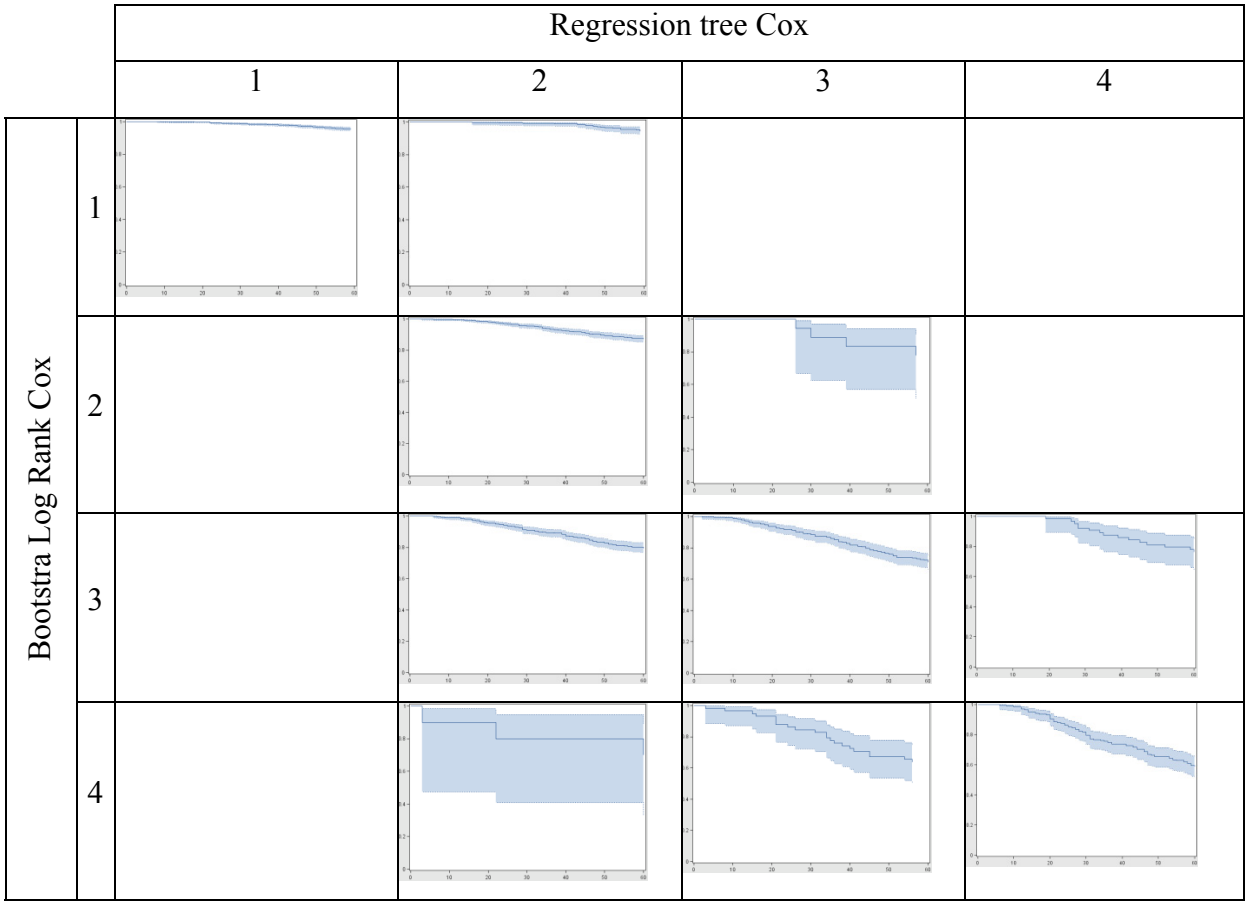

Figure 4.31 – Survival curves obtained for the patients' cross-tabulation. They were obtained with the regression trees stratification methodology and bootstrap log-rank stratification methodology for the Cox Proportional Hazards and for the validation data set.

In order to verify the patients' consistency allocated to the different stratification methodologies, regression tree and bootstrap log-rank, using prognostic risks, Cox proportional hazards and PLANN-ARD, the survival curves for the previously cross-tabulations are plotted on Figure 4.31 and Figure 4.32 , for the validation data set. Analysing the survival curves for both risk indexes, the obtained survival curves with the bootstrap log-rank stratification methodology are more consistent than the ones for regression trees methodology. This finding is more evident for Cox proportional hazards model.

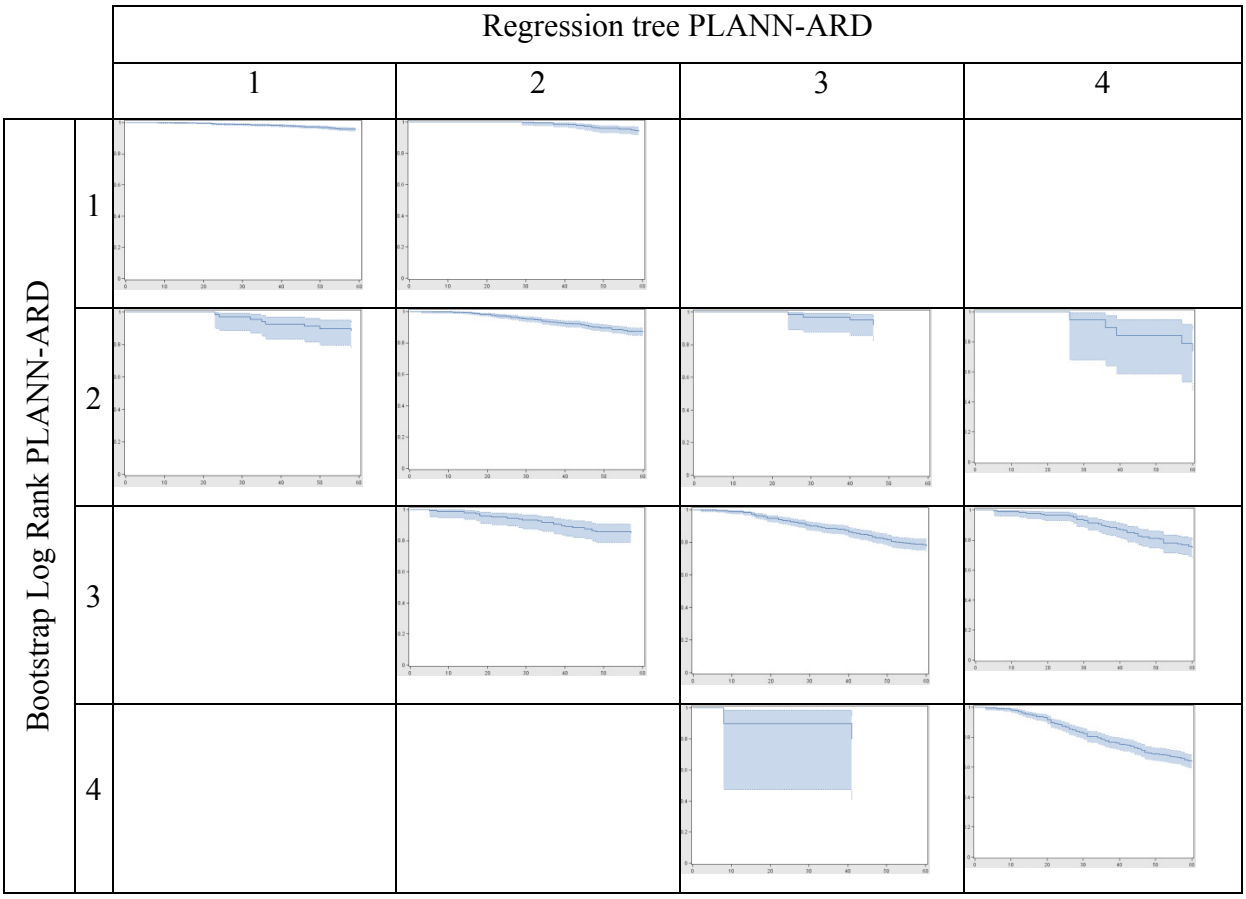

Figure 4.32 – Survival curves obtained for the patients' cross-tabulation. They were obtained with the regression trees stratification methodology and bootstrap log-rank stratification methodology for the PLANN-ARD and for the validation data set.

Comparing the KM curves for the PI_{Cox} and $PI_{PLANNARD}$ it can be confirmed that, for both algorithms (Bootstrap log-rank aggregation and regression tree), survival is lower for the risk groups obtained by using PI_{Cox}, for both training and validation data set. This conclusion is more evident for the training data set. However, the risk groups' patient allocation obtained with $PI_{PLANNARD}$ are more conservative than PI_{Cox}, for both training and validation data set,

because patients are allocated in higher risk groups, as it can be observed in Table 4.40 . This finding manifests itself more, using the regression tree stratification methodology.

		Cox				
		1	**2**	**3**	**4**	**Total**
PLANN-ARD	**1**	280	13	0	0	293
	2	12	235	0	0	247
	3	0	85	19	0	104
	4	1	3	58	37	99
	Total	293	336	77	37	743

		Cox				
		1	**2**	**3**	**4**	**Total**
PLANN-ARD	**1**	322	3	0	0	325
	2	43	137	1	0	181
	3	0	33	116	0	149
	4	0	0	49	39	88
	Total	365	173	166	39	743

		Cox				
		1	**2**	**3**	**4**	**Total**
PLANN-ARD	**1**	1329	181	0	0	1510
	2	64	1152	0	0	1216
	3	0	560	72	0	632
	4	5	26	360	267	658
	Total	1398	1919	432	267	4016

		Cox				
		1	**2**	**3**	**4**	**Total**
PLANN-ARD	**1**	1709	62	0	0	1771
	2	43	813	37	0	893
	3	0	64	824	7	895
	4	0	0	194	263	457
	Total	1742	939	1055	270	4016

Table 4.40 – Risk groups' cross tabulation between different models.
The left tables represents patients' cross tabulation for the regression tree method and the right tables represents patient's cross-tabulation using the Bootstrap log-rank aggregation, using the PI obtained with Cox and PLANN-ARD. The top tables are for the training data set and the bottom tables are for the validation data set.

In order to verify the patients' consistency allocated to the different prognostic risks, Cox proportional hazards and PLANN-ARD, using both regression tree and bootstrap log-rank stratification methodologies, the survival curves are plotted the Figure 4.33 and Figure 4.34. Analysing the survival curves it can be concluded that for each risk index, the survival curves obtained with the PLANN-ARD prognostic index are more consistent than the ones for Cox proportional hazards prognostic index. This finding is more evident for the regression tree stratification.

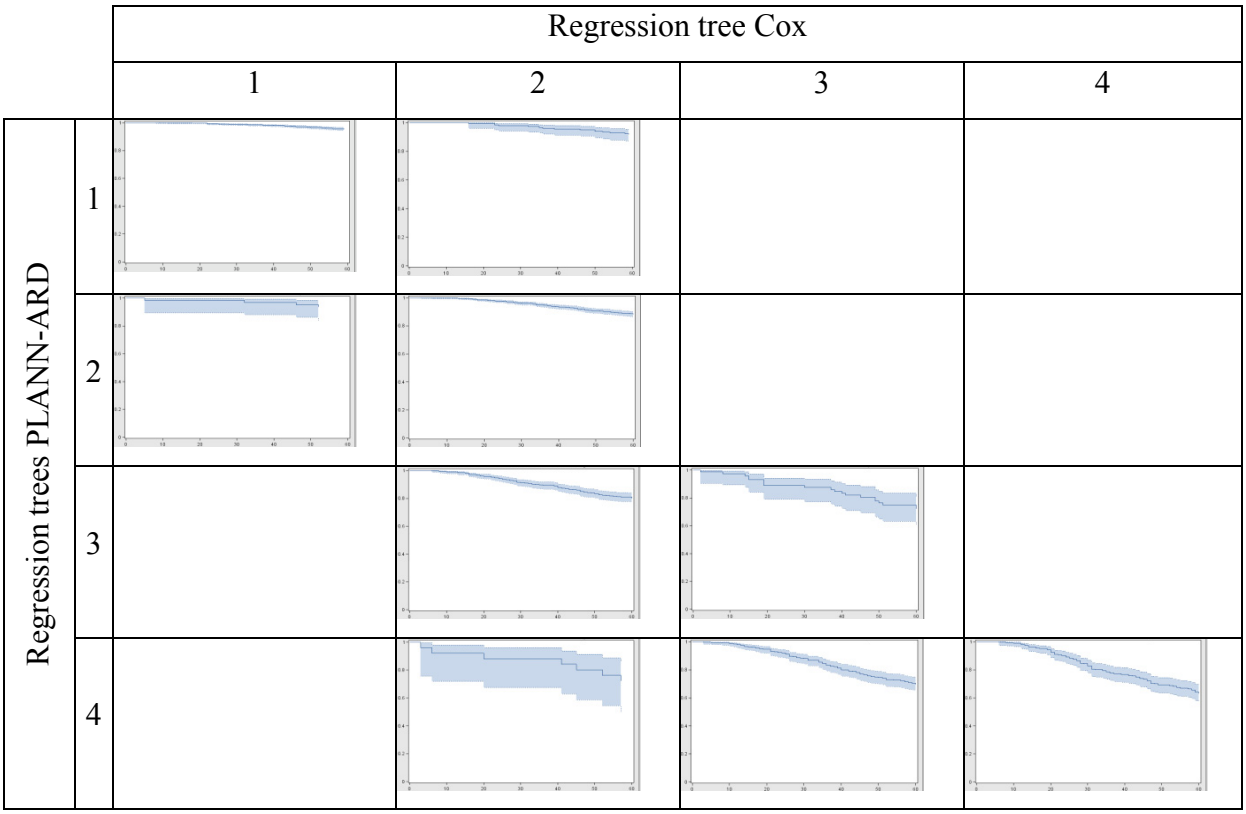

Figure 4.33 – Survival curves obtained for the patients' cross-tabulation. They were obtained with the regression trees stratification methodology for both indexes, Cox Proportional Hazards and PLANN-ARD, for the validation data set.

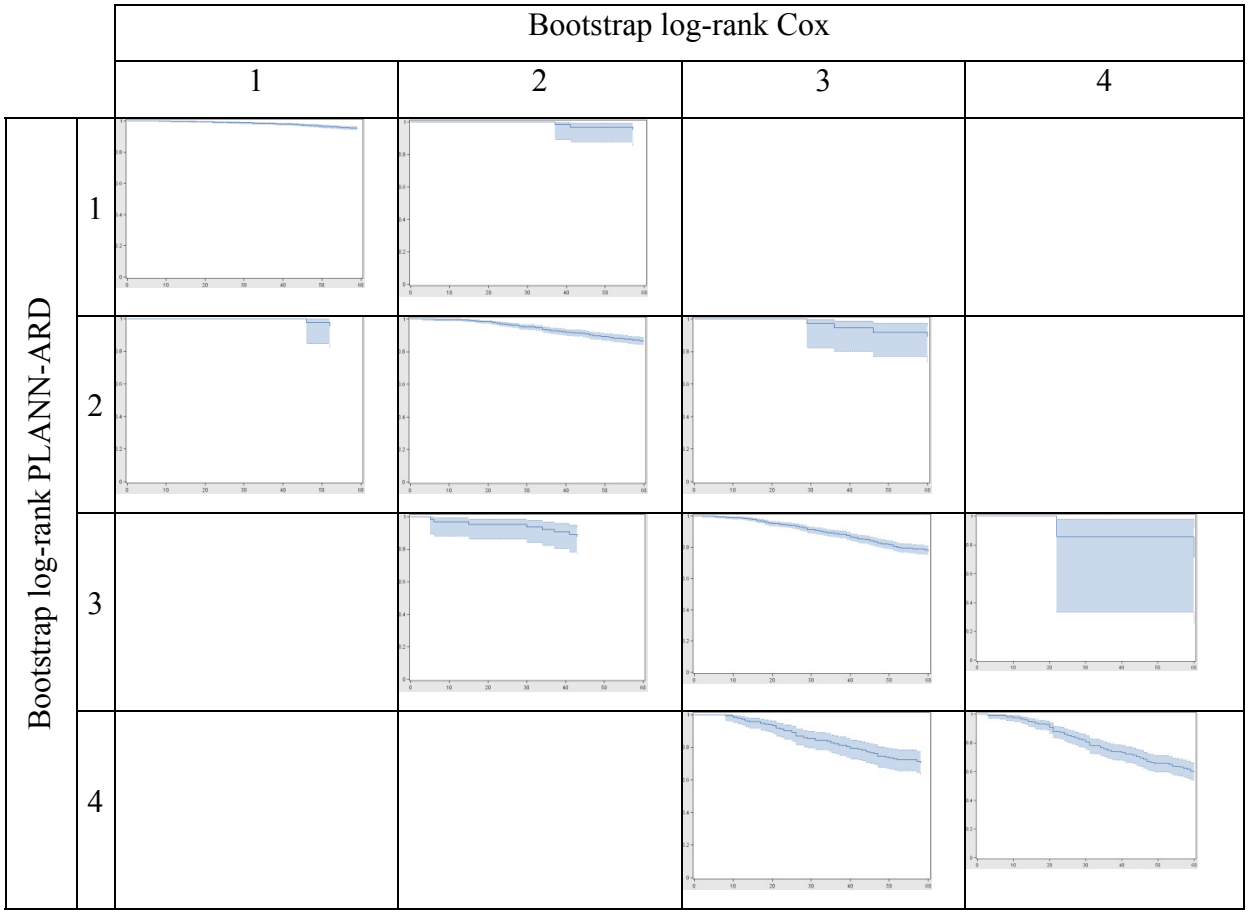

Figure 4.34 – Survival curves obtained for the patients' cross-tabulation. They were obtained with the bootstrap log-rank stratification methodology for both indexes, Cox Proportional Hazards and PLANN-ARD, for the validation data set.

Regarding all the risk groups, the most similar figures between the training and validation data set were achieved for Log-rank bootstrapping aggregation using PLANN as the prognostic model. The next more similar values are for the regression decision tree stratification methodology using also the PLANN as the prognostic model. Herewith there is a greater survival similarity at 5 years between the training and validation data set using the PLANN as a prognostic model.

4.7 - OSRE and CART rules comparison

For each rule extraction methodology, CART and OSRE, and for each model used, Cox proportional hazards modelling and PLANN-ARD, different rules can be obtained. These rules must be applied to new patients in order to obtain the risk group they belong.

For a specific model, all the rules obtained with regression trees methodology are mutually exclusive as opposed to the rules obtained with OSRE. In OSRE a patient can be classified for different rules and these rules can be in the same risk group or not. However, OSRE determines a rule hierarchy, which means that for each patient each rule is tested in turn and as soon as a rule is met for that patient profile it is not necessary going through the hierarchy. Even so, a patient can met the requirements of different rules of different risk groups. Here, it was chosen a conservative approach that is the patient must belong to the higher risk group. When OSRE rules were applied to the development data set, 22 patients were not classified for PLANN modelling and 23 patients were not classified for Cox modelling. When OSRE rules were applied to the validation data set, 227 patients were not classified for PLANN modelling and 205 patients were not classified for Cox modelling. This means that these patients were considered as outliers.

The rules obtained with both methodologies, OSRE and CART were compared and it was analysed that generally the methodologies derive the same number of rules. Therefore it cannot be confirmed that one methodology is more parsimonious than another.

For both rules extraction methodology it was necessary more rules to specify the patients belonging using the PLANN-ARD prognostic model than the Cox modelling. There were more similar rules between both methodologies when it is used the PLANN-ARD model rather than the proportional hazards modelling: 6 rules versus 2 rules for the development data set and 5 rules versus 1 rule for the validation data set.

As it was performed with patients' group risk membership, the rules' consistency can be also analysed, both for stratification methodologies and for the different prognostic models in order to verify more precisely which stratification methodology can perform better, in terms of rules.

The rules obtained with OSRE and CART methodology where compared for both Cox modeling and PLANN-ARD modeling, through the KM curves' analysis and the statistical difference between them. This analysis was performed for the development and validation data set, where for the second the 5 and 10 years of follow up were used.

Using the prognostic index obtained with Cox modelling, the CART rules are more consistent than OSRE rules, because there are more KM curves statistically different for the same rule for OSRE methodology than for CART methodology. For PLANN-ARD modelling this consistency couldn't be corroborated, as for development and validation there is not an evidence of more KM curves statistically different neither for OSRE nor for CART methodology.

The rules obtained with each stratification methodology were also compared in a different way, that is, the rules obtained using the Cox model and the rules obtained with PLANN-ARD model were compared for each OSRE and CART. Here it can be affirmed that the rules obtained with CART methodology are very similar between each other, 9 for development and 7 for validation data set. However, it was concluded for both, development and validation data set (5 and 10 years of follow up) that generally there is more consistency in rules obtained with PLANN-ARD than with Cox. The rules obtained with OSRE methodology are less similar, 7 for development and 5 for validation data set. Nevertheless the Cox rules are more consistent than the PLANN-ARD rules, for the development data set that and the contrary is verified for the validation data set, to both 5 and 10 years of follow up.

4.8 - Interval estimates of individual prognosis

Following the methodology previously explained on chapter 2 about the Individual prognostic predictions with confidence intervals using the PLANN-ARD Model, a survival distribution was obtained for the training data set, as it can be observed on Figure 4.35 . With this distribution a mean value as well as the 95% confidence intervals for each patient can be obtained.

The median survival estimates across all of the training data at the end of follow up is 0.8149 and the KM estimated survival is 0.8748. Box plots of personal survival estimates, split into the four PLANN-ARD prognostic groups obtained with the CART methodology are represented in Figure 4.36. The mean of the individual survival estimates in each group predicted by the PLANN-ARD model can be compared with the observed grouped mean survival estimated with the Kaplan-Meier method at 5 years of follow-up, shown for each risk group in Table 4.41. By the table inspection, model predictions are generally conservative, because these are generally lower for the different risk groups than the KM estimated values.

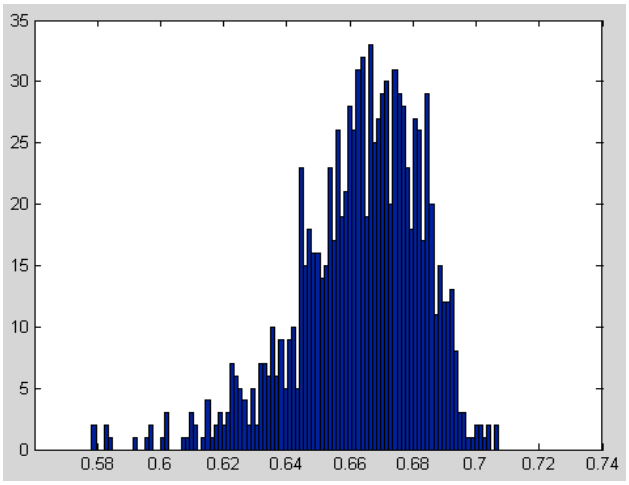

Figure 4.35 – Survival Distribution for an individual patient.
It was calculated from 1000 iterations of estimated survival for an individual patient for the training data set, where the mean survival and the 95% confidence intervals can be obtained.

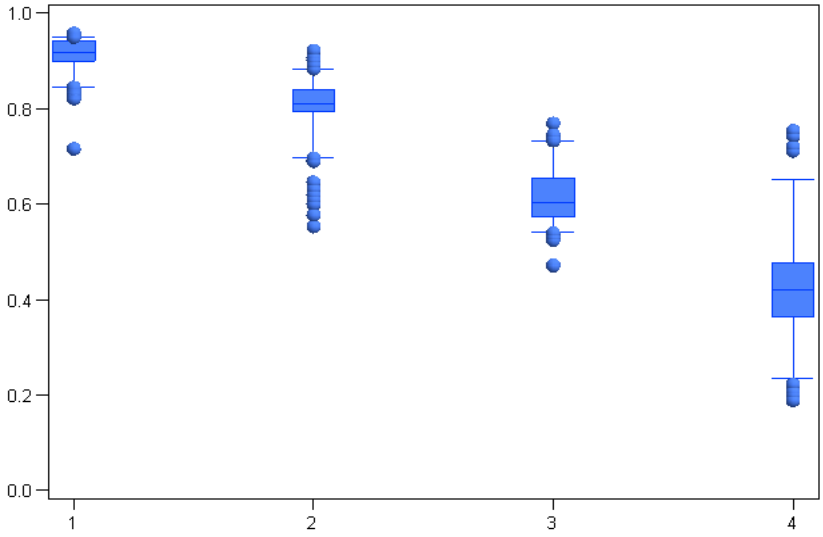

Figure 4.36 – Box plots of individual survival estimates to 5 years.
These are separated into PLANN-ARD CART risk groups.

Risk group	Mean predicted survival	95% Low individual survival estimate	95% High individual survival estimate	KM estimate	95 % Low KM estimate	95 % High KM estimate
1	0,917	0,846	0,949	0,983	0,958	0,99
2	0,809	0,697	0,883	0,903	0,851	0,93
3	0,603	0,541	0,732	0,798	0,695	0,857
4	0,421	0.235	0.65	0,566	0.448	0.646

Table 4.41 – Mean and 95% confidence intervals.
These values were computed for the individual PLANN-ARD survival estimate and the Kaplan-Meier estimated survival to 5 years, separated into PLANN-ARD CART risk groups.

4.9 - Comparison between the existent prognostic groups and the proposed ones

As previously mentioned, there are several clinical prognostic classification schemes proposed for breast cancer patients, some of which discriminate between the survival of different risk groups defined from the patient characteristics, such as the TNM staging system. The most widely used nowadays are the Nottingham prognostic index (NPI) and the consensus rules agreed by the St. Gallen group.

By cross-matching these prognostic classification schemes with the new prognostic indexes obtained with the Cox proportional hazards and PLANN-ARD followed by the regression tree stratification methodology it is possible to examine survival for patient sub-groups, using Kaplan Meier estimated survival curves, to uncover heterogeneity among the prognostic groups. This can be achieved to both training and validation data sets.

4.9.1. Comparison between NPI with Cox and PLANN-ARD modelling

Whereas data collected at Christie Hospital are categorical, descriptive statistics from a complementary data set also from Christie Hospital gave a mode for *pathological size* groups 1 and 2 as 1.1 and 2 cm, respectively. Therefore, these values were used to calculate the NPI score. In order to keep consistency between the Christie and BCCA data set, the NPI was also calculated with the same mode used before for the different groups of *pathological size*.

Values of the NPI index may be split into as many as five groups from excellent (group1) through good (2), moderately good (3a), moderately poor (3b) and poor (4) expected outcome. However, in this analysis we analyzed only 4 risk groups, where the moderate groups 3a and 3b are combined following common clinical practice resulting in three cut-off points at 2.41, 3.41 and 5.41.

Figure 4.37 represent the KM curves, applying the NPI to Christie and BCCA data sets. Comparing the NPI grouped survival with the proportional hazards modeling and PLANN-ARD modeling for both data sets, it can be verified that they produce almost identical grouped survival, with a higher evidence for PLANN modeling. For both data sets it is clear that the NPI groups have slightly higher survival to 5 years than the corresponding PLANN-ARD and Cox groups as well as different population sizes.

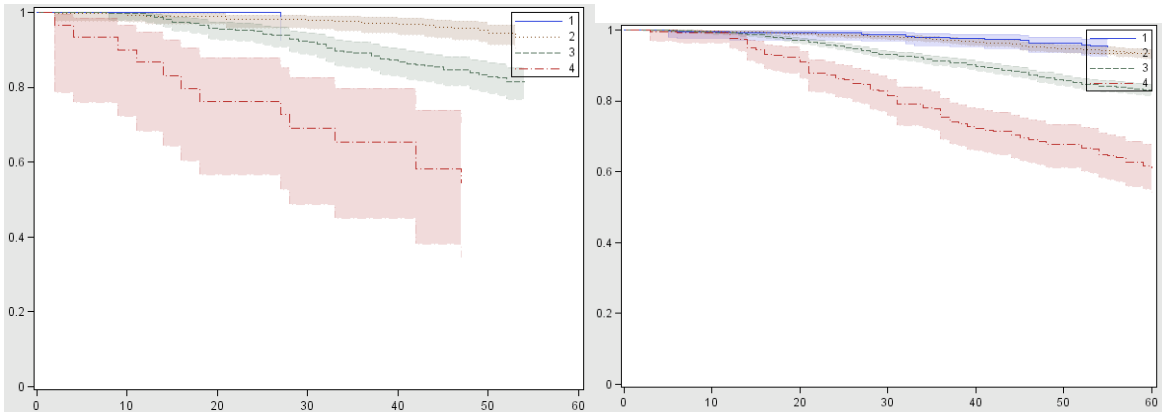

Figure 4.37 – KM survival curves for the NPI.
The NPI formula was applied to Christie Hospital and BCCA data sets, left and right picture respectively.

To investigate the correlation between the patients selected by the two methodologies, the three prognostic indexes were cross-tabulated in Table 4.42 . Inspecting these tables, there are a considerable number of patients that do not belong to the same risk group. However, it is difficult to define that one stratification methodology is more conservative in terms of patient allocation than the other. That conclusion is only valid for risk group 2 and 4, for both comparisons, NPI with Cox and NPI with PLANN-ARD, where it can be observed that PLANN-ARD is more conservative than NPI. This is observed for both training and validation data sets.

		Cox				
		1	**2**	**3**	**4**	**Total**
NPI	**1**	74	17	0	0	91
	2	168	113	6	1	288
	3	51	205	55	23	334
	4	0	1	16	13	30
	Total	293	336	77	37	743

		PLANN-ARD				
		1	**2**	**3**	**4**	**Total**
NPI	**1**	65	25	1	0	91
	2	165	101	21	1	288
	3	63	121	81	69	334
	4	0	0	1	29	30
	Total	293	247	104	99	743

		Cox				
		1	**2**	**3**	**4**	**Total**
NPI	**1**	274	6	0	1	281
	2	881	254	9	7	1151
	3	243	1494	252	52	2041
	4	0	165	171	207	543
	Total	1398	1919	432	267	4016

		PLANN-ARD				
		1	**2**	**3**	**4**	**Total**
NPI	**1**	238	37	1	5	281
	2	903	205	25	18	1151
	3	369	921	480	271	2041
	4	0	53	12	364	543
	Total	1510	1216	632	658	4016

Table 4.42 – Cross-tabulation between different classification schemes.
These classification schemes are NPI, Cox proportional hazards and PLANN-ARD for the training and validation data sets, top and bottom tables respectively.

In order to discover which of the prognostic models (comparing with NPI) had homogeneous groups of patients, indicated by consistent survival curves from one matrix plot to the next in either the rows or the columns of the matrix, the patient groups in terms of KM estimated survival within each matrix cell was examined, as in Figure 4.38 and Figure 4.39. For all three models, the higher and the lower survival groups have homogenous survival curves. Analogously for all models risk group 2 and 3 present a more heterogeneous set of survival curves with survival decreasing as the risk group in the corresponding prognostic index decrease.

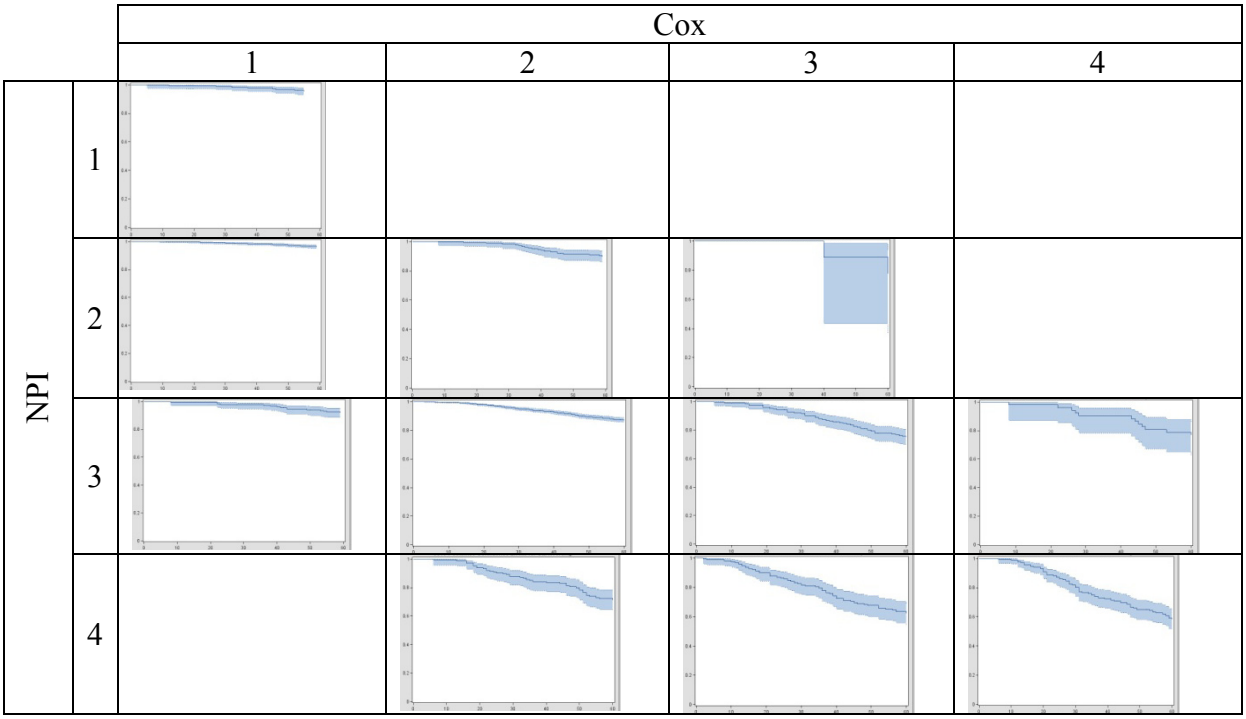

Figure 4.38 – Matrix of KM curves for NPI vs Cox for the validation data set.

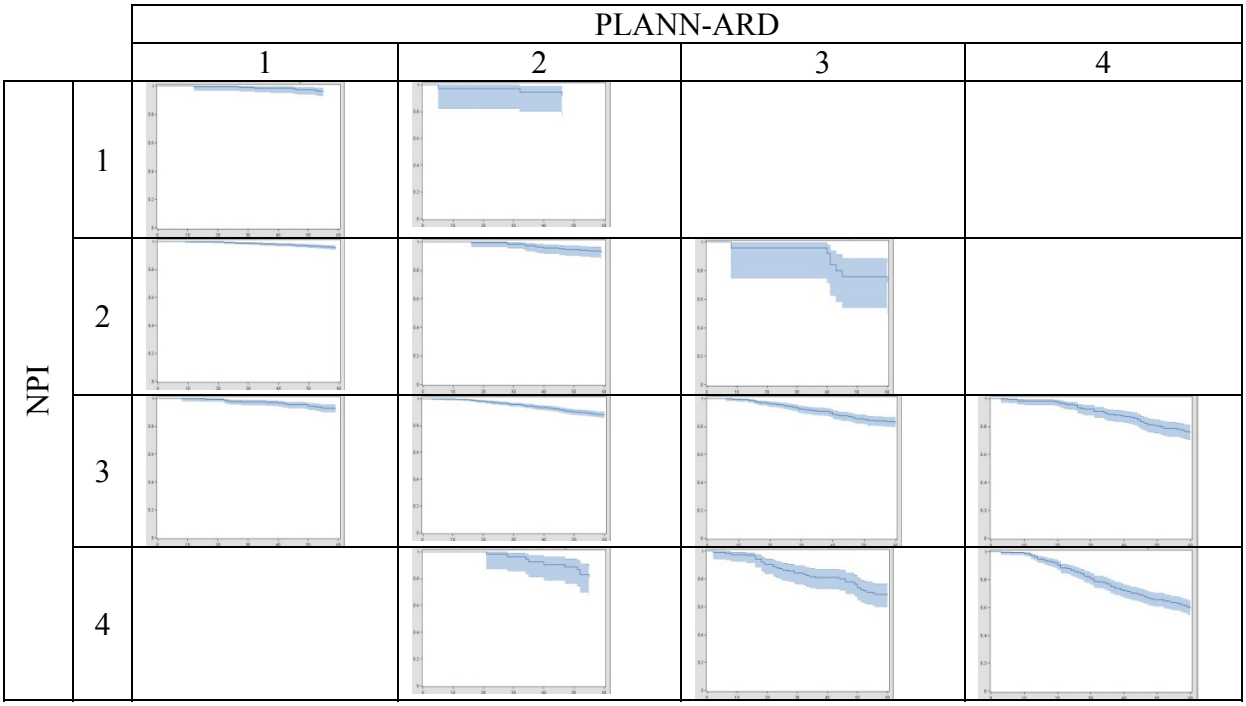

Figure 4.39 – Matrix of KM curves for NPI vs PLANN-ARD for the validation data set.

4.9.2. Comparison between TNM with Cox and PLANN-ARD modelling

The study developed for the NPI classification scheme was also applied to the TNM Classification of Malignant Tumours (TNM). Here three different risk groups were considered: Group 1, characterized by T1N0M0; Group 2 characterized by T2N0M0 and Group 3 characterized by T3N0M0 or T1N1M0, T2N1M0, or T3N1M0. Figure 4.40 represents the KM curves, applying the TNM to Christie data set and Table 4.43 the respective log-rank pairwise values. This staging system cannot be applied to the validation data set because the regional lymph nodes variable it is necessary to compute these risk groups, and it is not available for the BCCA data set. It can be examined that all the KM curves have significant distinct survival, existing however an overlap of the risk group 2 and 3 curves' from the start of the follow up until the 30th month. Table 4.44 corresponds to the cross-tabulation between the TNM risk groups with the ones established for Cox proportional hazards and PLANN-ARD modeling using the regression tree stratification methodology. It can be noticed that, as it exists less of a risk group, the membership is very sparse, being more remarkable for the second risk group, for both modeling.

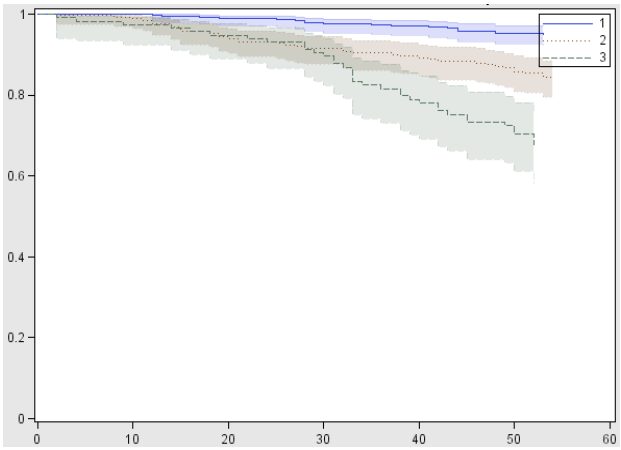

Figure 4.40 – TNM KM survival curves applied to the Christie Hospital data set.

	Risk Groups	1	2
		X ^2 (sig.)	*X* ^2 (sig.)
Log Rank (Mantel -Cox)	2	20.8866 (0.0000)	-
	3	67.6241 (0.0000)	12.0643 (0.0005)

Table 4.43 – Log-rank pairwise values for TNM applied to the training data set.

		Cox				
		1	**2**	**3**	**4**	**Total**
TNM	**1**	229	137	6	0	372
	2	45	144	47	19	255
	3	19	55	24	18	116
	Total	293	336	77	37	743

		PLANN-ARD				
		1	**2**	**3**	**4**	**Total**
TNM	**1**	237	118	8	9	372
	2	37	95	71	52	255
	3	19	34	25	38	116
	Total	293	247	104	99	743

Table 4.44 – Cross-tabulation between different classification schemes.
These classification schemes are TNM, Cox proportional hazards and PLANN-ARD using the regression tree stratification methodology for the training data set.

4.9.3. Comparison between St. Gallen with Cox and PLANN-ARD modelling

In order to achieve consistency between Christie Hospital and BCCA data set in the use of the consensus rules agreed by the St. Gallen group the Christie Hospital criteria was used (i.e. without using age), Table 4.45 . The Kaplan-Meier curves were obtained for 5 and 10 years of follow up for the BCCA data set and 5 years for Christie Hospital data set.

Figure 4.41 represents the KM curves, applying the consensus rules agreed by the St. Gallen group to Christie and BCCA data sets. All but risk group 1 and 2 from the training data set show distinct survival, as these have a p-value of 0.0599. The survival curves for the training and validation data set are very similar where for the 10 years of follow up, the survival is lower, as it was expected. The risk group membership was also compared with the Cox proportional hazards and PLANN-ARD modeling followed by the regression tree stratification methodology, presented in Table 4.46. It can be noticed that, as it exists less of a risk group, that the membership is very sparse, for both training and validation data set. However, the membership ratio presented on each cell of the training cross-tabulation is very similar to the ratio presented on the validation data set. Moreover, it can also be concluded that this membership relation is less sparse than the one obtained for the TNM classification scheme.

Low risk	Nodes involved=1 and Pathsize=1 and Histological grade=1
Intermediate risk	Nodes involved=2 and Oestrogen=2
	Nodes involved=1 and (Pathsize=2 or Histological grade=2 or 3)
High risk	(Nodes involved=2 and Oestrogen=1) or Nodes involved=3

Table 4.45 – Risk group consensus rules agreed by the St. Gallen group.
The known consensus rules were adapted in order to use the variables available at Christie Hospital and BCCA data sets.

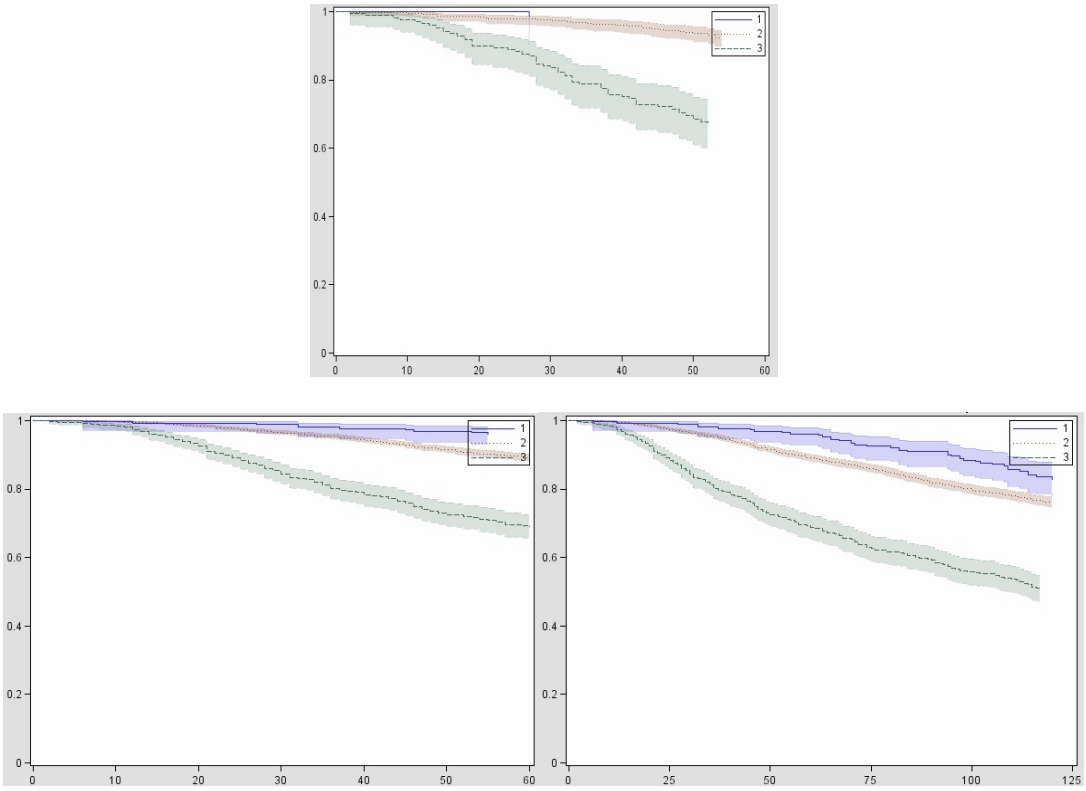

Figure 4.41 – Consensus rules agreed by the St. Gallen group KM survival curves.
The rules were applied to the Christie Hospital data set and BCCA data set, top and bottom figures, respectively. The bottom left picture represents a 5 year follow up and the bottom right picture a 10 year follow up.

		Cox				
		1	**2**	**3**	**4**	**Total**
St. Gallen	**1**	57	14	0	0	71
	2	223	248	24	5	500
	3	13	74	53	32	172
	Total	293	336	77	37	743

		PLANN-ARD				
		1	**2**	**3**	**4**	**Total**
St. Gallen	**1**	55	16	0	0	71
	2	227	190	61	22	500
	3	11	41	43	77	172
	Total	293	247	104	99	743

		Cox				
		1	**2**	**3**	**4**	**Total**
St. Gallen	**1**	248	5	0	0	253
	2	1116	1680	232	81	3109
	3	34	234	200	186	654
	Total	1398	1919	432	267	4016

		PLANN-ARD				
		1	**2**	**3**	**4**	**Total**
St. Gallen	**1**	225	24	0	4	253
	2	1252	1080	510	267	3109
	3	33	112	122	387	654
	Total	1510	1216	632	658	4016

Table 4.46 – Cross-tabulation between different classification schemes.
These classification schemes are the consensus rules agreed by the St. Gallen group, Cox proportional hazards and PLANN-ARD using the regression tree stratification methodology for the training and validation data sets, on the top and bottom tables respectively.

4.10 - PLANN-ARD prognostic indexes and comparison with Cox prognostic index

In prognostic modelling it is important to define an adequate prognostic index that ranks patients by the severity of illness. The output of the PLANN-ARD modelling is the hazard for each patient and for each time of follow up. Four different prognostic index calculations for the PLANN-ARD modelling were considered, analysed and compared in the next sections. Moreover, they were compared with the prognostic index obtained with Cox proportional hazards modelling.

4.10.1. Analysis of the different PLANN-ARD prognostic indexes calculation

The definition of PLANN-ARD prognostic index has been achieved after the analysis of different proposals. At the beginning there were four hypotheses to define the prognostic index, such as:

1. The mean of the hazard $h_p(t_k, x_p)$ of the 10 imputed data sets (characterized as PI A hereinafter).
2. The $\ln\left(\frac{h_p(x_p,t_k)}{1-h_p(x_p,t_k)}\right)e$, where $h_p(t_k, x_p)$ is the mean of the hazard of the 10 imputed data set (characterized as PI B hereinafter)
3. The mean of $S(t-1)\times\ln\left(\frac{h_p(x_p,t_k)}{1-h_p(x_p,t_k)}\right)$ of the 10 imputed data sets (characterized as PI C hereinafter).
4. The *ln(-ln(1-CCI)) = ln(-ln(S(t)))* (characterized as PI D hereinafter).

The different prognostic indexes calculations must be compared in order to verify the differences between them. However, patients must be stratified in different risk groups. This allocation into different risk groups was developed with a robust bootstrap log-rank aggregation method, where it identifies the *cutpoints* that separate the patients into statistically significant risk groups by overall mortality, based on the different prognostic index mentioned before. All the different prognostic indexes applied to the bootstrap methodology finalized with 4 different risk groups. This analysis was computed for the previously identified model,

with 6 significant variables and the KM curves are plotted in Figure 4.42. Comparing all the survival curves it can be concluded that they broadly agree in terms of survival.

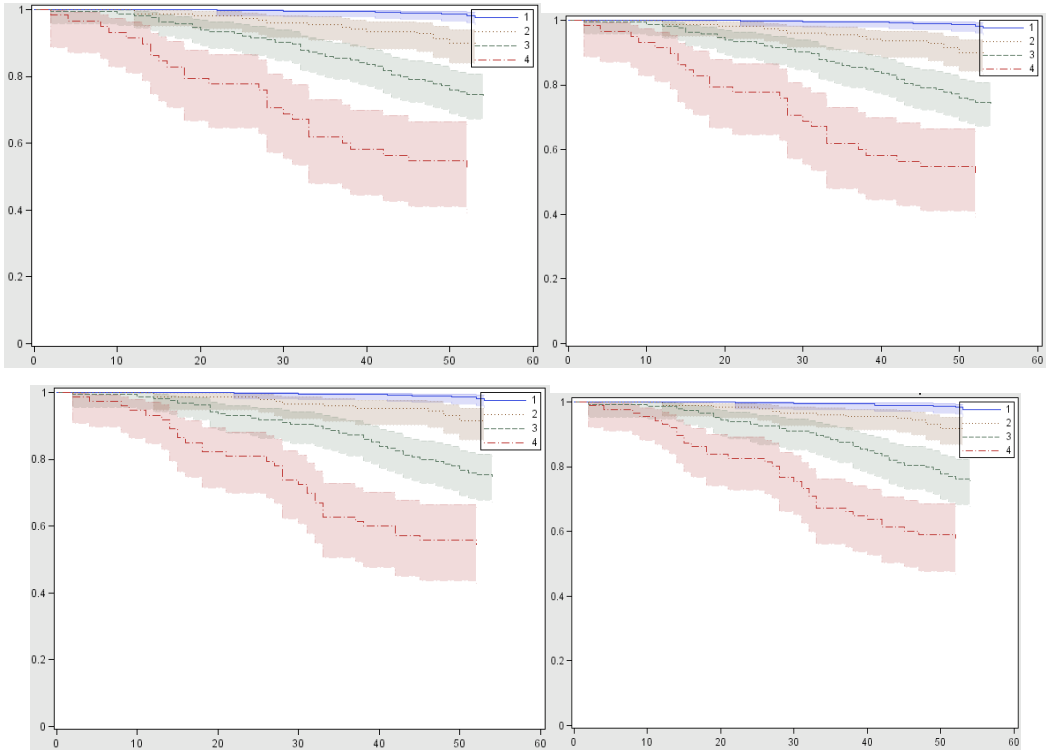

Figure 4.42 – KM curves for the different PI calculation for the training data set. The top left picture represents the mean hazard prognostic index. The top right picture represents the PI B prognostic index. The bottom left picture represents the PI C prognostic index and the bottom right picture represents the CCI prognostic index

The cross-tabulations for the different risk group membership was obtained (Table 4.47) . It can be noticed that the patients risk group allocation is very similar for all the prognostic indexes, especially for the prognostic index obtained as the mean of the hazard and the PI B, where it only exists 2 patients which do not belong to the same risk group. Both, the PI C prognostic index and the PI D prognostic index are more conservative than the others in terms of patient allocation. However, the PI D prognostic index is more conservative than PI C prognostic index. This conclusion increases our expectation that the CCI prognostic index is the one that must be used.

		Mean hazard				
		1	2	3	4	Total
Mean log (hazard)	1	357	0	0	0	357
	2	2	156	0	0	158
	3	0	0	169	0	169
	4	0	0	0	0	59
	Total	359	156	169	59	743

		Mean hazard				
		1	**2**	**3**	**4**	**Total**
Mean log hazard*survival	**1**	354	4	0	0	358
	2	5	142	3	0	150
	3	0	10	151	0	161
	4	0	0	15	59	74
	Total	359	156	169	59	743

		Mean Hazard				
		1	**2**	**3**	**4**	**Total**
CCI	**1**	325	0	0	0	325
	2	34	147	0	0	181
	3	0	9	140	0	149
	4	0	0	29	59	88
	Total	359	156	169	59	743

		Mean log hazard*survival				
		1	**2**	**3**	**4**	**Total**
CCI	**1**	325	0	0	0	325
	2	33	147	1	0	181
	3	0	3	146	0	149
	4	0	0	14	74	88
	Total	358	150	161	74	743

Table 4.47 – Cross tabulations between different PI calculations for the training data set.

4.10.2. Cox proportional hazards and PLANN-ARD prognostic indexes comparison

The previously mentioned prognostic indexes obtained with PLANN-ARD modelling were compared with the prognostic index obtained with Cox proportional hazards modelling. It must be mentioned that the model used to compute this prognostic index incorporate the missing impute data and was developed for the 6 established variables as the most predictive ones.

Figure 4.43 presents a scatter plot comparing the different prognostic indexes obtained with PLANN-ARD and Cox proportional hazards. It can be noticed that for all the prognostic indexes, the Cox prognostic index is non-linear related with PLANN-ARD. In particular it seems that the PLANN-ARD model compresses all but the mean of the hazard prognostic index calculation, in the extreme sectors but extends the dynamic range for the middle sector. This can be because the non-linear algorithm implicitly models interactions between covariates. However, it can be analysed that the prognostic index range obtained with Cox is equal to the PI C prognostic index and higher than both the PI B and the PI D prognostic index.

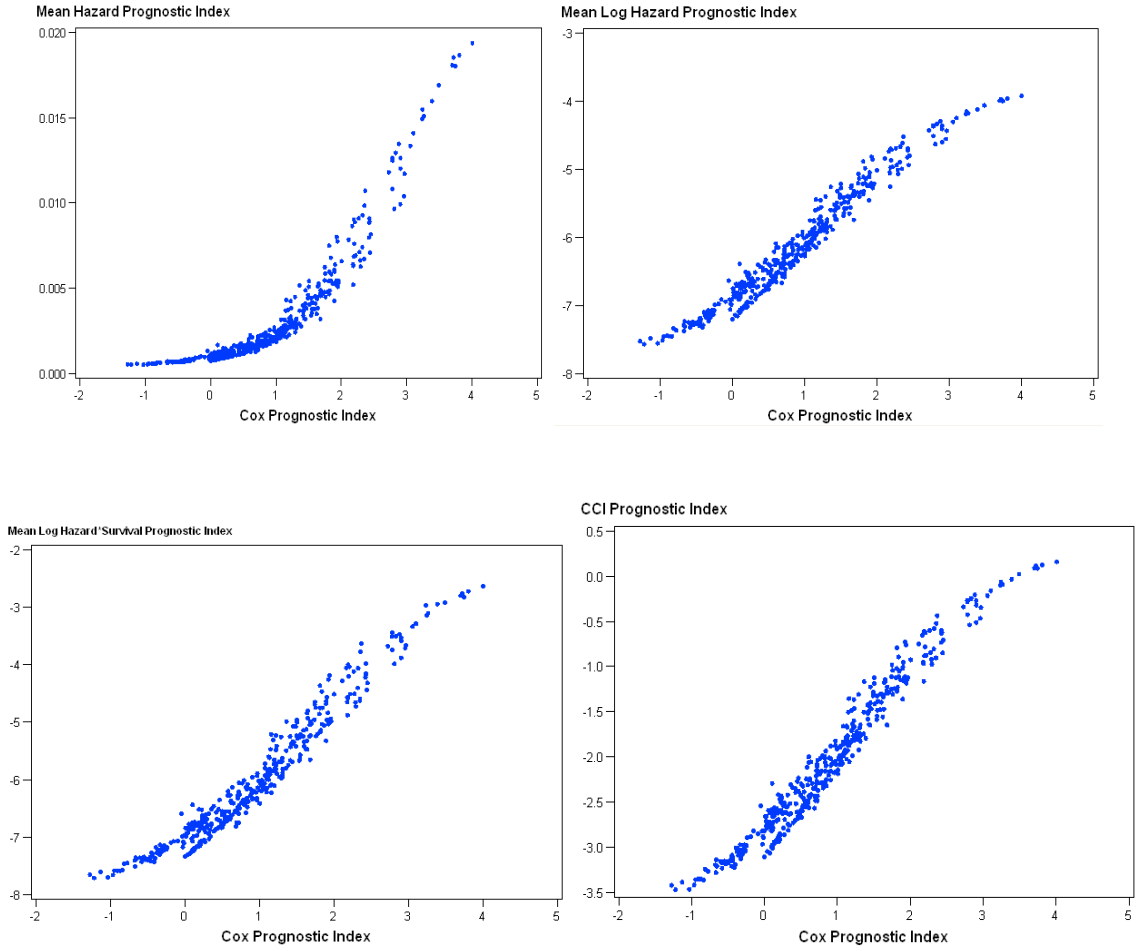

Figure 4.43 – Scatter plots comparing different prognostic indices.
Comparison between different PLANN-ARD prognostic index calculations with the Cox proportional hazards prognostic index for the training data set.

4.11 - Models with different variables' comparison

Four good breast cancer prediction models were previously identified, one with 5 variables (*Histological Grade*, *Histological Type*, *Oestrogen*, *Age*, *Nodes ratio*), one with 6 variables (*Histological Grade*, *Histological Type*, *Oestrogen*, *Age*, *Pathological size*, *Nodes Ratio*), one with also 6 variables (*Histological Grade*, *Histological Type*, *Oestrogen*, *Age*, *Pathological size*, *Nodes involved*) and another one with 7 variables (*Histological Grade*, *Histological Type*, *Oestrogen*, *Age*, *Pathological size*, *Nodes Ratio*, *Menopausal status*).

It was already demonstrated that all models have a high similarity on the prognostic indexes range, predict well, are very well calibrated and have a very good discrimination.

However, the 6 variable model including *Nodes Ratio* was considered the one that predicts better.

Nevertheless, the patients' risk group allocation must be also analysed for the different models. This allocation into different risk groups was developed using the robust bootstrap log-rank aggregation method, where it identifies the cutpoints that separate the patients into statistically significant risk groups by overall mortality, based on the prognostic index given by the βx of the Cox proportional hazards modelling. Table 4.48 represents the different cross-tabulations.

		6 variables Model (Nodes ratio)				
		1	**2**	**3**	**4**	**Total**
5 variables Model	**1**	292	82	2	0	376
	2	33	94	115	4	246
	3	0	5	31	52	88
	4	0	0	2	32	33
	Total	325	181	149	88	743

		6 variables Model (Nodes involved)				
		1	**2**	**3**	**4**	**Total**
5 variables Model	**1**	277	71	28	0	376
	2	50	71	121	4	246
	3	0	9	48	31	88
	4	0	0	9	24	33
	Total	327	151	206	59	743

		7 variables Model				
		1	**2**	**3**	**4**	**Total**
5 variables Model	**1**	285	85	6	0	376
	2	31	101	114	0	246
	3	0	1	81	6	88
	4	0	0	3	30	33
	Total	316	187	204	36	743

		6 variables Model (Nodes ratio)				
		1	**2**	**3**	**4**	**Total**
7 variables Model	**1**	288	28	0	0	316
	2	37	123	27	0	187
	3	0	30	122	52	204
	4	0	0	0	36	36
	Total	325	181	149	88	743

		6 variables Model (Nodes involved)				
		1	**2**	**3**	**4**	**Total**
7 variables Model	**1**	267	37	12	0	316
	2	58	80	47	2	187
	3	2	34	136	32	204
	4	0	0	1	25	36
	Total	327	151	206	59	743

		6 variables Model (Nodes ratio)				
		1	**2**	**3**	**4**	**Total**
6 variables Model (Nodes involved)	**1**	275	52	0	0	327
	2	41	76	33	1	151
	3	9	53	107	37	206
	4	0	0	9	50	59
	Total	325	181	149	88	743

Table 4.48 – Cross tabulation between patients' risk group allocation. This was performed for the 5, 6 and 7 variables different models. There are two different models with 6 variables, one including Nodes ratio variable and another one including Nodes involved variable

Analysing all the cross-tabulations in terms of risk allocation, it is verified the the 5 variables model is less conservative than the other 3 models, that is, the patients are allocated in lower risk groups, comparing with the other models. The 7 variables model is generally more conservative in terms of patient allocation than both the 6 variables models, with the exception of the first risk group. The patients in the cross-tabulation between the 6 variables model including *Nodes Ratio* and the 6 variables model including *Nodes involved* are very spread, making it impossible to define the most conservative model. There is however an exception, which is in the higher risk group, where the model including *Nodes Ratio* is more conservative than the model including *Nodes involved*. Figure 4.44 presents the KM curves for the different variables' models.

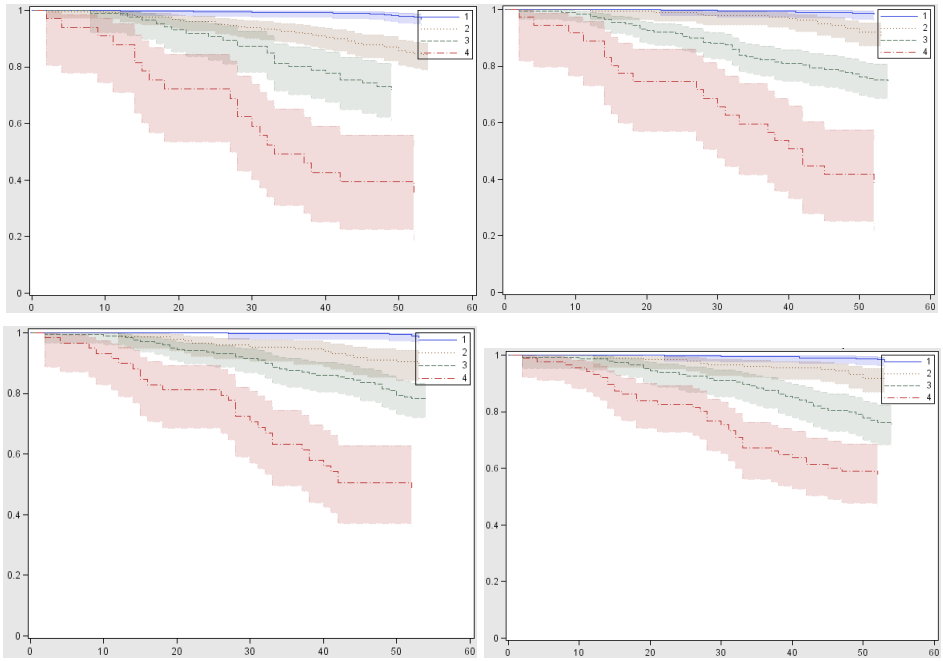

Figure 4.44 – KM curves for the different four variables models, for the training data set. The top left pictures represents the 5 variables model; the top right picture represents the 7 variables model. The bottom left picture represents the 6 variables model including Nodes involved and the bottom right picture the 6 variables model including Nodes ratio.

Analysing Figure 4.44 it can be supported that all curves have a distinct survival between the groups and have a very similar shape and survival between them, with the exception of the 4th risk group where the survival is higher for the 6 variables model including *Nodes Ratio*. This analysis increases the study already performed that has determined the 6 variables model including *Nodes Ratio* as the most predictive model.

4.12 - Treatments distribution

The treatments applied to the patients on the training and validation data set were recorded and divided in four categories: None, Hormone, Chemotherapy or Combined. Considering their distributions by the risk groups, applying the regression tree stratification methodology to PLANN-ARD and Cox prognostic models, it can be noticed any trend about the treatments that should be applied. Table 4.49 and Table 4.50 contain these distributions for the training data set, while Table 4.51 contains the distribution considering the NPI classification scheme, applied to the training data set.

	1	2	3	4	Total
None	170 (58%)	125 (37%)	7 (10%)	4 (11%)	306
Hormone	111 (38%)	161 (48%)	55 (71%)	18 (49%)	345
Chemo	11 (3%)	50(15%)	15 (19%)	15(40%)	91
Combined	1(1%)	0	0	0	1
Total	293	336	77	37	743

Table 4.49– Distribution of the different treatments for the different risk groups. These risk groups were obtained for Cox modelling followed by the regression tree stratification methodology, for the training data set.

	1	2	3	4	Total
None	166 (57%)	106 (43%)	24 (23%)	10 (10%)	306
Hormone	112 (38%)	115 (47%)	59 (57%)	59 (60%)	345
Chemo	14 (4%)	26 (10%)	21 (20%)	30 (30%)	91
Combined	1(1%)	0	0	0	1
Total	293	247	104	99	743

Table 4.50 – Distribution of the different treatments for the different risk groups. These risk groups were obtained for PLANN modelling followed by the regression tree stratification methodology, for the training data set.

	1	2	3	4	Total
None	67 (74%)	184 (64%)	55 (16%)	0	306
Hormone	24 (26%)	97 (34%)	207 (62%)	17 (57%)	345
Chemo	0	7 (2%)	71 (21%)	13 (43%)	91
Combined	0	0	1(1%)	0	1
Total	91	288	334	30	743

Table 4.51 – Distribution of the different treatments for the different risk groups. These risk groups were obtained for NPI classification scheme, for the training data set.

For the three risk group stratifications (Cox, PLANN-ARD and NPI), the treatment with the biggest patients' ratio in the 1st risk group is None, with 58%, 57% and 74%, respectively. For the 2nd risk group membership, Cox stratification has Hormone as the biggest ratio for treatment, NPI stratification has None as the biggest ratio and PLANN, has a very similar ratio for None and Hormone Treatment. For the 3rd and 4th risk groups membership, the three different stratifications (Cox, PLANN and NPI) have Hormone as the biggest treatment ratio.

Table 4.52 indicates, for each model, Cox, PLANN-ARD and NPI, the risk groups with the higher number of patients treated with each treatment.

	Cox	**PLANN-ARD**	**NPI**
None	Risk group 1	Risk group 1	Risk group 2
Hormone	Risk group 2	Risk group 1 and 2	Risk group 3
Chemo	Risk group 4	Risk group 4	Risk group 4
Combined	Risk group 1	Risk group 1	Risk group 3

Table 4.52 – Higher Treatments' Risk group Ratio for different models, for the training data set.

The biggest coherency for treatments is for Chemotherapy where the majority of patients are allocated in the same risk group for the different stratifications. On the other hand, Hormone treatment is the one where is more spread through the 4 risk groups.

The results previously obtained on the training data set were also obtained for the validation data set. Table 4.53 and Table 4.54 contain the treatment distributions for the validation data set, while table 4.54 contains the distribution considering the NPI classification scheme, applied to the validation data set.

	1	**2**	**3**	**4**	**Total**
None	893 (64%)	801 (42%)	89 (20%)	29 (11%)	1812
Hormone	362 (26%)	592 (31%)	171 (40%)	107 (40%)	1232
Chemo	83 (6%)	342(18%)	112 (26%)	70(26%)	607
Combined	60(4%)	184(9%)	60(14%)	61(23%)	365
Total	1398	1919	432	267	4016

Table 4.53 – Distribution of the different treatments for the different risk groups. These risk groups were obtained for Cox modelling followed by the regression tree stratification methodology, for the validation data set.

	1	2	3	4	Total
None	969 (64%)	550 (45%)	174 (28%)	119 (18%)	1812
Hormone	375 (25%)	398 (33%)	228 (36%)	231 (35%)	1232
Chemo	97 (6%)	160 (13%)	165 (26%)	185 (28%)	607
Combined	69(5%)	108(9%)	65(10%)	123(19%)	365
Total	1510	1216	632	658	4016

Table 4.54 – Distribution of the different treatments for the different risk groups. These risk groups were obtained for PLANN modelling followed by the regression tree stratification methodology, for the validation data set.

	1	2	3	4	Total
None	235 (84%)	823 (72%)	721 (35%)	33(6%)	1812
Hormone	39 (14%)	243 (21%)	747 (37%)	203 (37%)	1232
Chemo	7 (2%)	61 (5%)	371 (18%)	168 (31%)	607
Combined	0	24(2%)	202(10%)	139(26%)	365
Total	281	1151	2041	543	4016

Table 4.55 – Distribution of the different treatments for the different risk groups. These risk groups were obtained for NPI classification scheme, for the validation data set.

As cofirmed in training data set data set, for the three risk group stratifications (Cox, PLANN and NPI), the treatment with the biggest patients' ratio in the 1st risk group is None, with 64%, 64% and 84%, respectively. For the 2nd risk group membership, all the three different methodologies have a biggest ratio for None, with 42%, 45% and 72% respectively, which is not similar with the training data set for Cox and PLANN stratification. For the 3rd and 4th risk group membership, Cox and PLANN stratifications have Hormone as the biggest treatment ratio, with 40% and 36%, respectively for the 3rd risk group and 40% and 35% for the 4th risk group. NPI, opposed to the training data set, has for the 3rd group a similar ratio for None and Hormone and for the 4th risk group a similar ratio for Hormone and Chemotherapy.

Table 4.56 indicates, for each model, Cox, PLANN-ARD and NPI, the risk groups with the higher number of patients treated with each treatment.

	Cox	PLANN-ARD	NPI
None	Risk group 1	Risk group 1	Risk group 2
Hormone	Risk group 2	Risk group 1 and 2	Risk group 3
Chemo	Risk group 2	Risk group 2,3,4	Risk group 3
Combined	Risk group 2	Risk group 4	Risk group 3

Table 4.56 - Higher Treatments' Risk group Ratio for different models, for the validation data set.

Opposed to Christie, here there is not much coherency between the treatments and the different stratification methodologies, where the biggest one is for None treatment for Cox and PLANN.

These known distributions can help in treatments' decision making when obtaining a prognosis for a new patient, as it is acknowledged the ratio for each risk group.

Chapter 5 -

Online Breast Cancer decision support systems

Many studies noticed that many women with breast cancer are requesting more information about their disease and also suggest that they have an increasing desire to be involved in decisions about their care (Ravdin, Siminoff, Harvey,1998). However, in order to participate in decision making, the patient needs accurate information about the disease. A growing number of clinical tools have been developed in order to address the problems identified by clinical studies with both determining the risk of recurrence for individual patients and communicating this information in the clinical consultation. A number of these clinical tools are easy to use and some are accessible over the Internet. While studies have shown some of these instruments to improve patient knowledge and facilitate shared decision making, a number of basic questions still remain unresolved and need to be addressed with the aim to involve patients more on decision making regarding the breast cancer disease (Whelan, Loprinzi, 2005), (Fonseca, Mora, Barroso, 2006).

Several online breast cancer prognostic tools exist. However there are two that are widely known and used clinically, namely Adjuvant! (Olivotto, Bajdik, Ravdin, Speers, Coldman, Norris, Davis, Chia, Gelmon, 2005) and Numeracy (Loprinzi, Thome, 2001). These are computer-based programs designed to assist in adjuvant therapy decision making. Both programs determine a patient's baseline risk of recurrence and/or death at 10 years without adjuvant therapy, based on prognostic factors. Despite providing similar estimates of baseline

risk and absolute benefits, the instruments do differ (Whelan, Loprinzi, 2005).

Firstly in this chapter, it is explained a web decision making tool which can be accessible on the Internet, Adjuvant!. It has been chosen this particular interface because it appears to be readily accepted by practicing clinicians and it is used to validate the model, the same data set that it was used to validate the developed approaches in this thesis, the BCCA data set. Finally a Web decision support system is exhibited and its functionalities are explained. It is important to mention that this Web decision support system includes all the modelling and stratification methodologies improved and developed, presented on this thesis.

5.1 - Online breast cancer prognostic estimate – AdjuvantOnline

Adjuvant! is a computer program designed to produce prognostic estimates of outcome with and without therapy, based on estimates of individual patient prognosis. Adjuvant! Online was developed as a decision-making tool for health care professionals and patients with early cancer to discuss the risks and benefits of adjuvant therapy following surgery. The Web site states that its goal is to help health professionals estimate the risk of negative outcome (cancer-related mortality or relapse) without systemic adjuvant therapy, estimate the reduction of these risks afforded by therapy and estimate the risks of side effects of the therapy.

Version 2.2 (2001) (Ravdin, Sminoff, Davis, Mercer, Hewlett, Gerson, Parker, 2001) of this program includes the estimation of risk of breast cancer death at a 10-year follow-up on tumour size, the number of involved nodes and Oestrogen status. Patients included in the initial analysis were woman who had invasive, unilateral, noninflamatory disease, had undergone definitive surgery and had axillary staging with at least six nodes sampled. Furthermore, patients must not have known residual or metastatic disease. The parameters used for adjuvant therapy decision-making are entered in the online software. Age is used by the program to calculate the expected natural mortality and to produce the default estimate of menopausal status. Comorbidity is also a parameter that must be inserted and is an estimate of the general health of the individual for whom the estimates are being made. For Adjuvant! the number of positive nodes, together with tumor size are the main factors used to make estimates of patient prognosis.

The output shows the outcomes for survival in terms of Overall Survival at 10 years, estimates of remaining life expectancy and long term survival curves. It also shows outcomes

for DFS (disease-free survival) at 10 years. The existing bar-graphs show the percentage of patients died of breast cancer, the percentage died of non-breast cancer causes of death and an estimate of the increased percentage of patients alive at 10 years because of specific adjuvant therapy chosen. There are also bar graphs which show the percentage of patients who are alive without breast cancer at 10 years, what percentage are expected to relapse with breast cancer and what percentage died of other non-breast cancer causes of death. These estimates are shown for scenarios where adjuvant therapy is either used or not, allowing to view the additional percentage of patients who remain disease-free at 10 years because of adjuvant therapy. It is also possible to toggle between different adjuvant treatment options and examine the impact on DFS. A criticism of Adjuvant! is that it does not provide estimates with 95% confidence intervals.

The assumptions inherent in Adjuvant! and its applicability to woman beyond the range of ages used to develop the model were only independently validated in 2005 using a data base provided from BCCA (British Columbia Cancer Agency) (Olivotto, Bajdik, Ravdin, Speers, Coldman, Norris, Davis, Chia, Gelmon, 2005). 10-year predicted OS, BCSS and EFS values were determined from each patient using Adjuvant! version 5.0. Patient age, tumour size, number of positive nodes, grade, ER status and adjuvant systemic therapy used were entered in the model and 10-year OS, BCSS and EFS values were calculated. The default comorbidity assumption of "minor health problems" was used. The study demonstrated that the predicted outcomes of Adjuvant! were valid with the exception of a few specific subgroups of patients.

Other studies were carried out in order to validate this software, as it is widely used and consulted by clinicians and patients and has been shown to influence patient choices in the clinical setting (Peele, Siminoff, Xu, et al, 2005), (Ozanne, Braithwaite, Sepucha, Moore, Esserman, Belkora, 2009) tested the hypothesis that Adjuvant! predictions would be sensitive to comorbidity inputs. In contrast with the other inputs, the assessment of patient comorbidities is highly subjective questioning the reliability of Adjuvant! This variable inputs for Adjuvant! model includes: perfect health, minor problems, average for age, major problems (+10;+20;+30), where documentation offers little guidance regarding these definitions of comorbidity. (Ozanne, Braithwaite, Sepucha, Moore, Esserman, Belkora, 2009) concluded that comorbidity influences mortality predictions, specially for woman older than 60 years old it is the most influential input. In addition, the changes in the comorbidity outputs are significant enough that they are likely to affect patients' treatment choices.

It is however important to mention that, although the model is described quantitatively in (Ravdin, Sminoff, Davis, Mercer, Hewlett, Gerson, Parker, 2001), the true dynamics of the model remain unpublished. Therefore, all kind of validation is a first-order approximation, as there is no evidence of relationships between variables or interaction terms in the model.

5.2 - Proposed Breast cancer survival Web decision support system

A web clinical decision support system was developed in order to assist the clinicians to perform more accurate decisions about breast cancer treatments and prognostic outcome of survival, based on patients' characteristics. It is important to mention that the aim of this web system is to keep and expand current practices rather than replace them. The present decision support system makes an important contribution to both technical innovation and clinical application as several important novelties were added or changed to current practice. Previous developments have already presented a web decision support system as a relevant innovation (Jarman et al, 2008), (Lisboa, Etchells, Jarman, Ramsey, 2007). However, the proposed system improves upon these systems by resolving and improving some particular issues.

This web clinical decision support system incorporates three breast cancer prognostic methodologies previously mentioned, namely the Cox proportional hazards modelling, the PLANN-ARD and the NPI. The decision support system indicates, not only the prognostic index calculated for a single patient, but also its prognostic risk group, which is straightforward for NPI and obtained through a methodology explained previously, based on regression trees, when applied to the Cox proportional hazards and PLANN-ARD modelling. It is important to mention that missing data was incorporated in the prognostic models available in the web decision support system, which was overcome using multiple imputation techniques. This web system also has the advantage of saving patients' prognosis as well as their clinical data, providing a patient history based on prognosis and treatments. This history can be analysed and compared over time, which may help and improve clinicians' medical decisions.

Figure 5.1 – Home page of breast cancer decision support system

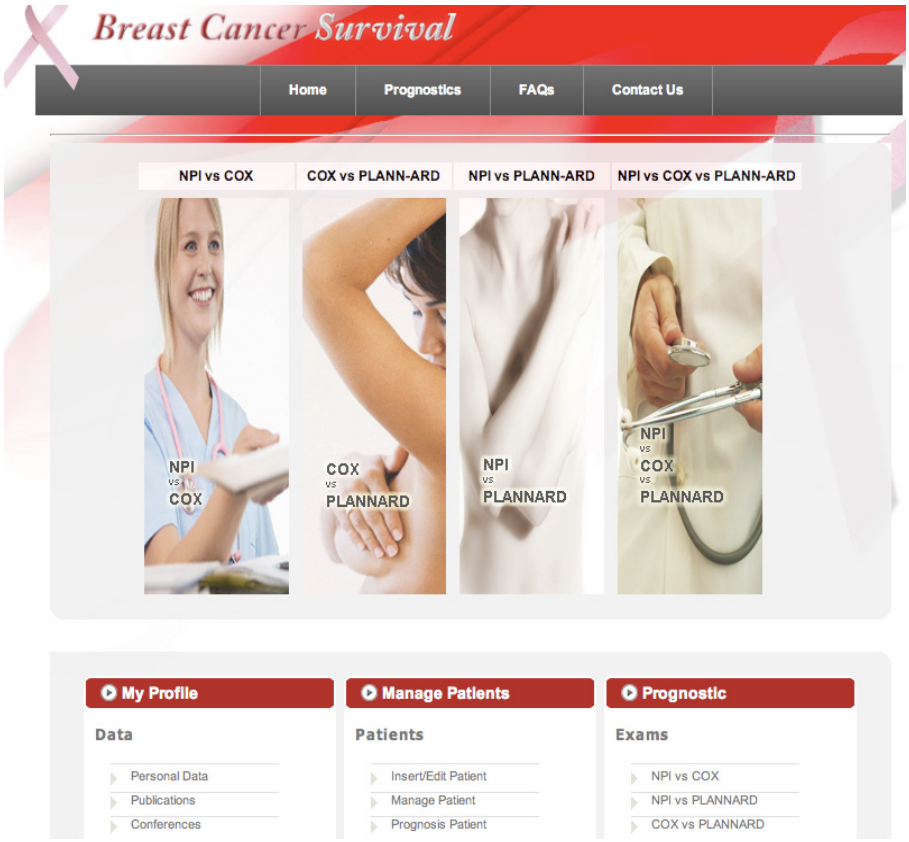

Figure 5.2 – *Home page of breast cancer decision support system after the introduction of a registration user*

The web site can be reached trough the link http://bcsurvival.pt.la/, where it is visualized the web page shown in Figure 5.1. In this web page it is possible to register a user and therefore access to the entire web site potential, first presented as in Figure 5.2 . The functionalities that are available for use are:

1. My Profile

   In this web-site section it is presented the user's personal information and his publications. All the information can be also updated.

2. Manage Patients

   This feature allows the user to insert or edit the information of a patient, manage all the patients who were previously created and visualise all predictions previously assessed by this web site, indicating the date that was saved.

3. Prognostic assessments

   In this web-site section the output of three prognostic models combined with the stratification methodology can be analysed together, that is NPI versus Cox, NPI versus PLANN-ARD and Cox versus PLANN-ARD. Cox and PLANN-ARD model estimate the risk of breast cancer death at a 5-year follow-up based on 6 predictive variables, namely *Age*, *Histological type*, *Nodes ratio*, *Oestrogen*, *Histological grade* and *Pathological size*. Patients included in this analysis are post-operative female cancer patients with primary invasive carcinoma of the breast at clinical stage T1-2 (tumours with maximum diameter of less than 5 cm), node stage N0-1 (no nodes affected in the axilla or mobile nodes) and metastasis stage M0 (no evidence of distant metastic spread of the tumour). The remaining patients were excluded from the study. The Cox proportional hazards and PLANN-ARD model were trained using a data set of 743 patients, collected at the Christie Hospital data set, from 1990-94.

   The different predictions models can be compared two by two. For this purpose the user must choose the ones he want to compare and analyse. After entering the patient variables', both the prognostic indices and the prognostic risk group are obtained. It is visualized a scatter-plot of the prognostic indices as well as the risk group of all patients used to train the models, for the two models chosen previously. In addition this section also presents the patient individual prognosis with 95% confidence intervals, derived by using the PLANN-ARD model with Monte Carlo methods.

Figure 5.3 presents an example of a web-page in which the NPI and PLANN-ARD prognostic models were chosen. After introducing the patient characteristics and pressing the *Prognosis* button, the prognostic index is obtained, the prognostic risk group, for each the selected models and the individual prognosis obtained for the patient.

Pressing the *View Treatments* button, the patient's information section is replaced by the percentages for each treatment that was received by the patients in the training data set, divided in hormone, chemotherapy, combined or none.

Pressing the *Switch View* button, a different view of the prognostic assessments can be observed. It can be examined survival for patients groups within each matrix cell using KM estimated survival curves, also indicating the mean survival estimates and the 95% confidence intervals, in order to discover heterogeneity in estimated survival for any of the models prognostic group. These differences in survival are an indication of the added value of cross-matching the different prognostic models. For each survival curve it is indicated the number of patients and deaths that were considered for the training data set. Figure 5.5 represents an example of this matrix, where it demonstrates how it is possible to draw misleading conclusions and shows the benefit of combining information.

The obtained prognostic as well as the patient information can be also saved for further analysis. This must be associated to a certain patient, previously inserted in the area "Manage patients", by selecting the sought patient followed by the button *Save values*.

By clicking above of each KM curve, it is possible to visualize it with more detail, as it is demonstrated on figure 5.6 .

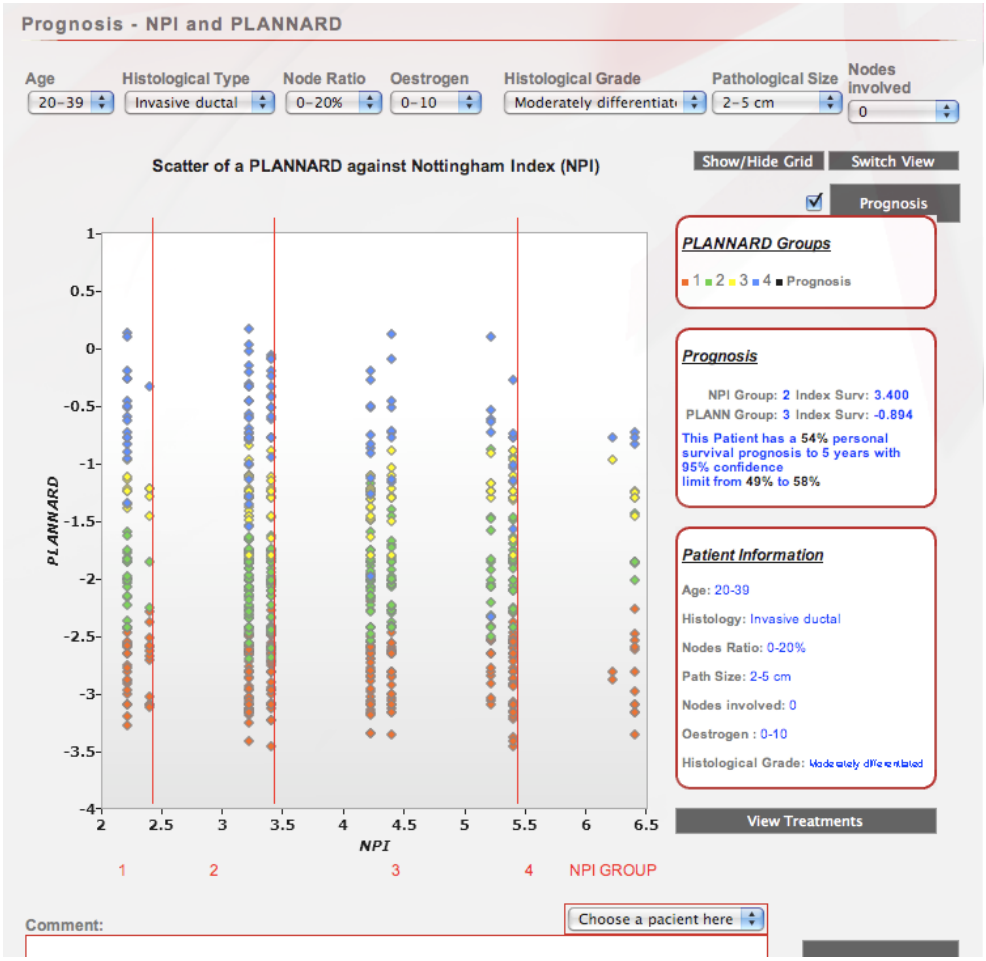

Figure 5.3 – Prognostic assessments for a particular patient.
Patient with age between 20 and 39 years, with histological type equal to invasive ductal, with Nodes ratio from 0 to 20%, with Oestrogen from 0 to 10, with Histological Grade moderately differentiated, with pathological size from 2 to 5 cm and with Nodes involved equal to 0. Here it is compared the NPI prognostic mode with PLANN-ARD prognostic model, followed by a stratification methodology.

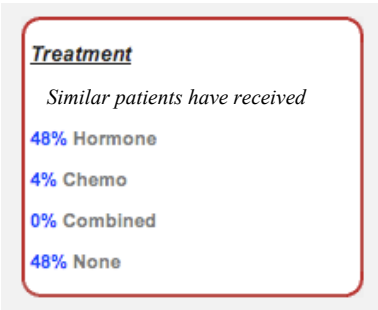

Figure 5.4 – Treatments information on the web-site.
This information substitutes the patient information in the web page, when the View Treatments button is pressed.

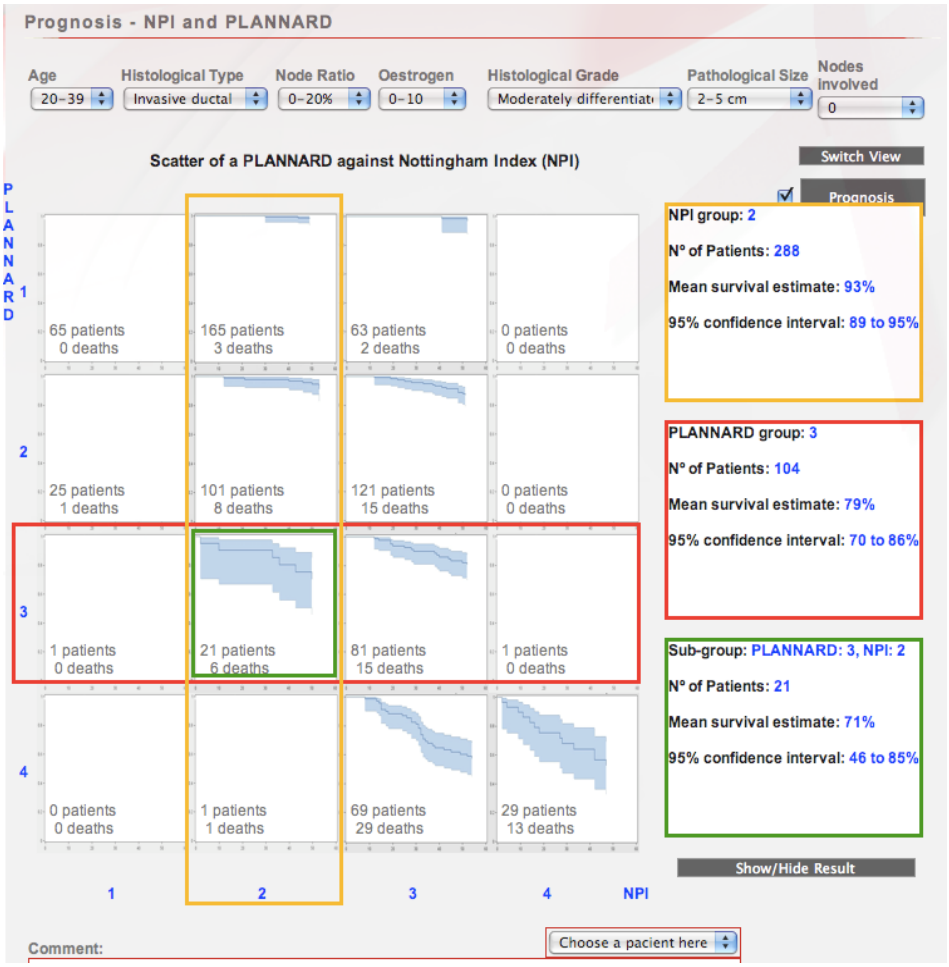

Figure 5.5 – Matrix with KM curves for a patient choosing NPI and PLANN-ARD. This comparison shows a significant difference in survival between the NPI (risk group 3) and PLANN-ARD modelling. Note that within NPI group 3, the top group experiences an incidence of death at 5 years of less than 5%, the next two groups together 15%, three times higher, and the last group over 40%, more than double the previous group. This raises the prospect of under- or over-treatment within this single clinical risk group.

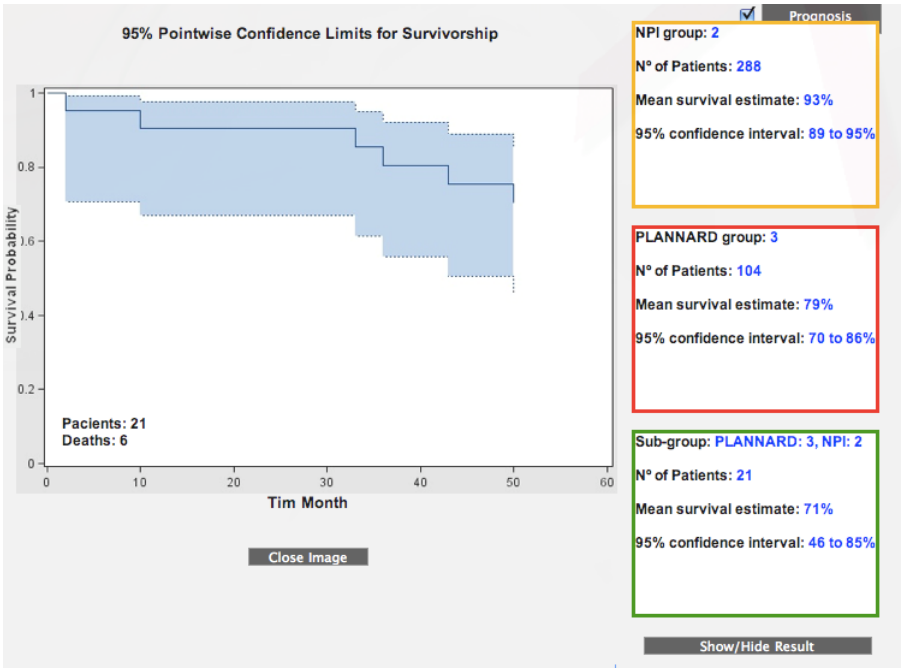

Figure 5.6 – KM curve after clicking in the cross-matching survival curves web-page.

Combining all the elements described above provides an integrated intelligent web support decision tool, achieved by the cross-matching matrix where each column represents patients in a prognostic risk group determined by a prognostic model and each row represents the risk group determined by another prognostic model. This can inform the user on a patient's survival outcome, allowing accessing the heterogeneity in survival within a prognostic risk group. The different perspectives given by this web decision tool can also indicate whether a particular patient is an outlier of the model when occupies an empty or sparsely populated cell or if the patient borders on other cells of the matrix.

Chapter 6 -

Conclusions and Future Work

This breast cancer prognostic modeling study followed a precise methodology and has identified six predictive variables, which are consistent with those used in clinical practice. The thesis proposes a methodology for incorporating missing data into generic non-linear modelling with the Partial Logistic Artificial Neural Network (PLANN) regularised within the evidence-based framework with Automatic Relevance Determination (ARD) by multiple imputation and model averaging over samples of the imputed distribution. This methodology is shown to be effective and enables predictions to be made on data sets with a different pattern of missing data, which is essential for external validation as well as to used data from training data with missing values to make inferences for future patients.

While the linear (Cox proportional hazards) and flexible models (PLANN-ARD) are comparable in their discrimination ability evaluated using the C^{td} index, there are also important differences between the two models. The PLANN-ARD is mainly proposed as a predictive model providing individual interval estimates of the hazard and survival for individual patients, overcoming the limitation of the proportionality of hazards. Furthermore, the dependence of the hazard prediction as a function of covariates and over time is estimated directly and can be visualised over time. Secondly, the calibration of the models in external validation shows a marginal advantage for the neural network. Concerning discrimination, the

results are not surprising since the intrinsic limitation in the available data was not expected to support a major improvement in the capability of ranking prognosis according to covariate values, especially given that they are available only after discretisation onto non=linearly spaced groups, which effectively turns the Cox regression model into a piecewise linear model. However, the calibration performances of PLANN-ARD over the standard Cox regression model suggest the relevance of non-additive and time dependent covariate effects for predicting on new patients.

It should be emphasised that when the data satisfy the assumptions of piecewise linear model such as Cox regression with discrete variables, then a well regularised non-linear model will behave similarly to a linear model. Residual interactions can result in marginal benefits for the non-linear model. Actually, the limited information available in the pre-categorised data from few predictor variables, as in the present case, should limit the final performance of any modelling tool.

Avoiding the binning of the individual covariates would help to remove unintended subjective effects currently recognised by the clinical community as damaging to the consistency in delivery of care for this important disease category. A relevant consideration here is the forced categorisation of histological tumour grade, which is subject to subjective effects in borderline cases between two grades, whose effect would be much reduced by reporting the underlying numerical score resulting from the histological observations.

The ability to avoid discretisation of continuous variables further enhances the advantage of flexible models over standard linear approaches.

A prognostic index was defined which may be used to stratify patients into risk groups with statistically significant grouped outcomes, utilizing the PLANN-ARD model. The proposed use of the Crude Cumulative Incidence rate as the basis of the prognostic index calculation means that this approach will extend from single to competing risks modelling.

Four stratification strategies were applied. Pure clustering of the population of covariates without reference to the prognostic indicator results in coherent patient groups but with relatively poor specificity for outcome, as measured by the separation between the group means of the overall mortality rates. Regressing the distribution of prognostic scores with rule-based trees (CART) succeeds in separating patient groups with statistically different mean survival and coherent membership within each group. Another approach stratifies the prognostic index directly, which can result in mixed populations within single risk groups,

which the application of an automatic rule extraction method (OSRE) identified and characterized. This showed the composition of the risk groups to be no more complex than for the groups identified by CART and the mean grouped survival show similar, statistically significant, separation. In contrast, informed clustering with the Fisher information matrix as a metric has the merit of permitting a specific definition of the patient population, from which to forward predict grouped survival, instead of inferring a threshold back from the log-rank separation index.

Therefore, while all methods generalized well to the validation data set, thresholding of the prognostic index using the regression tree methodology (CART) is the only methodology to offer both specificity to outcome and transparency of group composition. This method generalized in external validation as required by a staged methodology for the development of decision support systems in medicine and offers a possible route to a clinically useful patient stratification index that is expressed in the form of straightforward Boolean rules.

The work developed in this study has also enabled the combination of different breast cancer prognostic methodologies including those currently used in clinical practice, such as NPI, TNM and St. Gallen. This consensus approach helps in building a more robust allocation into different prognostic groups and consequently in the decision about the therapies to apply, by providing a triangulation of several plausible and validated prognostic indices.

Finally, the thesis describes a web decision support system for breast oncology which shows the value of the new prognostic models and stratification methodology to discriminate patients by risk of overall mortality. This tool starts with patient specific variables entered by the clinician and identifies the risk group allocation for the three prognostic models: NPI, Cox proportional hazards, PLANN-ARD, together with the Boolean rules that explain each risk group. A cross-matched matrix of grouped survival and the cell where the particular patient's parameters reside within the matrix is also presented, leading to better insights about the accuracy of the risk group allocation for each specific prognostic model.

Future work should compare the predictions obtained with Adjuvant! and the proposed methodologies, using the BCCA data set to validate the prognostic model, since it is the data utilized to validate Adjuvant!. However, prognostic predictions for 10 years of follow-up are required, which was not possible with the available data from the Christie Hospital since this has only 5 year follow-up. In addition, it was necessary to obtain the individual predictions from Adjuvant! for each patient in the validation data set, which has not been forthcoming

from the Adjuvant! group. Overcoming these two barriers, it would be extremely interesting and indeed important from a clinical point of view to compare, both historically for retrospective routinely acquired hospital data bases and for individual patients prospectively assessed, the prognostic inferences obtained with the range of widely used online breast cancer decision support system, as well as Adjuvantonline! and the predictions obtained from the proposed methodology proposed in this thesis.

References

(Altman, Lausan, Sauerbrei, Schumacher, 1994) Altman, Gouglas G., Lausan, Berthold, Sauerbrei, Willi and Schumacher, Martin, "Dangers of using "Optimal" cuptoints in the evaluation of prognostic factors", Journal of the National Cancer Institute, Vol. 86, No. 11, pp 829-835 (1994).

(Altman, Royston, 2000) Altman, Doug G. and Royston, Patrick, "What do you mean by validating a prognostic model?", Statistics in Medicine, Vol. 19, pp 453-473, (2000).

(Altman, Vergouwe, Royston, Moons, 2009) Altman, Gouglas, Vergouwe, Yvonne, Royston, Patrick, Moons, Karel, "Prognosis and prognostic research: validating a prognostic model", BMJ;338:b605, (2009).

(Amari, 1998) Amari, Shun-ichi"Natural gradient works efficiently in learning". Neural Computation, vol. 10, no. 2, pp. 251–276, (1998).

(Ambrogi, Biganzoli, Boracchi, 2008) Ambrogi, Frederico, Biganzoli, Elia, Boracchi, Patrizia, "Estimates of clinically useful measures in competing risks survival analysis", Statistics is Medicine, Vol. 27(30), pp. 6407-6425, (2008).

(Antolini, Boracchi, Biganzoli, 2005) Antolini, Laura, Boracchi, Patrizia and Biganzoli, Elia, "A time-dependent discrimination index for survival data", Statistics in Medicine, 24: pp 3927-3944 (2005).

(Bacciu, Starita, 2008) D. Bacciu and A. Starita, "Competitive repetition suppression (CoRe) clustering: A biologically inspired learning model with application to robust clustering". Neural Networks, IEEE Transactions, vol. 19, no. 11, pp. 1922–1941, (2008).

(Bacciu, Jarman, Etchells, Lisboa, 2009) Bacciu, Davide, Jarman, Ian H F., Etchells, T.A., Lisboa, P.J.G., "Patient stratification with competing risks by Multivariate Fisher distance", Proceedings of the 2009 international joint conference on Neural Networks, pp 3453-3460, (2009).

(Bakker, Heskes, 1999) Bakker, Bart and Heskes, Tom, "A neural-Bayesian approach to survival analysis", Artificial Neural Networks, Vol. 2, pp 832-837 (1999).

(Bar-Yam, 1997) Bar-Yam,Yaneer, "Dynamics of complex system", Addison-Wesley, Chapter 2 (1997).

(Biganzoli, Boracchi, Mariani, Marubini, 1998) Biganzoli, E., Boracchi, P., Mariani, L. and Marubini, E., "Feed forward neural networks for the analysis of censored survival data: A partial logistic regression approach", Statistics in Medicine, 17, 1169-1186 (1998).

(Biganzoli, Boracchi, Mariani, Marubini, 1998) Biganzoli, Elia, Boracchi, Patrizia,Mariani, L. and Marubini, Ettore, "Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach", Stat. Med, 17, pp 1169-1186 (1998).

(Biganzoli, Boracchi, Marubini, 2002) Biganzoli, Elia, Boracchi, Patrizia and Marubini, Ettore, "A general framework for neural network models on censored survival data", Neural Networks, 15, pp 209-218 (2002).

(Bishop, 2006) Bishop, C. M., "Pattern recognition and Machine learning", Springer (2006).

(Boracchi, Coradini, Antolini, Oriana, Dittadi, Gion, Daidone, Biganzoli, 2008) Boracchi, P., Coradini, D., Antolini, L., Oriana, S., Dittadi, R., Gion, M., Daidone, M.G., Biganzoli, E., "A prediction model for breast cancer recurrence after adjuvant hormone therapy", The international journal of biological markers, Vol. 23 no.4, pp 199-206, (2008).

(Bradburn M.J., Clark, Love, Altman, 2003) Bradburn, M.J., Clark, T.G., Love, S.B. and Altman, D.G., "Survival Analysis Part III: Multivariate data analysis – choosing a model and assessing its adequacy and fit". British Journal of Cancer, Vol. 89, pp 605-611, (2003).

(Bradburn, Clark, Love, Altman, 2003) Bradburn, M.J., Clark, T.G., Love, S.B. and Altman, D.G., "Survival Analysis Part IV: Further concepts and methods in survival analysis". British Journal of Cancer, Vol. 89, pp 781-786, (2003).

(Brand, 1998) Brand, J.P.L., "Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets", Academic thesis, Erasmus University, Rotterdam, (1998).

(Breiman, Friedman, Olsen, Stone, 1984) Breiman, L., Friedman, J.H., Olsen, A.R, Stone, C. J.,

"Classification and Regression Trees", The Wadsworth & Brooks, (1984).

(Burke, 1994) Burke, H.B., "Artificial neural networks for cancer research:outcome prediction", Seminars in surgical oncology; Vol 10 73-79 (1994).

(Clark, Stewart, Altman, Gabra, Smyth, 2001) Clark. T.G., Stewart, M.E., Altman, D.G., Gabra, H., Smyth, J.F., "A prognostic model for ovarian cancer", British Journal of Cancer, 85, 944–52 (2001).

(Clark, Altman, 2003) Clark, Taane G. and Altman, Douglas G., "Developing a prognostic model in the presence of missing data: an ovarian cancer case study". Journal of Clinical Epidemiology, Vol. 56, pp 28-37, (2003).

(Collet, 2003) Collet, David, "Modelling Survival data in medical research", Chapman & Hall/CRC, (2003).

(Collins, Altman, 2009) Collins, Gary, Altman, Gouglas, "An independent external validation and evaluation of QRISK cardiovascular risk prediction: a prospective open cohort study", BMJ, 339:b2584, (2009).

(Computation in the brain) Computation in the brain. Available at: http://www.willamette.edu/~gorr/classes/cs449/brain.html

(Concato, Peduzzi, Holford, Feinstein, 1995) Concato, John, Peduzzi, Peter, Holford, Theodore R. and Feinstein, Alvan R., "Importance of events per independent variable in proportional hazards analysis. I. Background, Goals and General Strategy", Journal of clinical epidemiology, Vol. 48, Nº12, pp 1495-1501, (1995).

(Concato John, Peduzzi, Holford, Feinstein, 1995) Concato, John, Peduzzi, Peter, Holford, Theodore R. and Feinstein, Alvan R., "Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and Precision of regression estimates", Journal of clinical epidemiology, Vol. 48, Nº12, pp 1503-1510, (1995).

(Concato, Peduzzi, Holford, Kemper, Feinstein, 1996) Concato, John, Peduzzi, Peter, Holford, Theodore R., Kemper, Elizabeth and Feinstein, Alvan R., "A Simulation Study of the number of

events per variable in Logistic Regression Analysis", Journal of clinical epidemiology, Vol. 49, Nº12, pp 1373-1379, (1996).

(Cox, 1972) Cox D. R., "Regression models and life tables". *Journal of the Royal Statistical Society*, B. Vol. 74, pp 187-220, (1972).

(D'Agostino, Nam, 2004) D'Agostino, R.B., Nam, B.H., "Evaluation of the performance of survival analysis models: discrimination and calibration measures". N. Balakrishnan, C. Rao (Eds.), *Handbook of Statistics*, 23rd ed., Elsevier, Amsterdam, 1-26, (2004).

(D'Eredita, Giardina, Martellotta, Natale, Ferrarese, 2001) D'Eredita, G., Giardina, C., Martellotta, M., Natale, T., Ferrarese, F., "Prognostic factors in breast cancer: the predictive value of the Nottingham Prognostic Index in patients with a long-term follow-up that were treated in a single institution", European Journal of Cancer, 37, pp 591-596, (2001).

(Delen, Walker, Kadam, 2005) Delen, Dursun, Walker, Glenn and Kadam, Amit, "Predicting breast cancer survivability: a comparison of three data mining methods", Artificial Intelligence in Medicine, 34, pp 113-127 (2005).

(Eleuteri, Aung, Taktak, Damato, Lisboa, 2007) Eleuteri, A., Aung, M.S.H., Taktak, A.F.G., Damato, B. and Lisboa, P.J.G., "Bayesian Neural Networks for survival analysis: A comparative study", Proceedings of CIMED 2007 ,(2007).

(Etchells, Lisboa, 2006) Etchells T.A. and Lisboa P.J.G., "Orthogonal Search-Based Rule Extraction (OSRE) for trained Neural Networks: A practical and efficient approach", IEEE Transactions on Neural Networks, Vol. 17, No. 2, (2006).

(Etchells, Fernandes, Jarman, Fonseca, Lisboa, 2008) Etchells, T.A., Fernandes, A.S., Jarman, I.H., Fonseca, J.M., Lisboa, P.J.G., "Stratification of severity of illness indices: a case study for breast cancer prognosis", *accepted KES2008 Zagreb, Croatia*, 3-5 September, (2008).

(Fernandes, Jarman, Etchells, Fonseca, Biganzoli, Bajdik, Lisboa, 2008) Fernandes A.S., Jarman I.H., Etchells T.A., Fonseca J.M., Biganzoli Elia, Bajdik Chris and Lisboa P.J.G., "Missing data imputation in longitudinal cohort studies – application of PLANN-ARD in breast cancer survival", Machine Learning and Applications, ICMLA08, pp 644-649, (2008).

(Fonseca, Mora, Barroso, 2006) Fonseca, José Manuel, Mora, André, Barroso, Pedro, "The web and the new generation of medical information systems", Outcome Prediction in Cancer, Section 5 – Chapter 14, Elsevier, Studies in Multidisciplinarity series, (2006).

(Galea, Blamey, Elston, Ellis, 1992) Galea, M.H., Blamey, R.W., Elston, C.E., Ellis, I.O., "The Nottingham Prognostic Index in primary breast cancer". Breast Cancer Res Treat, 22:207–19, (1992).

(Greenland, Finkle, 1995) Greenland, S., Finkle, W.D., "A critical look at methods for handling missing covariates in epidemiologic regression analysis", Am J Epidemiology, 142(12): 1255-64, (1995).

(Guerra, Algorta, Diaz de Otazu, Pelayo, Farina, 2003) Guerra, I., Algorta, J., Diaz de Otazu, R., Pelayo, A., Farina, J., "Immunohistochemical prognostic index for breast cancer in young women", Journal Clinical Pathology: Molecular Pathology 56, 323–327 (2003).

(Harbeck, Jakesz, 2007) Harbeck, Nádia, Jakesz, Raimund, "St Gallen 2007:Breast Cancer Treatment Consensus Report", Breast Care, 2:130-134, (2007).

(Harrel, Lee, Califf, Pryor, Rosati, 1984) Harrel, F.E., Lee, K.L., Califf, R.M., Pryor, D.B., Rosati, R.A., "Regression modelling strategies for improved prognostic prediction, Statistics in Medicine", Volume 13, 1501-1511, (1984).

(Harrell Jr., Lee, Mark, 1996) Harrell Jr., Frank E., Lee, Kerry L. and Mark, Daniel B., "Tutorial in Biostatistics: Multivariable Prognostic Models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors", Statistics in Medicine, Vol. 15, pp 361-387, (1996).

(Harrell, 2001) Harrell, F.E., "Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression and Survival Analysis"; Springer-Verlag; (2001).

(Haybittle, Blamey, Elston, Johnson, Doyle, Campbell, Nicholson, Griffiths, 1982) Haybittle, J.L., Blamey, R.W., Elston, C.W., Johnson, J., Doyle, P.J., Campbell, F.C., Nicholson, R.I. and Griffiths, K., "A prognostic index in primary breast cancer", British Journal of Cancer, 45, 3621 (1982).

(Heller, Venkatraman, 1996) Heller, Glenn and Venkatraman, E.S., "Resampling procedures to compare two survival distributions in the presence of right censored data", Biometrics, Vol.52, nº 4, pp 1204-1213, (1996).

(Hippisley-Cox, Coupland, Vinogradova, Robson, May, Bringle, 2007) Hippisley-Cox, Julia, Coupland, Carol, Vinogradova, Yana, Robson, John, May, Margaret and Bringle, Peter, "Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study", BMJ, 335:136, (2007).

(Hippisley-Cox, Coupland, Vinogradova, Robson, Bringle, 2008) Hippisley-Cox, Julia, Coupland, Carol, Vinogradova, Yana, Robson, John and Bringle, Peter, "Performance of the QRISK cardiovascular risk prediction algorithm in an independent UK sample of patients from general practice :a validation study", Heart, 94, 34-39, (2008).

(Jarman et al, 2008) Jarman, I.H., et al "An integrated framework for risk profiling of breast cancer patients following surgery", Artificial Intelligence in Medicine, Vol. 42, Issue 3 pp 165-188, (2008).

(Jordan, 1995) Jordan, Michael I., "Why the logistic function? A tutorial discussion on probabilities and neural networks; Massachusetts Institute of technology", Computational Cognitive Science, Technical report 9503 (1995).

(Kappen, Neijt, 1993) Kappen, H.J., Neijt, J.P., "Neural Network analysis to predict treatment outcome", Annals of oncology; Vol.4 Supplement S31-S34 (1993).

(Kaski, Sinkkonen, Peltonen, 2001) Kaski, S., Sinkkonen, J. and Peltonen, J. "Bankruptcy analysis with self-organizing maps in learning metrics". *Neural Networks, IEEE Transactions*, vol. 12, no. 4, pp. 936–947, (2001).

(Lisboa, 2002) Lisboa, P.J.G., "A review of evidence of health benefit from artificial neural networks in medical intervention", Neural Networks, Invited Paper, Vol. 15, issue 1, pp 9-37, (2002).

(Lisboa, Wong, Harris, Swindell, 2003) Lisboa, P.J.G., Wong, H., Harris, P. and Swindell, R., "A Bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer". Artificial Intelligence in Medicine, Vol. 28, issue 1, pp 1-25, (2003).

(Lisboa, Etchells, Jarman, Ramsey, 2007) Lisboa, P.J.G., Etchells, T.A., Jarman, Ian H F. and Ramsey, Philip; 'A prototype Integrated Decision Support System for Breast Cancer Oncology', Lecture Notes in Computer Science, 4507, pp. 996–1003, (2007).

(Lisboa, Etchells, Jarman, Aung, Chabaud, Bachelot, Perol, Gargi, Bourdès, Bonnevay, Négrier, 2008) Lisboa, P.J.G., Etchells, T.A., Jarman, Ian H,, Aung, M.S., Chabaud, Sylvie, Bachelot, Thomas, Perol, David, Gargi, T., Bourdès, V., Bonnevay, S., Négrier, Sylvie, "Time to event analysis with artificial neural networks: An integrated analytical and rule-based study for breast cancer", Neural networks, vol 21(2-3), pp. 414-426, (2008).

(Loprinzi, Thome, 2001) Loprinzi, C, Thome, S.D., "Understanding the utility of adjuvant systemic therapy for primary breast cancer", Journal of Clinical Oncology, 19, 972-979, (2001).

(Mackay, 1992) Mackay, D.J.C., "The evidence framework applied to classification networks", Neural computation; 4(5) 720-736(1992).

(Mackay, 1995) MacKay, D.J.C., "Probable networks and plausible predictions – a review of practical Bayesian methods for supervised neural networks. Network", Computation in Neural Systems.,6: 469-505 (1995).

(Marubini, Valsecchi, 1995) Marubini, E., Valsecchi, M., "Analysing survival data from clinical trials and observational studies", John Wiley and Sons, (1995).

(McGuire, Tandon, Allred, Chamness, Ravdin, Clark, 1992) McGuire, W.L., Tandon, A.K., Allred, D.C., Chamness, G.C., Ravdin, P.M., Clark, G.M., 2Treatment decisions in axillary node-negative breast cancer patients", Monograohs-National Cancer Institute; Vol.11 173-180 (1992).

(Neal, 2001) Neal, Radford M., "Survival analysis using a Bayesian Neural Network", Joint Statistical Meetings report, Atlanta (2001).

(Newgard, Haukoos, 2007) Newgard, Craig, Haukoos, Jason., "Advanced statistics: Missing Data in Clinical Research – Part 2: Multiple Imputation ", Society for Academic Emergency Medicine, 669-678, (2007).

(Ohno-Machado, Walker, Musen, 1995) Ohno-Machado, L., Walker, M.G. and Musen, M.A., "Hierarchical Neural Networks for Survival Analysis", Proceeding of the eight congress on Medical InformaGcs, 8: pp 828-832, (1995).

(Olivotto, Bajdik, Ravdin, Speers, Coldman, Norris, Davis, Chia, Gelmon, 2005) Olivotto, I.A., Bajdik, C.D., Ravdin, P.M., Speers, C., Coldman, A., Norris, B., Davis, G.J., Chia, S. and Gelmon K., "Population-based validation of the prognostic model ADJUVANT! for early breast cancer". *Journal of Clinical Oncology*, Vol. 23 (12), pp 2716-25, (2005).

(Ozanne, Braithwaite, Sepucha, Moore, Esserman, Belkora, 2009) Ozanne, Elissa, Braithwaite, Dejana, Sepucha, Karen, Moore, Dan, Esserman, Laura, Belkora, Jeffrey, "Sensitivity to Input Variability of the Adjuvant!Online Breast Cancer Prognostic Model", Journal of clinical oncology, vol. 27, no. 2, pp. 214–219, (2009).

(Peele, Siminoff, Xu, et al, 2005) Peele, P., Siminoff, L., Xu, X, et al.,"Decreased use of adjuvant breast cancer therapy in a randomized controlled trial of a decision aid with individualized risk information", Medical Decision Making, vol. 25, pp. 301–307, (2005).

(Pop, Hayward, Diederich, 2009) Pop, E., Hayward, R., Diederich, J., "RULENEG: extracting rules from a trained ANN by stepwise negation". QUT NRC technical report. Queensland, Austrália: Queensland University of Technology, Neurocomputing Research Centre, (2009).

(Ravdin, Clark, 1992) Ravdin, P.M., Clark, G.M., "A practical application of neural network analysis for predicting outcome of individual breast cancer patients; Breast cancer research and Treatment", Vol. 22, 285-293 (1992).

(Ravdin, Clark, Hilsenbeck, Owens, Vendely, Pandian, McGuire, 1992) Ravdin, P.M., Clark, G.M., Hilsenbeck, S.G., Owens, M.A., Vendely, P., Pandian, M.R., McGuire, W.L., "A demonstration that breast cancer recurrence can be predicted by neural network analysis", Breast cancer research and Treatment; Vol. 21 47-53 (1992).

(Ravdin, Siminoff, Harvey,1998) Ravdin, P.M., Siminoff, I.A., Harvey, J.A., "Survey of breast cancer patients concerning their knowledge and expectations of adjuvant therapy". Journal of Clinical Oncology, 16:515-521, (1998).

(Ravdin, Sminoff, Davis, Mercer, Hewlett, Gerson, Parker, 2001) Ravdin, Peter, Sminoff, Laura, Davis, Greg, Mercer, Mary, Hewlett, Joan, Gerson, Nancy and Parker, Helen., "Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer", Journal of clinical oncology, vol 19, No 4 (February 15), pp 980-991, (2001).

(Ripley, Ripley 1998) Ripley, B.D. and Ripley, R.M., "Neural Networks as Statistical Methods in Survival Analysis", In R. Dybowsky, & V. Gant (Eds.), Artificial neural networks: Prospects for medicine. Landes Biosciences (1998).

(Rubin, 1987) Rubin, D.B., "Multiple imputation for nonresponse in surveys", New York: John Wiley & Sons, Inc, (1987).

(SAS software) SAS software. Available at: http://www.sas.com/

(Satagopan, Ben-Porat, Berwick, Robson, Kutler, Auerbach, 2004) Satagopan, J.M., Ben-Porat, L., Berwick, M., Robson, M., Kutler, D. and Auerbach, A.D., "A note on competing risks in survival analysis". British Journal of Cancer, Vol. 91, pp 1229-1235, (2004).

(Schafer, 1999) Schafer, J.L., "Multiple imputation: a primer", Statistical Methods in Medical Research, Vol. 8, pp 3-15, (1999).

(Schumacher, Hollander, Sauerbrei, 1997) Schumacher, Martin, Hollander, Norbert and Sauerbrei, Willi, "Resampling and Cross validation techniques: a tool to reduce bias caused by model building", Statistics in Medicine, Vol. 16, pp 2813-2827 (1997).

(Schwarzer, Vach, Schumacher, 2000) Schwarzer, Guido, Vach, Werner and Schumacher, Martin, "On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology", Statistics in Medicine, Vol. 19, pp 541-561, (2000).

(Sebastian, Gonzalez, Paricio, Perez, Flores, Madrona, Romero, Tebar, 2000) Sebastian S. Ortiz, Gonzalez, J.M. Rodrıguez, Paricio, P. Parilla, Perez, J. Sola, Flores, D. Perez, Madrona, A. Piñero, Romero, P. Ramirez, Tebar, F.J., "Papillary Thyroid Carcinoma: Prognostic Index for Survival Including the Histological Variety", Archives of Surgery, 135 (2000).

(Shafer, 1997) Shafer, J. L., "Analysis of incomplete multivariate data", Chapman & Hall/CRC, (1997).

(Silverstein, Buchanan, 2003) Silverstein, M.J., Buchanan. C., "Ductal carcinoma in situ: Ductal carcinoma in situ: USC/Van Nuys Prognostic Index and the impact of margin status", The Breast, Volume 12, Issue 6, 8th International Conference on Primary Therapy of Early Breast Cancer, St Gallen, Switzerland, 457-471, (2003).

(Steyerberg, 2009) Steyerberg, Ewout, "Clinical Predictions Models", Springer, (2009).

(Taktak, Antolini, Aung, Boracchi, Campbell, Damato, Ifeachor, Lama, Lisboa, Setzkorn, Stalbovskaya, Biganzoli, 2007) Taktak, A., Antolini, L., Aung, M.H., Boracchi, P., Campbell, I., Damato, B., Ifeachor, E.C, Lama, N., Lisboa, P.J.G, Setzkorn, C., Stalbovskaya, V. and Biganzoli, E.M. 'Double-blind evaluation and benchmarking of survival models in a multi-centre study' *Computers in Biology and Medicine,* 37 (8): 1108-1120 (2007).

(Van Buuren, Boshuizen, Knook, 1999) Van Buuren S., Boshuizen H.C., Knook D.L.,"Multiple imputation of missing blood pressure covariates in survival analysis", *Statistics in Medicine***,** Vol 18 (6), pp 681-94, (1999).

(Whelan, Loprinzi, 2005) Whelan, Timothy, Loprinzi, Charles. "Physician/Patient Decision Aids for Adjuvant Therapy", Journal of Clinical Oncology, Vol. 23,pp. 1627-1630, (2005).

(Williams, Mandrekar, Mandrekar, Cha, Furth, 2006) Williams, B.A., Mandrekar, J.N., Mandrekar, S.J., Cha, S.S. and Furth, A.F., "Finding Optimal Cutpoints for Continuous Covariates with Binary and Time-to-Event Outcomes", Technical Report Series #79, Mayo Clinic, Rochester, Minnesota, June 2006.