Assessing flexible models and rule extraction from censored survival data

Paulo J.G. Lisboa,* Elia M. Biganzoli**, Azzam F. Taktak***, Terence A. Etchells*, Ian H. Jarman*, M.S. Hane Aung*, Federico Ambrogi**

Abstract— **Qualitative model operation description is useful for its direct validation using expert domain knowledge. A framework for this purpose uses low-order Boolean rules to approximate the response surfaces generated by analytical inference models. In the case of censored data, this approach serves to characterise the allocation of patients into risk groups generated by a risk staging index. Furthermore, the low-order rules define low-dimensional sub-spaces where individual data points can be directly visualised by reference to decision boundaries for risk group allocation. The well-known ROC framework has recently been extended to a threshold independent, time-dependent performance index to quantify the predictive accuracy of censored data models, termed the C^{td} index. Taken together, the quantitative performance index, Boolean explanatory rules and direct visualisation of the data, define a consistent and transparent validation framework based on triangulation of information. This information can be included in decision support systems.**

I. Introduction

ASSESSING the generality of flexible models, such as artificial neural networks, is complex and multifaceted task to ensure generality of the results obtained. When these models are integrated into decision support systems, especially in safety-critical domains, the assessment process must comply with well-known requirements for evaluation as part of the lifecycle of software development, which includes the stages of verification and validation [1].

Traditionally generic non-linear models are assessed by evaluating the generality of the estimated performance, using for instance the Receiver Operating Characteristic (ROC) framework. This framework serves to assure the validity of the performance claims made but it does not verify the extent to which the operation of the model, whether used for knowledge discovery or to make predictive inferences, is consistent with domain expertise. There several procedures that can be used for this purpose, from sensitivity analysis through group profiling to rule extraction and visualization of the data in low-dimensions to identify, for instance, the presence of outliers which can result in apparently accurate model predictions that are, nevertheless, unreliable. These issues, and how they integrated into accepted frameworks for the development of complex clinical interventions, have been reviewed elsewhere [2-3][Elia, your paper from EWADP in Pisa may be relevant here]

This paper provides an overview of two aspects of the assessment process where recent developments have been reported, namely:

- How to extend the ROC framework to evaluate the performance of time-to-event models in the presence of censored data, with reference to the model's ability to correctly fit outcome data for individual cases
- A framework to represent the operation of neural networks as low-order Boolean rules that may be tested for consistency with domain knowledge and use to visualize decision boundaries.

It is proposed that statistical performance estimation, rule generation and direct visualization of the data, form an integrated framework to triangulate relevant and complementary aspects of verifiability and validity, which are relevant to quality assurance when assessing flexible models.

In the interests of space, other aspects of the quality assurance will not be discussed in detail. However, it is worth noting that the performance evaluation of generic non-linear models also needs to take account of robustness in model design, for instance through appropriate regularization. When inferences are drawn for individual cases, which is important for the development of personalized health care systems, then there is a further need to quantify the predictive uncertainty, that is to say to specify confidence intervals for the predictions made and to evaluate their accuracy, also to accurately model the data density in order to accurately identify outliers, which do not always automatically trigger large predictive errors.

The remainder of the paper is focused on the overview of a time-dependent AUROC index for censored data, and efficient generation of explanatory rules in the context of risk stratification for prognosis. The overview will be illustrated with results from a benchmark study of prognostic modeling for patients with uveal melanoma together with further analysis for rule identification.

Manuscript received January, 31 2007. This work was supported in part by the Biopattern Network of Excellence FP6/2002/IST/1, No.508803.

PJG Lisboa, TA Etchells, IH Jarman and MSH Aung are with the School of Computing and Mathematical Sciences, Liverpool John Moores University, UK (e-mail: p.j.lisboa@ljmu.ac.uk).

EM Biganzoli and F Ambrogi are with the Unit of Medical Statistics and Biometry, Istituto Nazionale Tumori, Milan, Italy (e-mail: elia.biganzoli@istitutotumori.mi.it).

AF Taktak is with the Department of Clinical Engineering, Royal Liverpool University Hospital, UK (e-mail: afgt@liverpool.ac.uk).

II. TIME DEPENDENT AUROC - THE C^{TD} INDEX

The ROC is among the most powerful and widely used frameworks to quantify generalized performance for classification tasks. Confidence intervals have been generated both in respect of the AUROC and for individual points on the curve, to show the area of uncertainty around the performance of classifiers with specific thresholds for class allocation [4].

However, the standard AUROC index does not apply for time-to-event modeling with censored data [5]. To derive models for outcome prediction, a crucial aspect is the availability of appropriate measures of predictive accuracy for a general class of models. The Harrell's C discrimination index is an extension of the area under the ROC curve to the case of censored survival data, which owns a straightforward interpretability. 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 times, which is not generally addressed in most clinical applications. The time-dependent discrimination index C^{td} [5] exploits the whole predicted survival function as outcome prediction, to summarise over time the discrimination power among subjects having different outcome. C^{td} is based on a novel definition of concordance: a subject who developed the event should have a less predicted probability of surviving beyond his/her survival time than any subject who survived longer. The predicted survival function of a subject who developed the event is compared to: (1) that of subjects who developed the event before his/her survival time, and (2) that of subjects who developed the event, or were censored, after his/her survival time. Subjects who were censored are involved in comparisons with subjects who developed the event before their observed times. A confidence interval for C^{td} is derived using the jackknife method on correlated one-sample U-statistics.

III. UVEAL MELANOMA BENCHMARK

Uveal melanoma is a cancer of the eyeball. Approximately 50% of all patients with uveal melanoma ultimately die of metastatic disease, which usually involves the liver. The probability of metastatic death after treatment of uveal melanoma is increased with a range of generally known risk factors, including large basal tumour dimension, ciliary body movement, epithelioid cellularity, closed connective tissue loops, increased microvascular density and chromosome 3 deletions [6]. At present, only tumour diameter and extension are known in most patients, because few undergo local resection or enucleation and because tumour biopsy is not routinely performed before radiotherapy or phototherapy. Whereas Tumour, Node, Metastasis (TNM) Classification and other categorizations of ciliary body and choroidal melanomas have been developed to group patients according to their prognosis for survival at the time of their initial treatment there is recognized need for more specific ranking of survivorship with less variation within each risk group.

The outcome of interest for the benchmark study reported in [6] all-cause mortality. Follow-up times were measured from the date of primary ocular treatment, either to the date of death or to the date of the close of the study, which was the 8th of March 2005. All the subjects entered in this study were registered with the UK National Cancer Registry who informed us automatically when any subject died. Thus, we were confident that any subject who has not been flagged as dead, was still alive at the end of the study.

Patients were selected from the database of the Liverpool Ocular Oncology Center for the time period 1984 – 2004 if: (1) diagnosed with uveal melanoma, clinically or histopathologically; (2) primarily treated at the Tennent Institute of Ophthalmology, Glasgow, before January 1993 or at the Royal Liverpool University Hospital after that date; and (3) resident in the United Kingdom (UK). Patients were excluded because of: (1) bilateral melanoma; or (2) missing data regarding basal tumor dimension or anterior tumor extension.

The dataset was randomly split into training and test sets and the sets were stratified to include roughly equal proportion of events, with data from 1734 patients (490 events) and 1146 patients (305 events) respectively. The median follow-up time was 5.31 years (range: 0 – 35.66).The tenets of the Helsinki Declaration were followed and institutional ethical committee approval for a multi-center outcomes analysis was obtained.

A double-blind evaluation of the accuracy in out-of-sample prediction of overall mortality was carried out to compare the predictive performance of generic non-linear models for censored data, [7], a Partial Logist Radial Basis Function Network (PLRBF) fitted with a multi-objective evolutionary algorithm, and a Partial Logistic Neural Network, with the architecture of a Multi-Layer Perceptron and regularisation within the Bayesian framework with a normal approximation to the evidence, using Automatic Relevance Determination (PLANN-ARD) [8]. The performance of the flexible models was benchmarked against a Partial Logistic Spline Model (PLSPL), which is a generalized linear model of the discrete hazard based on partial logistic regression, as well as the more commonly used log-normal and Cox regression models.

Model selection was carried out separately for each model, resulting in the covariate subsets listed in Table 1. ANTMAR is a categorical variable representing the anterior margin of the tumour. ANTORA was derived from ANTMAR by applying a threshold to categorise tumours into pre-ora and post-ora. Note that the COX and LOGN methods applied their own categorization technique into ANTMAR to reduce its dimensionality instead of using ANTORA. The tumour dimensions measured from ultrasound images provided variables LUBD and UH representing the largest basal diameter (mm) and height respectively. EPI is a binary variable representing the presence or absence of epithelioid cells in the tumour tissue from histopathological slides.

TABLE 1 VARIABLES SELECTED BY EACH MODEL.

Model	Age	Sex	ANT MAR	ANT ORA	LU BD	UH	EPI
COX	✓	✓	✓[1]		✓		✓
LOGN	✓		✓[1]		✓	✓	✓
PLSPL	✓			✓	✓		✓
PLAN	✓[2]	✓		✓	✓[2]	✓[2]	✓
NARD							
RBF	✓	✓		✓	✓[3]	✓	

[1]Categorised
[2]Normalised
[3]Rounded

TABLE 2 C^{TD} INDEX ESTIMATES τ =3, 5 AND 10 YEARS. TYPE "SPECIFIC" CONSIDERS THE DIFFERENT MODEL SPECIFIC TEST DATASET (COX AND LOGN 1146 PTS, PLANNARD AND RBF 1069, PLSPL 505) WHILST TYPE "COMMON" DENOTES THE SUBSET FOR WHICH ALL OF THE REQUIRED VARIABLES ARE PRESENT IN THE TEST DATASET (498 PATIENTS).

	Values of C^{td}				
dataset	COX	LOGN	PLSPL	PLAN NARD	RBF
specific	0.687	0.708	0.681	0.716	0.683
common	0.700	0.716	0.684	0.716	0.655
specific	0.706	0.722	0.668	0.732	0.701
common	0.705	0.708	0.672	0.710	0.659
specific	0.714	0.737	0.699	0.747	0.718
common	0.726	0.737	0.701	0.738	0.687

The C^{td} provided a discriminating index which ranked models by order of predictive performance in a way that was consistent with additional accuracy measures also reported in this study. The latter included graphical assessment of the survival curves generated by each model against the Kaplan-Meier curves in different prognostic groups of patients from excellent to poor prognosis. Numerical calibration of each model was also carried out using a generalization method of the Hosmer-Lemeshow analysis comparing predicted against observed survival in groups of patients at predetermined time intervals.

IV. RULE GENERATION FOR SURVIVAL DATA

A. Definition of a prognostic index

A further study was carried out to stratify patients into risk groups for mortality. The data comprised of 1734 cases with a 20 year follow up with all events being censored thereafter. The unit time interval size quieried to the PLANN-ARD is a tenth of a year. The explanatory variables used are listed in Table 1. Following the methodology introduced in [8], a prognostic index was defined from the logit of the hazard estimate for each individual case, averaged over the follow-up period. This score corresponds to the argument, βx, of logistic regression and Cox regression functions. The log-rank test was applied to the risk score, identifying three groups with are significantly different values of the prognostic index, which was verified by cross-validation resulting in the grouped Kaplan Meier curves shown in fig 1:

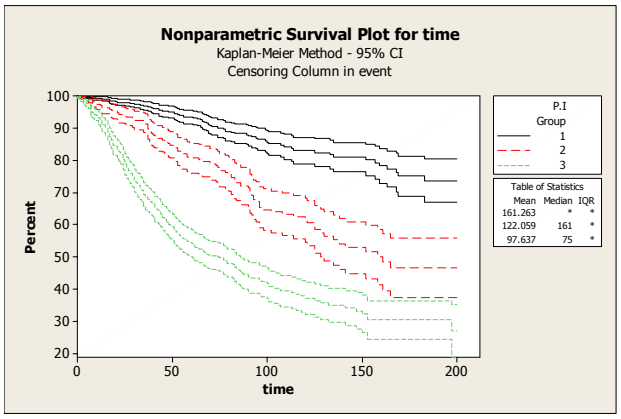

Figure 1 Kaplan-Meier survival characteristics for three groups of patients defined by significantly different values of prognostic index. The unit time interval is a tenth of year showing a follow up time of 20 years. Group 1 has 720 cases, group 2 has 345 and group 3 has 669 cases. The upper and lower lines for each group show 95% confidence intervals.

B. Rule extraction with OSRE

The clinical interpretation of risk groups relies on a mathematical characterization of their content. While the profile of average or median values of the covariates carry information about the overall pattern of data within each group, it comprises a series of univariate measurements which do not generally provide a method to verify consistency with domain expertise, other than in very broad terms. Consequently, it is of interest to consider strictly multivariate descriptions of the composition of risk groups, preferably translated into the language of the medical domain, often represented as logical rules in terms of set membership defined by thresholding the covariates.

Orthogonal Search Rule Extraction (OSRE) [9] is a computationally efficient algorithm to search for hypercubes in data space, since they map directly onto Boolean rules. This is achieved by assigning a separate indicator label for each risk group and treating risk group allocation as a classification task, to which a regularised Multi-Layer Perceptron can be fitted to generate a suitable response surface.

Once the noise in the risk assignment data is smoothed out by the response surface, the second stage is a recursive search process starting at each data point and traversing outwards from that point to the extreme value of each individual covariate, one at a time, keeping all other covariates fixed. A list is kept of the directions emanating from the data point along which the response surface crosses the response threshold, set for this study at a value of 0.5.

This methodology initially returns a rule for each data point, which requires a pruning process to keep only those rules

which represent large proportions of the data in each risk group, i.e. show high sensitivity for group membership, and do so with minimal mixing between groups, i.e. also have high specificity. The result is a set of multivariate rules involving typically few covariates, that is to say, low-order rules containing the covariates that characterize sub-group contained in each risk group.

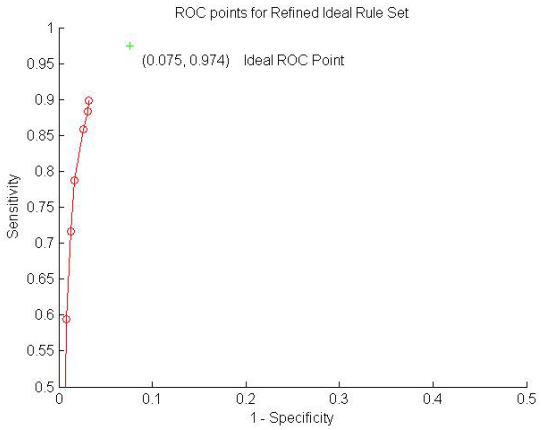

Figure 2 shows the effect of adding individual rules to the disjunction list. The overall sensitivity and specificity point of the disjoint of these rules steps closer to the cross which is the target ROC point. The rules in Table 3 can be seen as the lower three circles. It can be seen that addition of further rules provides less significant gain.

TABLE 3 CONTAINS THREE WELL PERFORMING RULES TARGETING MEMBERSHIP OF PROGNOSITIC GROUP 3 AS IN CLASS. THE FIRST TWO COLUMNS SHOW THE NUMBER OF CASES THAT ARE TRUE FOR THIS RULE FROM A TOTAL OF 1602 CASES (CASES WITH 'LUBD' AND 'UH' MISSING ARE NOT QUERIED INTO OSRE). THE NUMBER OF CASES IN GROUP 3 ARE 626. THE OTHER COLUMNS SHOW INDIVIDUAL SENSITIVITY, SPECIFICITY AND POSITIVE PREDICTIVE VALUE FOR EACH RULE.

Conjuntive Rule Statement	Cases In class	Cases Out class	Sens	Spec	PPV
Rule 1: 10.315 <= lubd <= 27.370 epi = 1	372	8	0.59	0.99	0.98
Rule 2: 12.935 <= lubd <= 27.635 6.500 <= uh <= 20.000 54.375 <= age <= 103.425 antora = 1	183	7	0.29	0.99	0.96
Rule 3: 7.040 <= lubd <= 27.885 2.000 <= uh <= 20.000 epi = 1 62.475 <= age <= 103.350	307	6	0.49	0.99	0.98

Note that OSRE contrasts with widely used rule induction methods in two ways: there are no univariate cut-offs for groups of data, as in OSRE a sequential univariate search is carried out at the level of each individual data point and returns a multivariate hyperbox around that point, without the need to partition the data along a sequence of univariate covariates; and secondly, the price paid for this methodology is that the rules are overlapping, rather than constrained to mutual exclusivity as is the case in rule tree induction. Mutually exclusive trees can be readily derived from overlapping rule sets by sequential conjunctions of each rule and the complement of the previous rule along the tree branch, but this loses the benefit of the simplicity of interpretation that comes with the derivation of low-order rules.

The rules generated by OSRE for the uveal melanoma data set are listed in table 3. These rules were presented to the surgeon leading the multicentre trial that acquired the data and it was judged that the sub-groups identified for each risk category were consistent with current understanding.

Representing neural network decision boundaries as Boolean rules has two important benefits. First, it provides a method for diagnosing and correcting class individual assignment failures during development and during potential prospective tests. This feature of the rule-based approach is as important as the ability to explain the drivers underpinning accepted inferences.

Secondly, the rules can capture a substantial proportion of the discriminant power of the neural network, providing a possible white-box replacement for it. This is illustrated by the Kaplan-Meier curves for the risk groups, expressed now transparently using Boolean for the risk allocations derived from the prognostic neural network model.

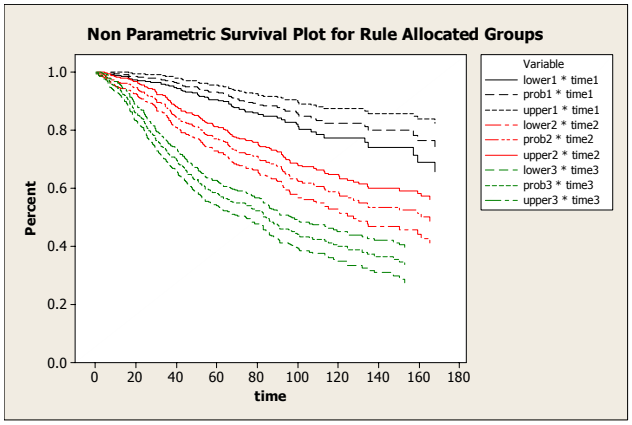

Figure 3 shows the Kaplan-Meier survival characteristic of the patients allocated into groups according to the criteria defined by OSRE extracted rules that target each PI group. It can be seen that the functions are similar to the corresponding trends in figure 1. Memberships for these groupings are not necessarily mutually exclusive. There are 608 cases that satisfy the disjunctive rule list for P.I group 1, 468 cases true for the rules explaining P.I group2 and 684 for P.I group 3.

V. VISUALISATION OF DATA AND DECISION BOUNDARIES

A claimed strength of the OSRE method is the low-order of the rules that are typically obtained even for complex, real-world medical data. This is largely a consequence of lifting the restriction of mutual exclusivity. Mutually exclusive rule trees can be trivially obtained where the nodes at each level of tree are multivariate.

Low-order rules lend themselves to direct visualization of the data at the dimensionality of the rule order. So, for

instance, where a rule involves 3 covariates, this obtains a 2-dimensional data subspace where the data separate by risk group. Given that projecting to higher dimensional spaces will not cause further missing, the visualization map reflects also separation of the full dimensional data.

However, the covariates absent from the rule will have different values. For this reason, their values are sampled by Monte Carlo on the basis of univariate distributions, generating 95% confidence intervals for the location of the decision boundary in its low-dimensional projection to the data subspace. [Terence: can you confirm this description?] A well-separated data subspace will be expected to show decision boundaries with a narrow confidence interval, while poorly separating sub-spaces, which form the vast majority of possible covariate combinations, will be almost entirely covered by the width of the confidence interval for the decision surface.

Note that the proximity of individual cases to the decision surface can now be directly ascertained. This is potentially of value both to understand erroneous assignments, should they occur, and for interpretation of borderline cases.

Once the models have been fitted, it is straightforward to apply them to out-of-sample data, for new patients, through risk stratification, application of the describing rules and projection onto the visualization maps.

VI. Conclusion

An assessment framework was described comprising complementary components to evaluate performance and describe the operation of prognostic models. This framework applies to standard generalized linear models, but equally as well to flexible models, of which neural networks were taken as an example.

The time-dependent extension of the AUROC is a rigorous method to quantify modeling accuracy for time-to-event models with censored data. This method was extended by the application of automatic rule generation and direct data visualization in separating subspaces.

This triangulation of available evidence from outcome data and domain expertise supports robust quality assurance through verification and validation of survival models and risk stratification. The methods are scalable and efficient for real-world clinical data.

Acknowledgment

The authors acknowledge funding from the Biopattern Network of Excellence FP6/2002/IST/1; N° IST-2002-508803, www.biopattern.org.

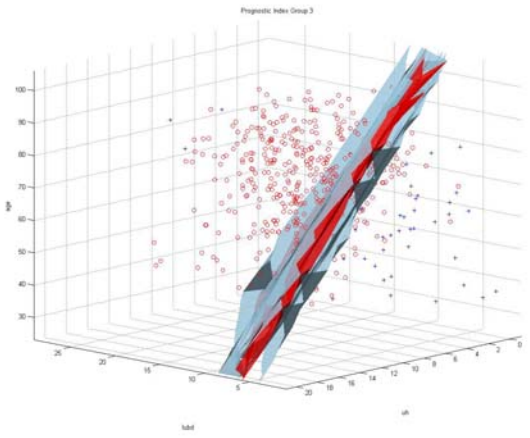

Figure 4 shows the data space defined by the three continuous variables from Rule 3 from table 3. The figure contains the decision surface of the MLP-ARD which classifies the cases in P.I group 3.

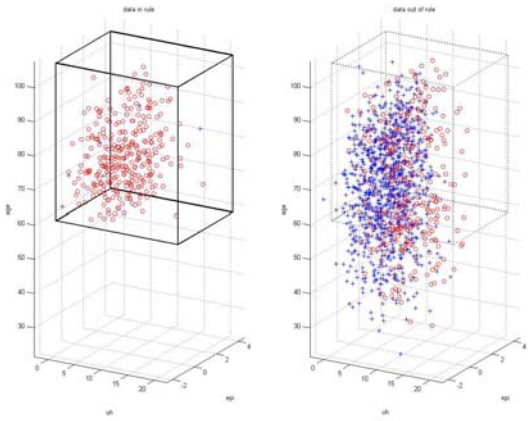

Figure 5 shows the 4 dimensional hyper-cube that is Rule 3. The left cube shows the data space when epi=1, this cube contains predominantly in-class data. The limits of the cube are defined by the other statements in rule 3 which are by definition parallel to each axis. The right cube shows the data space when epi=0, the majority of out-class cases appear on this side.

References

[1] P.J.G. Lisboa 'Industrial use of safety-related artificial neural networks' HSE CR 237/2001, HMSO 2001. http://www.hse.gov.uk/research/crr_pdf/2001/crr01327.pdf

[2] Lisboa, P.J.G. 'A review of evidence of health benefit from artificial neural networks in medical intervention', Neural Networks, 15, 1, 9-37, 2002.

[3] Lisboa, P.J.G. and Taktak, A.F.G. 'The use of artificial neural networks in decision support in cancer: A Systematic Review', Neural Networks, 19: 408-415, 2006

[4] Tilbury, J., Van-Eetvelt, P., Garibaldi, J., Curnow, J., and Ifeachor, E.C., "Receiver operator characteristic analysis for intelligent medical systems - a new approach for finding confidence intervals", IEEE Transactions Biomedical Engineering, Vol. 47 No. 7, pp. 952-963, July 2000.

[5] Antolini L, Boracchi P, Biganzoli E. A time-dependent discrimination index for survival data. *Stat.Med.* 2005;24:3927-44.

[6] 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 prognostic models in a multi-centre study' *accepted for* Computers in Medicine and Biology

[7] Biganzoli E, Boracchi P, Mariani L, Marubini E. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Stat.Med.* 1998;17:1169-86.

[8] 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, 28, 1, 1-25, 2003.

[9] Etchells, T.A. and Lisboa, P.J.G. 'Rule extraction from neural networks: a practical and efficient approach' IEEE Transactions on Neural Networks, 17 (2):374-384, 2006