Development of a five-year mortality model in systemic sclerosis patients by different analytical approaches

L. Beretta[1], A. Santaniello[1], F. Cappiello[1], N.V. Chawla[2], M.C. Vonk[3], P.E. Carreira[4], Y. Allanore[5], D.A. Popa-Diaconu[3], M. Cossu[1], F. Bertolotti[1], G. Ferraccioli[6], A. Mazzone[7], R. Scorza[1]

[1]Referral Centres for Systemic Autoimmune Diseases, Fondazione IRCCS Ospedale Maggiore Policlinico and University of Milan, Italy; [2]Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, USA; [3]Department of Rheumatology, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands; [4]Rheumatology Department, '12 de Octubre' University Hospital, Madrid, Spain; [5]Rheumatology A, Paris Descartes University, Cochin Hospital, AP-HP, Paris, France; [6]Division of Rheumatology, Department of Internal Medicine and Geriatrics, Catholic University of the Sacred Heart, Rome, Italy; [7]UO Internal Medicine and Oncology, Ospedali Civili di Legnano, Legnano, Italy.

Lorenzo Beretta, MD
Alessandro Santaniello, MD
Francesca Cappiello, MD
Nitesh V. Chawla, PhD
Madelon C. Vonk, MD, PhD
Patricia E. Carreira, MD
Yannick Allanore, MD
Delia A. Popa-Diaconu, MD
Marta Cossu, MD
Francesca Bertolotti, MD
Gianfranco Ferraccioli, MD
Antonino Mazzone, MD
Raffaella Scorza, MD

Please address correspondence to:
Dr Lorenzo Beretta, Referral Centres for Systemic Autoimmune Diseases, Fondazione IRCCS Ospedale Maggiore Policlinico and University of Milan, Via Pace 9, 20122 Milan, Italy.
E-mail: lorberimm@hotmail.com

Received on November 17, 2009; accepted in revised form on October 5, 2009.

Clin Exp Rheumatol 2010: 28 (Suppl. 58): S18-S27.

Key words: Systemic sclerosis, survival, Cox regression, data mining, Naïve Bayes.

Competing interests: none declared.

ABSTRACT

Objective. *Systemic sclerosis (SSc) is a multiorgan disease with high mortality rates. Several clinical features have been associated with poor survival in different populations of SSc patients, but no clear and reproducible prognostic model to assess individual survival prediction in scleroderma patients has ever been developed.*

Methods. *We used Cox regression and three data mining-based classifiers (Naïve Bayes Classifier [NBC], Random Forests [RND-F] and logistic regression [Log-Reg]) to develop a robust and reproducible 5-year prognostic model. All the models were built and internally validated by means of 5-fold cross-validation on a population of 558 Italian SSc patients. Their predictive ability and capability of generalisation was then tested on an independent population of 356 patients recruited from 5 external centres and finally compared to the predictions made by two SSc domain experts on the same population.*

Results. *The NBC outperformed the Cox-based classifier and the other data mining algorithms after internal cross-validation (area under receiving operator characteristic curve, AUROC: NBC=0.759; RND-F=0.736; Log-Reg=0.754 and Cox= 0.724). The NBC had also a remarkable and better trade-off between sensitivity and specificity (e.g. Balanced accuracy, BA) than the Cox-based classifier, when tested on an independent population of SSc patients (BA: NBC=0.769, Cox=0.622). The NBC was also superior to domain experts in predicting 5-year survival in this population (AUROC=0.829 vs. AUROC=0.788 and BA=0.769 vs. BA=0.67).*

Conclusion. *We provide a model to make consistent 5-year prognostic predictions in SSc patients. Its internal validity, as well as capability of generalisation and reduced uncertainty compared to human experts support its use at bedside. Available at: http://www.nd.edu/~nchawla/survival.xls.*

Introduction

Systemic sclerosis (SSc) is a complex autoimmune disease with multiorgan involvement that results in significant disability and morbility (1). Increased mortality ratios for SSc patients have been extensively reported across different countries (1-7) and many studies have focused on the analysis of factors that may eventually be associated with poor survival in scleroderma patients (5, 6, 8-16). For multivariate survival analysis, these studies have almost invariably relied on semi-parametric models, such as the Cox regression model (17). This approach, is the most widely used survival analysis method in the medical literature that is, however, not devoid of drawbacks and limitations (18). The Cox proportional hazard method relies on assumptions that can be easily violated, especially in presence of time-dependent covariates and, also when these are taken into account, the choice of covariate form has great potential for bias and does not lead to individual predictions (19). Even in the simpler Cox model with fixed covariate values, the task of making individual prediction may be computationally expensive (17, 20). Furthermore, it is well known that the Cox proportional hazards method produces poor results with many inputs (21), may suffer from multicollinearity of data and may produce unstable results when continuous and dichotomised results are entered together in the model (22). Finally, this model may not be capable of modelling interaction terms when data are scarce and dispersed through the multidimen-

sional space or in presence of non-linear interactions among variables (23).

Data mining methodologies (18, 24, 25) are also emerging as useful tools for medical researchers to build models on larger number of variables and generate disease predictions. The core of these techniques is the ability to find patterns and relationships among large amounts of data to eventually build models that can accomplish the task of assigning class label to unlabelled instances. The main advantage of data mining techniques is the capability of discriminating amongst a range of putative risk factors or causative agents and random variance, that is the ability of identifying predictive factors in absence of main independent effects and in case of multicollinearity or of non-linear interactions among variables even in relatively small datasets. Data mining techniques have, thus, successfully been applied in a variety of survival prediction tasks, including but not limited to predicting survival in breast cancer patients (26, 27), in cirrhotic subjects who underwent transjugular intrahepatic portosystemic shunt (TIPS) (28) or in critically-ill patients admitted to intensive care units (29).

In this paper, we develop a 5-year prognostic model in SSc patients using either data mining methods or a traditional Cox proportional Hazard model. The models are developed and internally validated on the large population of SSc patients from our referral centres. The generalisation of our model is finally evaluated on a panel of independent SSc patients recruited from four external centres.

Materials and methods

Patient selection and variables definition

Five hundred and fifty-eight consecutive patients with a diagnosis of SSc referred to our outpatient clinic between 1982 and 2008 were considered for analysis. The majority of our patients (88%) fulfilled the preliminary criteria for the classification of SSc proposed by the American college of rheumatology (ACR) (30), yet we also considered a proportion of patients with definite SSc who do not fulfill these criteria (5, 31).

All the clinical and laboratory variables were either selected on the basis of well-standardised definitions or their inclusion was motivated by pre-existing medical literature. These variables can be roughly divided in variables that do not change over time or "time-dependent" variables (*e.g.* clinical features that may develop anytime during the course of the illness). In the latter case, if the feature was ascertained within one year from diagnosis it was considered as "present", otherwise as "absent" even if it developed later on. These variables are hereafter briefly described.

All causes of deaths related to or possibly-related to SSc, according to the definition of Ferri *et al*. (5) were considered.

The patients were categorised as having the limited cutaneous (lcSSc) or the diffuse cutaneous (dcSSc) subset of the disease, according to LeRoy *et al*. (32) and the patients' autoantibody profile was determined by reviewing the patients' medical records.

Observation was started from the year of diagnosis (*e.g.* referral) and the age of 45 at referral was used as a threshold to categorise the patients (12). Referral delay was defined as the time elapsing between diagnosis and the time of the appraisal of the first non-Raynaud symptom (33). Restrictive lung disease was defined as a forced vital capacity (FVC) <70% of predicted values plus a forced expiratory volume in 1 second >70% of the FVC (6). A severe impairment of the diffusing capacity for carbon monoxide (DLco) was considered present for values ≤55% of the predicted (34); DLco values were corrected for the patient's haemoglobin concentrations. A right ventricular systolic pressure (RVSP) ≥45 mmHg on echocardiography was considered as threshold to estimate the presence of pulmonary hypertension (35). Renal involvement was defined as a serum creatinine >2.5 mg/dL or a creatinine clearance <40 mL/min on at least two consecutive determinations or the presence of "scleroderma renal crisis" (6). Gastroesophageal reflux disease was defined as the presence of heartburn or dysphagia alleviated by the use of proton pump inhibitors or anti-H_2 drugs (6). Systemic inflammation was defined as the presence of an erythrocyte sedimentation rate (ESR) ≥25 mm/h on at least two consecutive occasions without evidence of concurrent infection (10). Anaemia was ascertained for haemoglobin levels <12.5 g/dL (10). Digital ulcers were defined as the loss of surface epithelialisation (36). Muscle weakness, arthralgias or arthritis (5), were also considered.

Statistical analysis

1. Cox regression

We used first univariate Cox regression analysis, controlling that the proportionality of hazards was not violated (17), to test the different variables for their ability to predict mortality at 5 years from disease onset. *P*-values were adjusted by Bonferroni correction (p_c=p*16) and variables with a p_c<0.05 were then deemed interesting and inserted in a multivariate forward-stepwise Cox regression model; in this model a *p*-value less than 0.01 was selected as entry criterion at each step. The hazard ratios (HR), which can be interpreted as the relative risk of dying due to SSc, are presented with their 95% confidence intervals (CI_{95}).

We then built a Cox-based classifier to make predictions from the Cox proportional hazard model (20). As a first step, the mortality probability $m(t_5)$ for any case at 5 years from diagnosis was calculated by the following formula:

$$m(t_5) = 1 - \exp [-H_0(t_{10}) \times (X_1B_1 + X_2B_2 + \ldots X_nB_n)]$$

where $H_0(t_5)$ is the baseline cumulative hazard at 5 years and B_n are the regression coefficients for the X_n covariates included in the final model. Individual $m(t_5)$ were then used to build the receiving operator characteristic (ROC) curve (37). If the individual $m(t_5)$ was equal or fell above a threshold T, as defined under point 3.a, the patient was classified as "dead", otherwise as "alive".

Cox analysis was performed by the SPSS ver. 17.0 software (SPSS Inc, Chicago, IL).

2. Data mining

The following trivial rules were applied to handle censorship with data mining classifiers: patient alive ≥5 years from

disease onset were classified as "alive", patients who died within 5 years from disease onset were classified as "dead" while patients alive but with a disease duration <5 years were excluded from analysis. Thus, 353 cases were available for analysis by data mining classifiers.

We incorporated the following three steps in building our data mining based model: a) attribute selection; b) comparative evaluation of different classifiers or learning algorithms to select one classifier; and c) use of resampling techniques to counter the issue of high class imbalance in our data. The classifiers are re-learned from the resampled datasets to further improve the performance on the SSc survival task.

2a. Selection of attributes

The selection of the most interesting attributes (or conversely, the removal of noisy variables) among a pool of possible candidates may improve the quality of the signal, reducing the chance of classifier overfitting and increasing its overall accuracy (27, 36, 37).

To filter the dataset, we used information-theory based measures, such as Gain Ratio, a normalised variant of information gain (40). Gain Ratio estimates feature weights by examining the training data and determines for each feature how much information it contributes to the knowledge of the classes of the training data items. The open-source Orange data mining software (available at: http://www.ailab.si/orange) was used to calculate Gain Ratios and the top 5 attributes were selected for the further analyses.

2b. Data mining learning algorithms

The following learning algorithms/data mining classifiers were used for analysis: logistic regression models, Naïve Bayes classifier (NBC) and Random Forests (RND-F). We chose these three classifiers as they are each with different inductive biases, providing us with a broad coverage for evaluation and comparison. We used their implementations in the Orange data mining software for all the analyses.

For the NBC, Laplace correction was used as smoothing method for estimating posterior probabilities. In the logistic regression model, the forward stepwise procedure with *p*-values equal to (0.01) for entry and (0.1) for removal was used. For the RND-F, 10 forests of 500 trees each were built and then the forest with the best performance was used as the final classifier. Finally, nomograms were used to visualise the NBC results and to expose the quantitative information on the effect of attribute values to class probabilities (39).

2c. Resampling of the dataset

The natural distribution of data is often regarded as non-optimal for learning a classifier as it would overestimate the importance of the majority class and reduce the rate of correct detection in the minority class, especially when the dataset is highly unbalanced (42). To reduce the cost to misclassify the minority class we used the wrapper method based on the synthetic minority over-sampling technique (SMOTE) introduced and implemented by Chawla *et al*. (43-45). The wrapper method consists of a combination of under-sampling the majority class and over-sampling the minority class by creating synthetic minority class examples. Undersampling implies randomly removing examples from the majority class (alive patients). SMOTE generates new synthetic examples for the minority class (in our case the SSc mortality class) to improve the predictive capacity of the classifiers. Generating new examples provides additional information to the classifiers, improving the overall true positive rate in the testing set. While this may also increase false positives, our conjecture is that the increase in true positives overwhelms the relative increase in false positives, especially in comparison to the other methods, including Cox regression and domain expert. The wrapper method generated optimal levels of sampling (under-sampling and SMOTE) such that the performance of each of the three classifiers is independently optimised. Note that we can optimise this performance on the data from our referral centre, as our real test is on the three external centres. Nevertheless, we used different internal cross-validation procedure to optimise the sampling levels and evaluate the performance of classifiers.

3. Measures for performance evaluation

– 3a Accuracy, sensitivity, specificity and ROC curves

The performance of the classifiers was evaluated by means of sensitivity, specificity, accuracy and area under ROC curve (AUC) (37). To label the prediction form each classifier, among all the possible thresholds T that constitute the coordinates of the ROC curves, the value that resulted in the highest classification accuracy was then chosen (37).

When binary classifiers were finally evaluated in the external population (see point 3c) T was pre-defined and thus we used the balanced accuracy (BA) function introduced by Velez *et al*. (46) to establish the trade-off between sensitivity and specificity for that given T. The BA, -defined as the mathematical mean of sensitivity and specificity-, is an appropriate measure for performance evaluation in unbalanced datasets and it is mathematically equivalent to the raw accuracy in datasets with a 1:1 cases to controls ratio.

3b. Internal cross-validation (k-fold cross-validation)

Internal cross-validation is used to determine how well a learning algorithm will fit in independent datasets (47). The principles of k-fold cross validation are here briefly described: the dataset is divided into k mutually exclusive subsets of approximately equal size, the learning algorithm is then trained on each k-1 subset (the training subset) and its prediction are then verified on the corresponding k subset (the testing subset). The performance measures across all k trials are computed and then averaged to determine the performance of the k-fold cross-validation. The average of the performance measure provides an estimate of the performance of the classifier constructed from the whole dataset. For the current analysis we used the 5 stratified-fold cross-validation method.

For cross-validation of the SMOTE-resampled dataset, the original dataset was first divided in 5 training and 5 testing fold and then the training folds were resampled. Each training fold was independently re-sampled to optimise

the performance of a classifier, and then the classifier was evaluated on the corresponding testing fold. The procedure for this is detailed in Chawla *et al*. (43). The corresponding testing fold was never included for optimising the levels of sampling. This ensures that the reported performance in this paper is true "unseen" testing performance even for our referral centres data.

3.c External cross-validation

The predictions of the classifiers were tested in an independent population which consisted of 356 cases recruited from 5 external centres: Rome (FGF), 23 cases; Legnano (AM), 38 cases; Paris (YA), 60 cases; Madrid (PEC), 105 cases and Nijmegen (MCV and DAPD), 130 cases. These cases were selected in each centres from consecutive SSc patients and by applying the general inclusion criteria and the trivial rules described under point 2.

3d Domain experts

The predictive ability of the classifier with the highest performance after external cross-validation, that is the model with the highest reproducibility, was finally tested by comparison with two domain experts (RS, YA) as described by Razavi *et al*. (48). The raw data of the cases used for external cross-validation were blindly presented to the domain experts, and for each case, they were asked to rate the probability, from 0 to 100%, the patient will not be alive after 5 years from disease onset. Individual probabilities were averaged and then used to build ROC curves; performance measures were then calculated as described under point 3a. For classification purposes, the optimal threshold on the ROC curve was derived from the survival probability in this population.

Results

Clinical and demographic characteristics of the 558 patients referring to our centres (training population) and of the 356 patients referring to the external centres (testing population) are reported in Table I. No statistical differences between the populations were observed for most clinical parameters, albeit a reduction of SSc-specific autoantibodies (ACA or Scl70) was observed in the testing population. Similarly, mortality was comparable in the two groups of patients.

Univariate analysis sorted out the following variables as relevant to 5-years survival: male gender (χ^2=17.508, p_c<0.001); age ≥45 years at disease onset (χ^2=8.58, p_c<0.05); the presence of a FVC ≤70% of predicted (χ^2=8.186, p_c<0.05); the presence of a DLco ≤55% of predicted (χ^2=16.678, p_c<0.0001); an increased ESR (χ^2=13.929, p<0.01); the presence of pulmonary hypertension on echocardiography (χ^2=11.949, p_c<0.05) and the presence of renal involvement (χ^2=29.492, p_c<0.0001).

The multivariate analysis showed an increased mortality risk for the following variables: age >45 years at disease onset; the male gender; a markedly decreased DLco and/or the presence of renal involvement (Table II). The same variables plus a reduced FVC had the highest Gain Ratio after entropy-based analysis and were used to build data-mining classifier-based prediction models.

Cross-validation

The performance of the Cox-based classifier, along with that of the other learning algorithms after internal 5-fold cross-validation is reported in Table III. All the data mining algorithms had higher AUCs than the Cox-based classifier, either in the unbalanced or in the SMOTE-resampled datasets. Resampling improved the performance of the NBC and RND-F classifier, increasing both AUC and sensitivity, penalising only marginally the specificity and the overall accuracy. But, improved AUC and sensitivity is of a bigger concern here, as accuracy is misleading given the highly imbalanced nature of the dataset. The NBC trained on the SMOTE-resampled dataset ranked first (as for AUC) among all the classifiers we built after internal cross-validation (Table III), that is it was the model that was most likely to generalise to independent datasets. The capability of generalisation and overall predictive ability of this model is indeed confirmed after external cross-validation as illustrated in Table IV.

Table I. Demographics.

Variable	Training population* (n=558)	Testing population** (n=356)
Females, n (%)	499 (89.4)	177 (87.8)
dcSSc, n (%)	150 (26.9)	74 (29.5)
Fatalities, n (%)	59 (10.6)§	36 (10.1)
Age at onset >45 yrs, n (%)	355 (63.3)	224 (62.9)
Autoantiboides, n (%)		
ANA	541 (97)	331 (92.9)
ACA	191 (34.2)	104 (29.2)
Scl70	244 (43.7)	115 (32.3)
FVC ≤70% predicted, n(%)	78 (14)	36 (10.1)
DLco ≤55% predicted, n (%)	137 (24.6)	58 (16.3)
Renal involvement, n (%)	24 (4.3)	17 (4.8)
RVSP ≥45 mmHg	5.4 (9.7)	36 (10.1)
Systemic inflammation, n (%)	162 (29)	Nr
Oesophageal involvement, n (%)	409 (73.3)	Nr
Anemia, n (%)	91 (16.3)	Nr
Digital ulcers, n (%)	301 (53.9)	Nr
Arthralgias, n (%)	57 (10.2)	Nr
Arthritis, n (%)	33 (5.3)	Nr
Weakness, n (%)	47 (8.4)	Nr

Clinical and demographic characteristics in patients from the Milan centres (training population) and in the population pooled from the Rome, the Legnano, the Paris, the Madrid and the Nijmegen centres (testing population). *This population includes 203 patients with a disease duration <5 years (truncated data); **This population does not include patients with a disease duration <5 years; §At 5 years, frequency calculated excluding truncated data; Nr, not required for external cross-validation/not assessed. dcSSc, diffuse cutaneous subset; ANA, antinuclear antibodies; ACA, anti-centromere antibodies; Scl70, anti-topoisomerase I; FVC, forced vital capacity; DLco, diffusing capacity for carbon monoxide; RVSP, right ventricular systolic pressure.

Table II. Cox regression results.

Variable	Alive, n	Alive, %	Dead, n	Dead, %	B	HR	CI_{95}	*p*-value
Age at onset								
≤45 years	192	94.6	11	5.4		1 (ref)		
>45 years	307	86.5	48	13.5	1.124	3.08	1.59 – 5.97	<0.001
Gender								
Females	455	91.2	44	8.8		1 (ref)		
Males	44	74.6	15	23.4	1.101	3.01	1.66 – 5.45	0.001
DLco								
>55% predicted	391	92.9	30	7.1		1 (ref)		
≤55 predicted	108	78.8	29	11.2	0.94	2.56	1.57 – 4.29	<0.001
Renal involvement								
No	485	90.8	49	9.2		1 (ref)		
Yes	14	58.3	10	41.7	1.483	4.405	2.22 – 8.74	<0.001

Variables associated with poor 5-year survival in our population of systemic sclerosis patients. $H_0(t5)$, baseline cumulative hazard at 5 years = 0.033; B, regression coefficient. All the other definitions as in Table I. Patients are classified as "dead" at 5 years when the estimated mortality probability at 5 years, calculated as defined in the text, is ≥0.074.

Table III. Internal cross-validation.

Classifier	ROC-AUC	Accuracy	Sensitivity	Specificity
COX-based	0.724 ± 0.044	0.851 ± 0.234	0.165 ± 0.123	0.989 ± 0.168
Logistic regression				
Original	0.747 ± 0.111	0.844 ± 0.019	0.218 ± 0.06	0.969 ± 0.022
SMOTE	0.754 ± 0.091	0.793 ± 0.038	0.508 ± 0.195	0.85 ± 0.057
Naïve Bayes				
Original	0.75 ± 0.111	0.839 ± 0.012	0.216 ± 0.083	0.936 ± 0.022
SMOTE	*0.759 ± 0.101*	*0.782 ± 0.046*	*0.626 ± 0.175*	*0.813 ± 0.028*
Random Forest				
Original	0.732 ± 0.079	0.838 ± 0.01	0.084 ± 0.096	0.989 ± 0.017
SMOTE	0.736 ± 0.005	0.765 ± 0.008	0.539 ± 0.016	0.809 ± 0.011

Performance of the different classifiers after internal 5-fold cross-validation either on the original or on the synthetic minority oversample technique (SMOTE)-resampled datasets. ROC-AUC, area under receiving operator characteristics curve. Values expressed as mean ± standard deviation of 5 cross-validations; for the Random Forests, values are the average of 5 cross-validations run on 10 forests with 10 different random seeds. The best model that maximises the ROC-AUC is indicated in italicface type.

Table IV. External cross-validation.

Classifier	Accuracy	Balanced accuracy	Sensitivity	Specificity
COX-based	0.874	0.622	0.305	0.936
Logistic regression				
Original	0.888	0.592	0.222	0.963
SMOTE	0.803	0.755	0.694	0.816
Naïve Bayes				
Original	0.888	0.592	0.222	0.963
SMOTE	*0.777*	*0.769*	*0.722*	*0.816*
Random Forest				
Original	0.883 ± 0.008	0.547 ± 0.007	0.128 ± 0.025	0.967 ± 0.001
SMOTE	0.788 ± 0.002	0.751 ± 0.008	0.705 ± 0.002	0.797 ± 0.00

Performance of the different classifiers trained either on the original or on the synthetic minority oversample technique (SMOTE)-resampled datasets after external cross-validation in 356 cases. The best Model that maximises the balanced accuracy is indicated in italic face type.

The performance of the NBC in the single testing populations was as follows: Rome & Legnano, AUC=0.883, BA=0.865;Paris,AUC=0.847,BA=0.777; Nijmegen, AUC=0.892, BA=0.902; Madrid, AUC=0.836, BA=0.698.

Nomograms for the NBC models built on SMOTE-resampled dataset plotted in Fig. 2 allow the quantification of the relative importance of each variable in class prediction. By this graphical tool, the predictive ability of the model can eventually be tested in any SSc population by calculating the scores from each individual case and the corresponding predicted *p*-values. Nomograms can also allow class prediction when the value of a state variable is unknown, which is given a neutral value. To assist the reader and to ease the prediction from our model, we also provide a simple calculator which is illustrated in Appendix and is freely available at: http://www.nd.edu/~nchawla/survival.xls.

Domain experts

The ROC curve obtained from averaging the predictions from two domain experts, blinded to the previous results (RS and YA) is plotted in Fig. 2, along with the ROC curve of the NBC classifier trained on the SMOTE-resampled dataset. On the testing population, the NBC clearly outperformed the averaged prediction from the two-domain expert (AUC=0.829 *vs*. AUC=0.787). This resulted in a much better trade-off between sensitivity and specificity either (BA=0.769 for the NBC *vs*. 0.67 for domain experts).

Discussion

Ever since SSc was recognised as one of the rheumatic diseases with the worst prognosis (2-7), several studies across different countries have been conducted to find the prognostic factors associated with scleroderma poor outcome (5, 6, 8-15). These analyses have consistently described an increased mortality-risk in patients with the dcSSc subset of the disease along with any major organ involvement, leading to the concept that different SSc populations share similar prognostic factors and overall prognosis. It is nonetheless surprising to observe that none of the SSc survival studies conducted so far fully answered to a simple, yet often neglected, question: what is the individual mortality-risk of a patient when the combinations or the interactions of prognostic factors, considered both in term of presence and absence, is taken into account? The models described Bryan

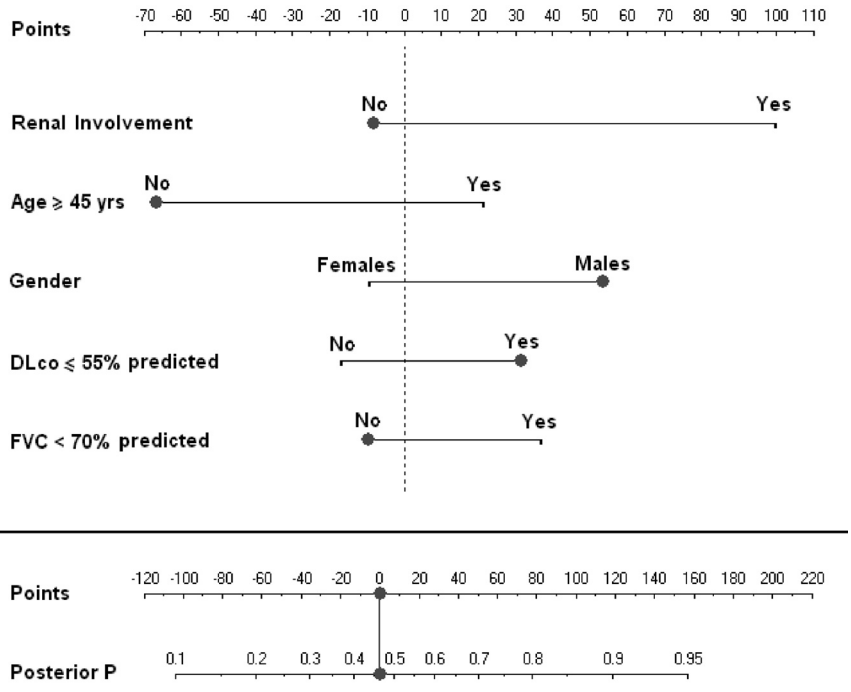


Fig. 1. Nomogram for the NBC. Individual point scores and corresponding posterior probability calculated from the naïve Bayes classifier trained on the dataset resampled by the Synthetic Minority Oversample Technique. To make a prediction, the contribution of each attribute is measured as a point score (topmost axis in the nomogram), and the individual point scores are summed to determine the probability of survival (bottom two axes of the nomogram). Patients with overall *p*-values ≥0.588, are classified as "dead" after 5 years from diagnosis, otherwise as "alive". The dots indicate the characteristics of an example male patient who aged <45 years at onset, without renal involvement a forced vital capacity (FVC) ≥70% of predicted values and a diffusing capacity for carbon monoxide ≤55% of predicted values. The posterior probability to be dead at 5 years from disease onset is calculated to be 0.46 (score=0.13) and hence the patient is predicted to be alive.

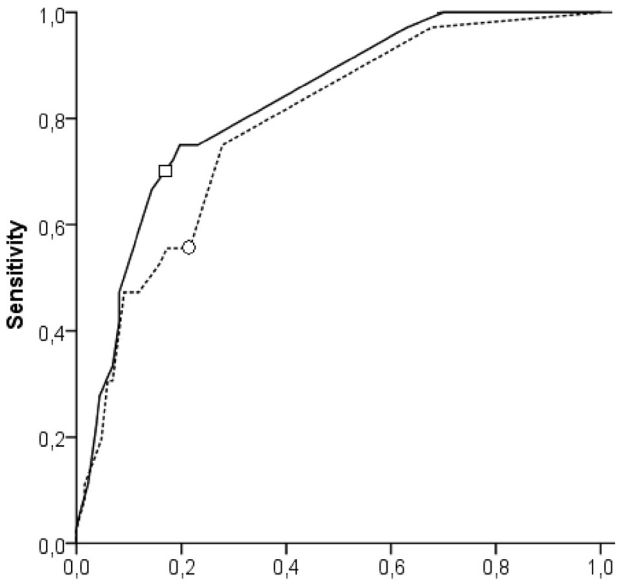


Fig. 2. Performance comparison of domain experts and NBC
Receiving operator characteristic (ROC) curve in the testing population plotted considering either the averaged predictions from two domain experts (dotted line) or the predictions from the naïve Bayes classifier (NBC) trained on the resampled dataset (black line). NBC area under curve (AUC)=0.829; domain experts AUC=0.788. The square point indicates the optimal threshold for the NBC derived from the training population; the circle indicates the threshold for domain experts derived from the observed mortality (=10.1%).

et al. (13) and by Scussel-Lonzetti *et al*. (10), do represent a first tentative in this direction, yet ultimately fail to solve this important issue. The finding that, in absence of an internal validation method, by increasing the number of predictors the expected mortality increases, does not add any clue about the way these variables interact and would represent a clear case of adaption of the model to the data (overfitting) (49). Thus, the significance of the reported results may be overestimated and misleading, as well as difficult to interpreter. If, for instance, we consider the model proposed Scussel-Lonzetti *et al*. (10), which identified older age, anemia, inflammation, an impaired DLco and the dcSSc subset as predictors, we cannot conclude anything but that aged patients with anemia have the very same mortality risk of dcSSc patients with lung involvement, even if it seems unlikely.

The prevalence of organ involvement in our population is largely comparable to the testing population and to other case-series (8, 10), including a large meta-analysis conducted in 1645 SSc patients (6). Accordingly, the prognostic factors we sorted out do not differ much from those described in previous reports, yet compared to previous reports, our model allows a better use of the available information. Given the inherent probabilistic nature of the NBC, instead of attributing a generic risk-pattern to individuals on the basis of the single clinical characteristics of theirs (*i.e*. males have a 3-fold reduced survival compared to females), we can calculate for each patient a precise survival probability that can be transformed into a yes/no answer to the question: "will the patient survive after 5-years from diagnosis?". One of the main advantages of Bayesian classifiers is that they are robust to real data noise and missing values and that they perform efficiently well when the sample size is small. This resulted in a consistent 5-year survival prediction, that is a better balance between sensitivity and specificity compared to the other statistical approaches we employed, including Cox regression (Tables III and IV).

Whilst the development of our prognostic model may seem quite sophisticated and complex, its practical use is surprisingly simple. Nomograms depicted in Fig. 1, illustrate how it possible to calculate individual survival probabilities on the basis of the clinical characteristics of a patient also in presence of missing values (see also Appendix for a description of an excel-based prognostic calculator). Noteworthy, our model was both internally validated and tested

on an independent external population, confirming the robustness and capability of generalisation of the provided predictions. Yet, as with any computational tool used in problem-solving, the ultimate goal of decision makers is to reduce uncertainty and to outperform their counterparts, embodied in human and domain experts, that is those persons who possess special knowledge or skill in the field. Indeed, our model appeared to be more robust than humans in its ability to objectively analyse patients' data and to reach unbiased conclusions. Compared to the clinicians' predictions, the NBC model had a much higher sensitivity (0.722 *vs*. 0.556) as well as a higher specificity (0.816 *vs*. 0.784). From a practical point of view, the improved ability of our model to discriminate patients at risk, otherwise dampened by over-optimism by health professionals, may facilitate better patient care and treatment. Furthermore, it has been observed that both patients and their families may later regret being over-optimistic about their prognosis (48) and that patients are willing to have access to accurate prognostic information (49).

In the evaluation of our prognostic model its inherent "static" nature should be carefully considered. Due to inclusion/exclusion criteria and variables definition, this model do not take into account the evolution of the disease, that is, it is assumed that during the first 5 years of the disease the mortality risk for a given patient is linear given the "early" (*e.g*. within the 1st year from diagnosis) clinical characteristics of his. Furthermore, it is assumed that both in the training and testing centres the optimal therapy based on the current practice is prescribed (52).

I summary, we described a framework to build a prognostic model to predict 5-year mortality in SSc patients by data mining algorithms and we demonstrated, for the first time, that it is possible to make accurate individual predictions. The advantages of the model we created include its logical simplicity, its biological plausibility as well as its capability of generalisation and applicability that would ultimately support its diffusion in the clinical practice.

Acknowledgements

We would like to thank David A Cieslak at the Department of Computer Science and Engineering, University of Notre Dame, for his assistance in preparing the SMOTE-resampled datasets.

We are also grateful to Dr Jaap Fransen at the Department of Rheumatology, Radboud University Nijmegen Medical Centre for his help in collecting the data from non-Italian centres.

References

1. VILLAVERDE-HUESO A, SÁNCHEZ-VALLE E, ALVAREZ E *et al*.: Estimating the burden of scleroderma disease in Spain. *J Rheumatol* 2007; 34: 2236-42.
2. BRYAN C, HOWARD Y, BRENNAN P, BLACK C, SILMAN A: Survival following the onset of scleroderma: results from a retrospective inception cohort study of the UK patient population. *Br J Rheumatol* 1996; 35: 1122-6.
3. MENDOZA F, DERK CT: Systemic sclerosis mortality in the United States: 1999-2002 implications for patient care. *J Clin Rheumatol* 2007; 13: 187-92.
4. KRISHNAN E, FURST DE: Systemic sclerosis mortality in the United States: 1979-1998. *Eur J Epidemiol* 2005; 20: 855-61.
5. FERRI C, VALENTINI G, COZZI F *et al*.; Systemic Sclerosis Study Group of the Italian Society of Rheumatology (SIR-GSSSc). Systemic sclerosis: demographic, clinical, and serologic features and survival in 1,012 Italian patients. *Medicine* (Baltimore) 2002; 81: 139-53.
6. IOANNIDIS JP, VLACHOYIANNOPOULOS PG, HAIDICH AB *et al*.: Mortality in systemic sclerosis: an international meta-analysis of individual patient data. *Am J Med* 2005; 118: 2-10.
7. MAYES MD, LACEY JV JR, BEEBE-DIMMER J *et al*.: Prevalence, incidence, survival, and disease characteristics of systemic sclerosis in a large US population. *Arthritis Rheum* 2003; 48: 2246-55.
8. SIMEÓN CP, ARMADANS L, FONOLLOSA V *et al*.: Mortality and prognostic factors in Spanish patients with systemic sclerosis. *Rheumatology* (Oxford) 2003; 42: 71-5.
9. ALTMAN RD, MEDSGER TA JR, BLOCH DA, MICHEL BA: Predictors of survival in systemic sclerosis (scleroderma). *Arthritis Rheum* 1991; 34: 403-13.
10. SCUSSEL-LONZETTI L, JOYAL F, RAYNAULD JP *et al*.: Predicting mortality in systemic sclerosis: analysis of a cohort of 309 French Canadian patients with emphasis on features at diagnosis as predictive factors for survival. *Medicine* (Baltimore) 2002; 81: 154-67.
11. CZIRJÁK L, KUMÁNOVICS G, VARJÚ C *et al*.: Survival and causes of death in 366 Hungarian patients with systemic sclerosis. *Ann Rheum Dis* 2008; 67: 59-63.
12. STEEN VD, MEDSGER TA: Changes in causes of death in systemic sclerosis, 1972-2002. *Ann Rheum Dis* 2007; 66: 940-4.
13. BRYAN C, KNIGHT C, BLACK CM, SILMAN AJ: Prediction of five-year survival following presentation with scleroderma: development of a simple model using three disease factors at first visit. *Arthritis Rheum* 1999; 42: 2660-5.
14. HESSELSTRAND R, SCHEJA A, AKESSON A: Mortality and causes of death in a Swedish series of systemic sclerosis patients. *Ann Rheum Dis* 1998; 57: 682-6.
15. JACOBSEN S, HALBERG P, ULLMAN S: Mortality and causes of death of 344 Danish patients with systemic sclerosis (scleroderma). *Br J Rheumatol* 1998; 37: 750-5.
16. KARASSA FB, IOANNIDIS JP: Mortality in systemic sclerosis. *Clin Exp Rheumatol* 2008; 26 (Suppl. 51): S85-93.
17. COX DR, OAKES D: Analysis of survival data. London: Chapman and Hall, 1984.
18. BATH PA: Data mining in health and medical information. *In* CRONIN B (Ed.) *Annual Review of Information Science and Technology* (ARIST) 2004; 38: 331-68.
19. FISHER LD, LIN DY: Time-dependent covariates in the Cox proportional-hazards regression model. *Annu Rev Public Health* 1999; 20: 145-57.
20. OHNO-MACHADO L: A comparison of Cox proportional hazards and artificial neural network models for medical prognosis. *Comput Biol Med* 1997; 27: 55-65.
21. BAKKER B, HESKES T, NEIJT J, KAPPEN B: Improving Cox survival analysis with a neural-Bayesian approach. *Stat Med* 2004; 23: 2989-3012.
22. BAKKER SJ: Crippling of inflammatory markers as predictors of death by dichotomization and multicollinearity. *Nephrol Dial Transplant* 2006; 21: 2990-1.
23. FAYYAD U, PIATETSKY-SHAPIRO G, SMYTH P: From Data Mining to Knowledge Discovery in Databases (a survey). *AI Magazine Fall* 1996; 17: 37-54.
24. BERETTA L, CAPPIELLO F, MOORE JH, SCORZA R: IL-1 Gene Complex Single Nucleotide Polymorphisms in Systemic Sclerosis: a Further Step Ahead. *Human Immunol* 2008; 69: 187-92.
25. LAVRAC N: Selected techniques for data mining in medicine. *Artif Intell Med* 1999; 16: 3-23.
26. JONSDOTTIR T, HVANNBERG ET, SIGURDSSON H, SIGURDSSON S: The feasibility of constructing a Predictive Outcome Model for breast cancer using the tool of data mining. *Expert Systems with Applications* 2008; 34: 108-118.
27. DELEN D, WALKER G, KADAM A: Predicting breast cancer survivability: a comparison of three data mining methods. *Artif Intell Med* 2005; 34: 113-27.
28. BLANCO R, INZA I, MERINO M, QUIROGA J, LARRAÑAGA P: Feature selection in Bayesian classifiers for the prognosis of survival of cirrhotic patients treated with TIPS. *Biomed Inform* 2005; 38: 376-88.
29. RAMON J, FIERENS D, GÜIZA F *et al*.: Mining data from intensive care patients. *Advanced Engineering Informatics* 2007; 21: 243-56.
30. Preliminary criteria for the classification of systemic sclerosis (scleroderma). Subcommittee for Scleroderma Criteria of the American Rheumatism Association Diagnostic and Therapeutics Criteria Committee. *Arthritis Rheum* 1980; 23: 581-90.

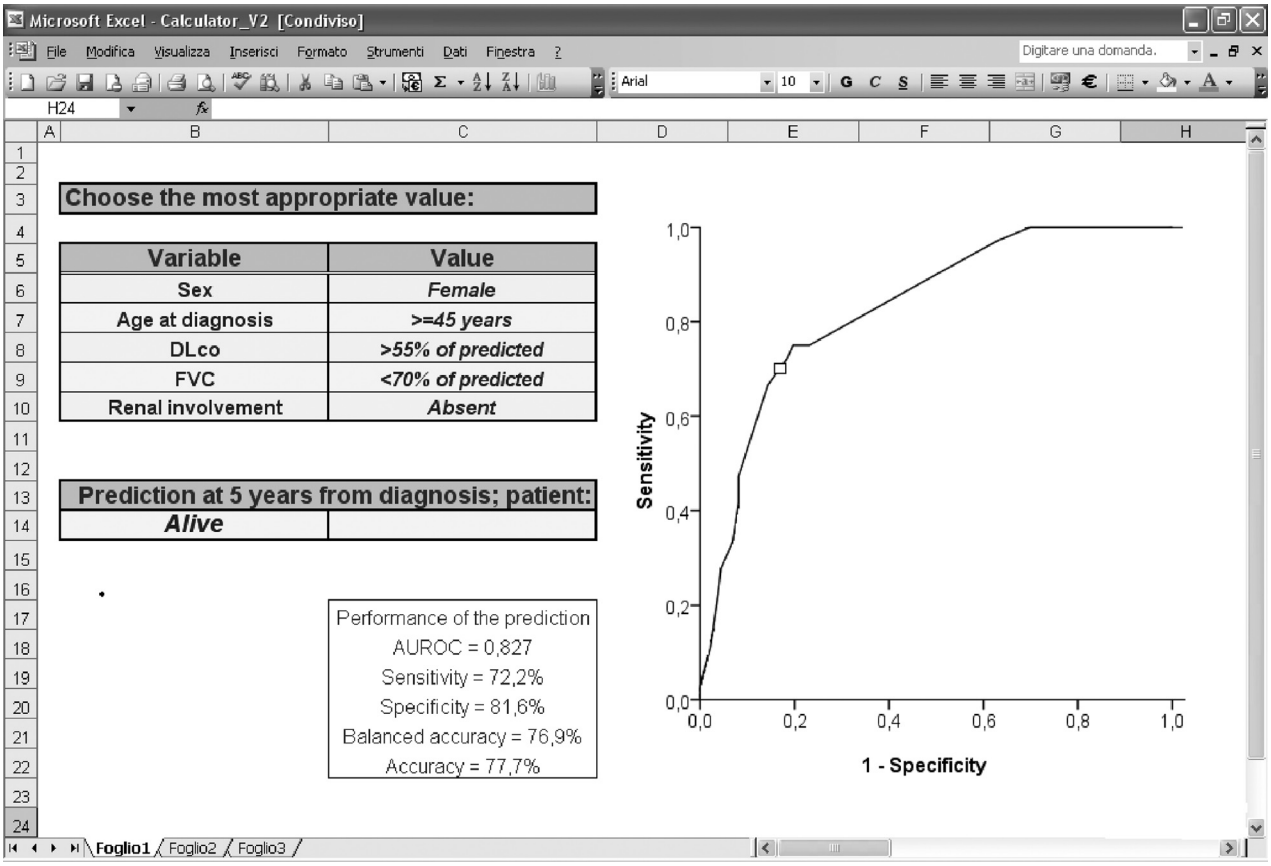
31. HUDSON M, TAILLEFER S, STEELE R *et al*.: Improving the sensitivity of the American College of Rheumatology classification criteria for systemic sclerosis. *Clin Exp Rheumatol* 2007; 25: 754-7.
32. LEROY EC, BLACK C, FLEISCHMAJER R *et al*.: Scleroderma (systemic sclerosis): classification, subset and pathogenesis. *J Rheumatol* 1988; 15: 202-5.
33. WHITE B, BAUER EA, GOLDSMITH LA, HOCHBERG MC, KATZ LM, KORN JH: Guidelines for clinical trials in systemic sclerosis (scleroderma). I. Disease-modifying interventions. The American College of Rheumatology Committee on Design and Outcomes in Clinical Trials in Systemic Sclerosis. *Sem Arthritis Rheum* 1995; 38: 351-60.
34. STEEN VD, GRAHAM G, CONTE C, OWENS G, MEDSGER TA JR: Isolated diffusing capacity reduction in systemic sclerosis. *Arthritis Rheum* 1992; 35: 765-70.
35. MUKERJEE D, ST GEORGE D, KNIGHT C *et al*.: Echocardiography and pulmonary function as screening tests for pulmonary arterial hypertension in systemic sclerosis. *Rheumatology* (Oxford) 2004; 43: 461-6.
36. KORN JH, MAYES M, MATUCCI CERINIC M *et al*.: Digital ulcers in systemic sclerosis: prevention by treatment with bosentan, an oral endothelin receptor antagonist. *Arthritis Rheum* 2004; 50: 3985-93.
37. VUK M, CURK T: ROC curve, Lift Chart and Calibration Plot. *Metodološki Zvezki* 2006; 3: 89-108.
38. BERETTA L, CAPPIELLO F, MOORE JH, BARILI M GREENE CS, SCORZA R: Epistatic interactions of cytokine single nucleotide polymorphisms predict susceptibility to disease subsets in systemic sclerosis patients. *Arthritis Rheum* 2008; 59: 974-83.
39. KONONENKO I: Inductive and bayesian learning in medical diagnosis. *Appl Artif Intell* 1993; 7: 317-37.
40. SAEYS Y, INZA I, LARRAÑAGA P: A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007; 23: 2507-17.
41. MOŽINA, M, DEMŠAR J, KATTAN M, ZUPAN B: 2004. Nomograms for visualization of naive Bayesian classifier. *In* PROCEEDINGS OF THE 8TH EUROPEAN CONFERENCE ON PRINCIPLES AND PRACTICE OF KNOWLEDGE DISCOVERY IN DATABASES (Pisa, Italy, September 20-24, 2004). BOULICAUT J, ESPOSITO F, GIANNOTTI F, PEDRESCHI D (Eds.) *Lecture Notes In Computer Science,* vol. 3202. Springer-Verlag New York, New York, NY, 337-48.
42. CHAWLA, NV, JAPKOWICZ N, KOTCZ A: Editorial: special issue on learning from imbalanced data sets. *SIGKDD Explor Newsl* 2004; 6: 1-6.
43. CHAWLA NV, BOWYER KW, HALL LO, KEGELMEYER WP. SMOTE: Synthetic Minority Over-sampling Technique. *J Artif Intell Res* 2002; 16:321-357.
44. CHAWLA NV, CIESLAK DA, HALL LO, JOSHI AJ: Automatically countering imbalance and its empirical relationship to cost. *Data Mining and Knowledge Discovery Journal* 2008 Feb 17; [Epub ahead of print]. DOI 10.1007/s10618-008-0087-0.
45. BATISTA GEAPA, PRATI RC, MONARD MC: A study of the behaviour of several methods for balancing machine learning training data. *SIGKDD Explorations* 2004; 6: 20-9.
46. VELEZ DR, WHITE BC, MOTSINGER AA *et al*.: A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genet Epidemiol* 2007; 31: 306-15.
47. KOHAVI R: A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, 1137-43. San Mateo, CA: Morgan Kaufmann, 1995.
48. RAZAVI AR, GILL H, ÅHLFELDT H, SHAHSAVAR N: Predicting metastasis in breast cancer: comparing a decision tree with domain experts. *J Med Syst* 200; 31: 263-73.
49. CONCATO J, FEINSTEIN AR, HOLFORD TR: The risk of determining risk with multivariable models. *Ann Intern Med* 1993; 118: 201-10.
50. WEEKS JC, COOK EF, O'DAY SJ *et al*.: Relationship between cancer patients' predictions of prognosis and their treatment preferences. *JAMA* 1998; 279: 1709-14.
51. DEGNER LF, KRISTJANSON LJ, BOWMAN D *et al*.: Information needs and decisional preferences in women with breast cancer. *JAMA* 1997; 277: 1485-92.
52. ALLANORE Y, KAHAN A: Treatment of systemic sclerosis. *Joint Bone Spine* 2006; 73: 363-8.

APPENDIX

This section illustrates the use of a simple prognostic calculator in Excel format, to predict 5-year survival (from diagnosis) in SSc patients. Follow the steps below to download and use the calculator

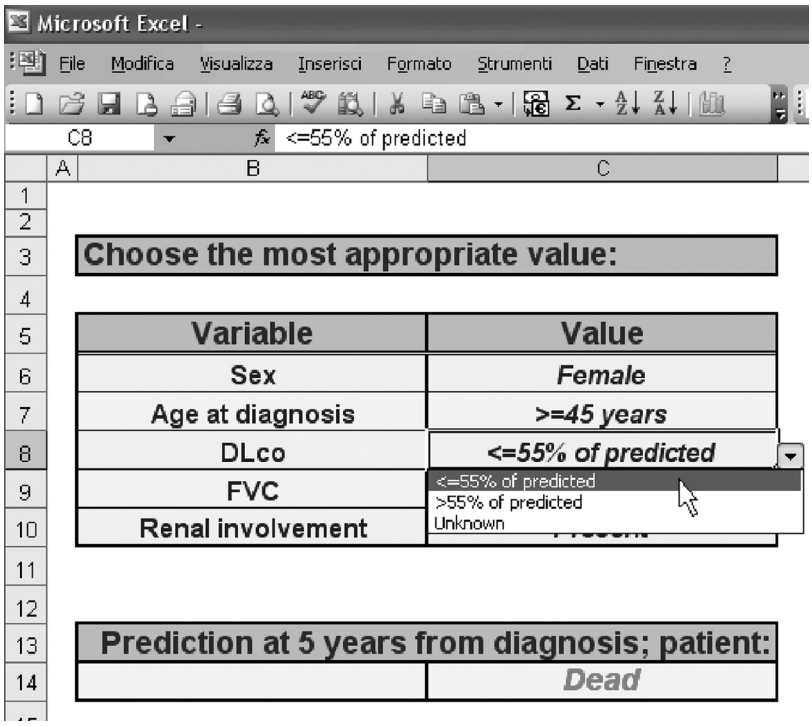
Step 1.
Download the calculator from the site:
http://www.nd.edu/~nchawla/survival.xls

Step 2.
Open the calculator, the following screenshot will appear:



There are three main features:
-The clinical characteristics of the patient to be used to make the prediction (topmost left)
-The prediction (mid left)
-The performance of the prediction as assessed in a population of 356 SSc patients recruited from France (60 cases), Spain (105 cases), the Netherlands (130 cases) and Italy (61 cases) (graph on the right and bottom left). The square on the graph indicates the optimal threshold as determined by the analysis conducted on 558 Italian subjects.

Step 3.
Choose the desired value from the scroll-down menu:



Step 4.
The estimated 5-year survival probability is displayed in Cells B14 or C14 and changes whenever a selection is made in Step 3.