A Comparison of Two Computer-Based Prognostic Systems for AIDS

Lucila Ohno-Machado, MD, MHA, and Mark A. Musen, MD, Ph.D.
Section on Medical Informatics, Stanford University School of Medicine
Stanford, CA 94305 – machado, musen @camis.stanford.edu

We compare the performances of a Cox model and a neural network model that are used as prognostic tools for a cohort of people living with AIDS. We modeled disease progression for patients who had AIDS (according to the 1993 CDC definition) in a cohort of 588 patients in California, using data from the ATHOS project. We divided the study population into 10 training and 10 test sets and evaluated the prognostic accuracy of a Cox proportional hazards model and of a neural network model by determining the number of predicted deaths, the sensitivities, specificities, positive predictive values, and negative predictive values for intervals of one year following the diagnosis of AIDS. For the Cox model, we further tested the agreement between a series of binary observations, representing death in one, two, and three years, and a set of estimates which define the probability of survival for those intervals. Both models were able to provide accurate numbers on how many patients were likely to die at each interval, and reasonable individualized estimates for the two- and three-year survival of a given patient, but failed to provide reliable predictions for the first year after diagnosis. There was no evidence that the Cox model performed better than did the neural network model or vice-versa, but the former method had the advantage of providing some insight on which variables were most influential for prognosis. Nevertheless, it is likely that the assumptions required by the Cox model may not be satisfied in all data sets, justifying the use of neural networks in certain cases.

PROGNOSTIC CLASSIFICATION

Quantitative tools that model disease progression and that provide predictions of survival are essential for the provision of care and the design of adequate health-care policies. The simplest models for survival analysis use actuarial life tables or Kaplan–Meier product-limit estimators. Parametric models of survival are also been used, where the restrictive assumption of a fixed distribution is traded for efficiency and facility of mathematical manipulation. These methods include Markov and accelerated failure time models.[1] A few authors used more complex nonparametric methods, such as classification trees and neural networks.[2] Traditional survival analysis tools, such as the Cox proportional hazards method[3], are often used in the exploratory phase of the development of staging systems. However, the adequacy of these methods for the development of prognostic classification of individual patients has not been effectively assessed. Most of these methods are aimed at *explaining the progression of the disease*, by selecting influential variables, rather than the *predicting survival* for populations or individual patients. Neural networks have been used successfully for building prognostic systems, and provide an alternative to the Cox model. The performances of both methods have not yet been experimentally compared.

An important requirement for a classifier is its ability to use as much information as it can on an individual case to make accurate predictions. In the domain of HIV infection, the Cox proportional hazards model is frequently used to study the importance of covariates for survival, but is seldom used to produce survival prognoses. The Cox model is a multiple logistic regression semi-parametric model that allows modeling of continuous covariates, and involves the assumptions that there is a simplifying transformation of the initial data and that the hazards for the different groups are proportional. The model assumes a baseline hazard and hazards for individuals with certain variable values are multiples of that baseline, as shown in

$$h_i(t) = h_0(t)e^{z_i\beta}$$

where $h_i(t)$ is the hazard for individual i at time t, $h_0(t)$ is the baseline hazard at t, z is the vector of explanatory variables and β is the vector of coefficients for each variable. In order to provide predictions of survival for individual patients, a baseline hazard that is common to all patients has to be estimated. This estimation represents no trivial task and the choice of the wrong baseline hazard can change the results of the predictions dramatically.

When the task is to establish predictions of survival for individual patients, neural networks constitute good alternatives for classical statistical methods. Neural networks used for this purpose are nonparametric models that allow modeling of a large number of non-linear functions.[4] Their advantage is that few assumptions have to be made about the distribution of the data. Their main disadvantage is that they provide little insight on which variables are most influential in the model. Escalating health-care costs related to the prevention and treatment of HIV infection call for policies that are derived from existing data using reliable quantitative tools.[5] In this paper, we show how we used data from an existing database of AIDS patients and used two mathematical models to determine prognosis of death due to AIDS.

Prognosis of AIDS

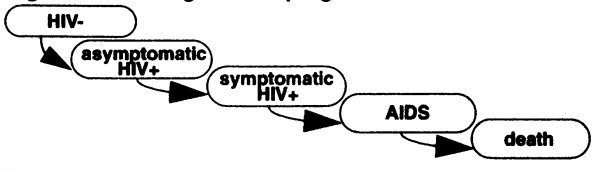
The economic effects of the AIDS epidemic have deeply influenced the development of health care policies and the practice of medicine. HIV infection is one of the most challenging public health problem that has arisen in the second half of the twentieth century. It is important to know the natural history of the disease, to predict the development of the disease in HIV+ individuals, to establish national policies to slow disease progression, and to establish guidelines for adequate medical intervention. Certain markers both may be easy to obtain in primary-care settings, and may provide valuable information for making prognostic distinctions.

Prognostic evaluations of patients who are HIV+ can help both patients and physicians to plan treatment and

0195-4210/95/$5.00

to allocate resources. Prognostic distinctions can help researchers to design clinical trials, by allowing them to stratify groups according to one measure of potential therapeutic benefits.[6] The classification of patients according to different disease-progression profiles also helps the policy maker to determine the nation's health-care needs and to assess the global influence of preventive and therapeutic interventions. Several authors have modeled transitions from seronegative to seropositive HIV, from asymptomatic HIV+ status to symptomatic HIV+ status, from HIV+ status to AIDS, and from AIDS to death, as shown in Figure 1. In this study, we have modeled the transition from AIDS to death.

Figure 1. Modeling disease progression in HIV infection.



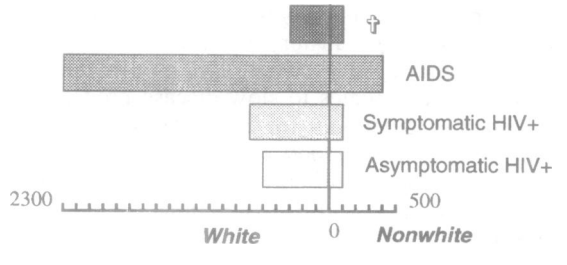
Mathematical models have been constructed to model every state transition in HIV infection. In this paper, we built survival models for the transition from AIDS to death.

MATERIAL AND METHODS

Data source: ATHOS database

The ATHOS database is a prospective, longitudinal, primary data set of HIV+ and at-risk subjects, collected from 10 clinics in California (3 private practices in the San Francisco Bay Area, 2 private practices in Los Angeles, and 5 community clinics associated with the Owen Clinic at UCSD).[7] The ATHOS database was built to provide a national HIV data resource that permits the systematic study of (1) disease costs and financing, (2) drug effectiveness, toxicity, and cost, (3) delivery systems and practice variations, (4) health status and quality of life, and (5) disease transitions and modeling. The data collection began in 1989 at Stanford University. Some authors have used the ATHOS database to assess the socioeconomic impact of the AIDS epidemic,[8] as well as to investigate medical issues related to HIV infection. Researchers of the ATHOS project have developed quality-control protocols that assure the reliability of ATHOS data. Approximately 50 percent of the patients have AIDS (1993 CDC definition[9]), 25 percent are HIV+ but do not have AIDS, and 25 percent are HIV-, but at risk for HIV infection. There were 290 deaths and 572 diagnoses of AIDS through mid-1993. Figure 2 shows the distribution of patient cases according to clinical stage and ethnicity.

Figure 2. Approximate distribution of ATHOS patients according to clinical stage and ethnicity.



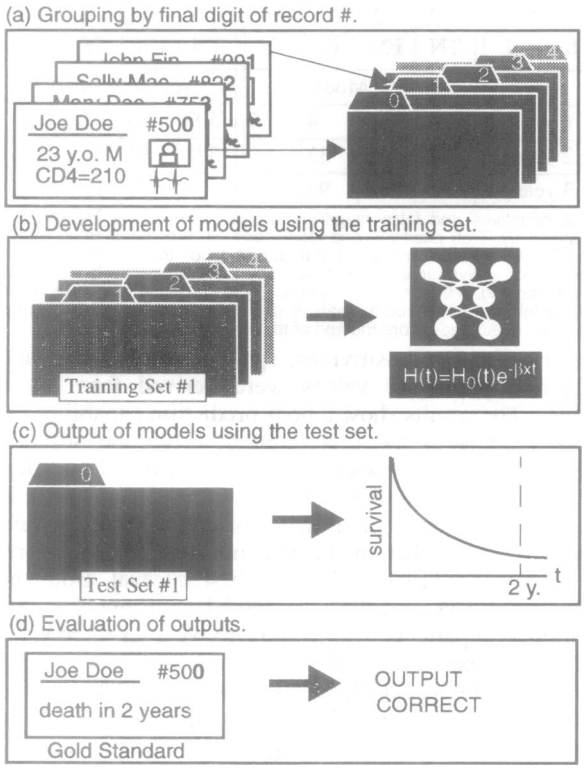
Data from 1056 patients who had AIDS were available for this study and over 700 variables were represented. The variables included diagnoses, signs and symptoms, results of laboratory tests, medications, and variables to assess functional status, quality of life, insurance coverage, medical resources utilization, side effects, and multiple health outcomes. Data were collected in three-month intervals, by means of questionnaires and summarization of clinical progress notes.

Statistical Methods

Specification of covariates and outcome. The major endpoint in this analysis was mortality due to AIDS-related conditions from the date of diagnosis by the 1993 CDC definition. Only 588 records were selected for this study because they had the necessary information on the date of AIDS diagnosis by the CDC definition. Demographic and socioeconomic explanatory variables included age, gender, race, risk group, AIDS-defining diagnoses, hospitalizations, educational level, and average income. Clinical findings included fatigue, weight loss, diarrhea, mental status, and Karnofsky scores. Laboratory test results included CD4 counts, CD4/CD8 ratio, hemoglobin, erythrocyte sedimentation rate, erythrocyte and platelet counts, white blood cell counts, serum p24 antigen, serum β-2 microglobulin, total cholesterol, HDL, and albumin levels. Variables indicating antiretroviral and prophylactic medications for opportunistic infections were used. Continuous variables were represented as standard deviations from their means. Dummy coding was used for categorical variables, with "0" indicating absence and "1" indicating presence of a certain characteristic. In cases where there were missing data, values representing either the mean or the mode were entered.

Data Set Construction. Because the number of observations was not very large, we used the *leave-n-out* technique to assess the predictive ability of our model. To do so, we divided the whole set of patients into 10 groups, by systematically classifying them according to the last digit of the database record number. Therefore, all patients with record numbers containing "0" as the last digit were grouped in the first set (Group 0), all patients with record numbers containing "1" as the last digit were grouped in the second set (Group 1), and so on. The training set was composed of those records left behind after each group was selected. For example, when Group 0 was selected as the test set, all other records containing last digits different from "0" would compose the training set and would be used to build the model. This process was repeated 10 times, so that every group was used once as a test set, and all patients were tested. We consequently had 10 different training sets, and needed 10 different models for each method (i.e., Cox Proportional Hazards, and Neural Network), as shown in a simplified form in Figure 3. For the Cox model, a survival curve for each patient in the test set was created. Using an arbitrary cutoff value (0.5), we considered a patient dead at a given interval if the probability of survival fell below the threshold. The "gold standard" was the information about survival contained in the actual records.

Figure 3. Training and test sets.



Cox Model. We built a Cox proportional hazards model using the PHREG procedure in SAS.[10] Variable selection was done using a stepwise process, in which the significance level of both entry and permanence of a given variable in the model was 0.05. Ties were handled by the Breslow method. Estimation of the survivor functions at event times were performed using the baseline statement in SAS. Ten models were constructed using the Cox model to evaluate the ten test sets. A full model, containing all cases, was also constructed in order to determine whether the combined results of the ten models were compatible with the whole data set with respect to the variables selected, their coefficients, and relative risks. The full model was not used for evaluation.

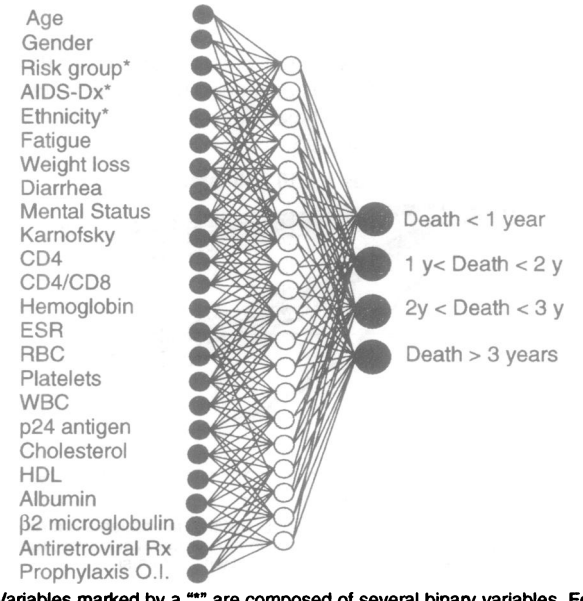
Neural Network. The neural network used exactly the same sets as the Cox model, except that the output was categorical: Instead of providing a duration of survival, represented by a real number, the output was composed of four binary variables indicating death at a given interval, as shown in Figure 4.

The PDP software[4] was used to develop the neural network model. Initial weights were random. The learning rate was 0.01. The network was trained by backpropagation using a cross-entropy error function for each output node, represented by

$$E = -\sum_i [Y_i \log p_i + (1 - Y_i) \log (1 - p_i)]$$

where Y_i is the target, or desired output, and p_i is the actual output produced by the network for each individual i. Minimizing this function is the same as performing maximum likelihood estimation.[11] The outputs of this network can be interpreted as a crude estimate of the probability that a given patients *dies* in a given interval (and not the probability that a patient *is dead* at this interval, which is the number that is easily derived from a survival curve, and therefore is the number that we are ultimately seeking). The probability that the patient *is dead* at a certain interval would then be the probability that the patient died at that given interval, plus the probability that the patient died in the interval that preceded the one in question. For example, if the network produced probabilities of 0.1, 0.3, 0.2, 0.4 that a patient would die in the first, second, third and fourth interval, respectively, then the actual probabilities of that patient being dead in those intervals would be 0.1, 0.4, 0.6, and 1, representing a monotonically increasing function. We analyzed the results for the first three intervals. The networks were trained until the error in the test set started to increase.

Figure 4. Neural network architecture.



Variables marked by a "*" are composed of several binary variables. For example, *Risk Group* is composed of *Gay/Bisexual, IVDU, Heterosexual,* and *Transfusion recipient. AIDS-Dx* is composed of *PCP pneumonia, Kaposi sarcoma,* and so on.

Evaluation. The methods were compared for total accuracy, sensitivities, specificities, and positive and negative predictive values for each interval. All results reflect a significance level of 0.05 for the combined results of all ten test sets, representing 588 patients. The Cox model was also assessed by a χ^2 test, described below.

RESULTS

Explanatory Variables. The stepwise procedure in the Cox model selects variables according to their χ^2 statistics, by allowing them to enter the model if their p-values are below a certain threshold. The variables selected by the stepwise procedure are shown in Table 1, were β represents the mean coefficient obtained from the ten models, and β' represents the coefficient for the full model. The first variable to enter the model was *Length of Stay in Hospital*, or *LOS*, representing the previous hospitalizations of a patient. This variable was entered in all models, together with *PCP Pneumonia*, and *Other*

AIDS-defining Diagnoses. *CD4 count* was entered in nine of the ten models, and *p24 antigen* in eight. The variable that indicated Hispanic ethnicity was entered in six models, and was not entered in the full model.

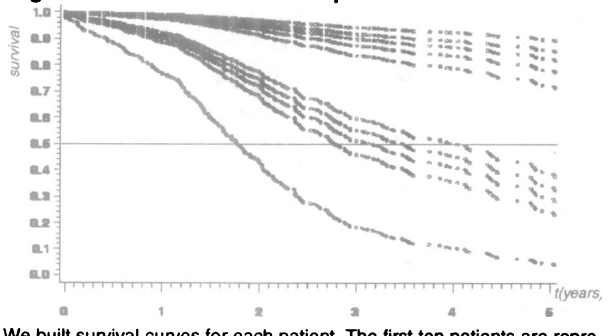
Table 1. Variables selected by the Cox model.

Variables	β	β'	RR	RR'
PCP pneumonia	1.12	1.32	3.08	3.73
Kaposi sarcoma	0.74	0.86	2.09	2.36
Other AIDS-defin. Dx	0.90	1.06	2.47	2.89
LOS	0.18	0.18	1.19	1.19
CD4 count	-0.33	-0.35	0.66	0.70
p24	0.36	0.55	1.44	1.73

The average coefficient for each selected variable of the ten models is shown under β. The coefficients for the full model are shown under β'. The average relative risk for each variable of the ten models in shown under *RR*. The relative risks of the full model are shown under *RR'*.

Missing data for laboratory values of *CD4 count* and *p24 antigen titers* in a three month interval around the diagnosis of AIDS were present in 200 and 483 cases, respectively. We built survival curves for each patient using the survivor estimates produced using the baseline hazard. An example of these curves for the first ten patients is shown in Figure 5.

Figure 5. Survival curves for 10 patients.



We built survival curves for each patient. The first ten patients are represented in this example. Three groups with similar survival curves, representing similar prognoses, were found in the whole data set.

Performance on a population. In the assessment of disease progression and health-care needs for the whole population, it is sufficient that the prognostic tool provides prediction on *how many* cases are likely to die at each interval, as opposed to *which* specific patients are likely to die. To assess the performance of the systems on the average of patients, rather than on each individual patient, it is enough to use a statistic that compares the true number of deceased in each interval and the estimate provided by each model. This statistic is simply the difference between the true number of deaths ΣY and the sum of probabilities derived from each system Σp, divided by the square root of the variance $p(1-p)$, and has an approximately normal distribution. For all intervals and both systems, the differences were not statistically significant (for $\alpha = 0.05$), showing that both methods were able to provide accurate numbers on the average population.

Performance on individual cases. We obtained probabilities of death at the end of a given interval and built a classification table for each interval and each method. The results are shown in Table 2.

Table 2. Performance of both methods on individual cases.

Interval	TN	FN	TP	FP	TN	FN	TP	FP
	Cox Model				Neural Network			
< 1 year	519	60	4	5	507	56	8	17
< 2 years	422	118	32	16	388	93	57	50
< 3 years	307	135	94	52	287	99	130	72

True negatives and false negatives are represented by *TN* and *FN*, respectively. True positives and false positives are represented by *TP* and *FP*. A prediction was counted under *TN* when the system predicted survival for a given interval, and the patient indeed survived the period corresponding to that interval. A prediction was counted under *TP* when the system predicted that the patient would be dead in a given interval, and he indeed died before the end of that interval, and so on.

Total accuracy, sensitivities, specificities, positive and negative predictive values were derived from these tables. The results show a poor prediction capability for both models in the first interval. From the survival curves of the Cox model, it is easy to see that the prediction of prognosis in the first year is indeed more difficult than in the second or third years (the curves are very close to each other in the first interval). Furthermore, few patients (11%) died in the first interval. Both the neural network and the Cox model were good at discriminating patients who would survive after three years of disease (percentage of correct cases of 0.68 and 0.70, PPVs of 0.69 and 0.74, NPVs of 0.64 and 0.64, for the Cox and Neural Network models, respectively). For the interval of two-years after diagnosis, the performance of both systems was also good (percentage of correct cases of 0.77 and 0.76, PPVs of 0.66 and 0.53 and NPVs of 0.78 and 0.80, for the Cox and Neural Network models, respectively).

Assessment of Probabilities. Since the Cox model produces true probabilities of survival for each interval, we could also assess its predictive ability on individual cases by using a technique provided by that author.[12] Let p represent the probability of survival in a given interval, as predicted by the model. We simply form k subgroups with nearly constant p and compare the proportion of successes in each subgroup with the true number of successes in that subgroup, by using a χ^2 test, with k-1 degrees of freedom. We obtained χ^2s of 13.49, 13.85, and 16.97, indicating that the predictions obtained were compatible with the data for all intervals.

DISCUSSION

Survival analysis methods are seldom used to establish prognostic classification of individuals or of groups of patients. Using a Cox model and an estimate of a baseline hazard, we were able to provide reliable predictions for the number of patients who were likely to die at intervals of one year after diagnosis of AIDS. The Cox model was also used to explain which variables were most important for the prognosis of AIDS. The ten Cox models obtained by the *leave-n-out* procedure seemed to agree with the full model regarding which variables are most important for the prognosis of AIDS and in which order they entered the model. Especially for the most influential variables, such as *LOS* and *PCP pneumonia*, the values for the coefficients did not differ by a large amount. The abundance of missing values for laboratory

markers may have determined the absence of other variables in the model. The survival curves provided by the Cox model seemed to define three distinct prognostic groups. It is still necessary to test these models in different samples of the population to determine whether the results are generalizable.

Both the Cox model and the neural network methods produced good estimates on the average number of individuals who were likely to die at each interval. However, the predictions of both methods for the prognosis of a particular individual were only reliable for the intervals of two and three years. The number of cases in this study was not small, but we believe that the significant amount of missing data contributed to the somewhat poor predictions provided by the systems for the first interval, where the number of examples was smaller. A more complete data set may be able to answer this question. An additional problem with the Cox model may have been the departure from the proportionality assumption. Hanson et al.[13] described a cohort of HIV+ patients for whom the assumption of hazard proportionality did not hold. For these cases, fully nonparametric models of survival that allow non-proportional hazards are needed, and neural networks may indeed perform better. We plan to stratify the data according to the variables selected by the Cox model and test the proportionality assumption. Another approach for establishing prognosis is to use the pooled-logistic regression method.[14]

A problem with the neural network was that it was presented with less information than was the Cox model, due to the transformation of the output into 4 discrete categories. Nevertheless, its sensitivities and specificities were not too different from those of the Cox model. Both methods presented very low figures for sensitivities, especially for the first interval. By using a hierarchical system of neural networks,[15,16] it may be possible to enhance the sensitivity to that pattern without decreasing specificity. Since each method has its own advantages and disadvantages, it seems natural to try to combine them to build a more accurate prognostic tool, so that, for example, we use the Cox model to select the variables and then apply a neural network that uses only those variables deemed significant by the Cox model to provide prognostic distinctions.

CONCLUSION

Current survival analysis methods are rarely used to provide prognostic predictions of survival for individual patients. New methods, such as neural networks, can constitute good alternatives for the development of prognostic systems, especially when certain assumptions about the distribution of the data cannot be verified. We developed two computer-based systems, one based on the Cox proportional hazards model and another based on neural networks to do the prognosis of people living with AIDS. We showed that the differences in performances of both methods for predictions for the whole population and predictions for individual cases were not significant, and concluded that both methods can be reliably used for prognostic classification.

Acknowledgments

We thank David Rumelhart, Michael Walker, Byron Brown, Daniel Bloch, Edward Shortliffe, Alan Garber, and Liping Wei for useful discussion in different aspects of the present work. This work has been funded by a grant from the Agency for Health Care and Policy (AHCPR) and by the Conselho Nacional de Pesquisa (CNPq), Brazilian Ministry of Education. Computing facilities were provided by CAMIS, through grant LM05305 from the National Library of Medicine.

Reference

[1] Collet D. *Modelling Survival Data in Medical Research.* Chapman and Hall, London, 1994.

[2] Ravdin PM; Clark GM. A practical application of neural network analysis for predicting outcome of individual breast cancer patients. *Breast Cancer Research and Treatment*, 1992, 22(3):285–93.

[3] Cox DR. *Analysis of Survival Data.* London, Chapman and Hall, 1984.

[4] McClelland J; Rumelhart DE. Explorations in Parallel Distributed Processing. MIT Press, 1989.

[5] Seage GR 3d; Landers S; Lamb GA; Epstein AM. Effect of changing patterns of care and duration of survival on the cost of treating the acquired immunodeficiency syndrome (AIDS). *American Journal of Public Health*, 1990 Jul, 80(7):835–9.

[6] Justice AC; Feinstein AR; Wells CK. A new prognostic staging system for the acquired immunodeficiency syndrome. *NEJM*, 1989 May 25, 320(21):1388–93.

[7] Fries J. *An HIV Data Resource for Systematic Health Policy Study.* Grant Proposal for the AHCPR, 1992.

[8] Lubeck DP; Fries JF. Health status among persons infected with human immunodeficiency virus. *Medical Care*, 1993 Mar, 31(3):269–76.

[9] CDC. 1993 revised classification system for HIV infection and expanded surveillance case definition for AIDS among adolescents and adults. *MMWR* 1992 Dec 18, 41(RR-17):1–19.

[10] SAS Institute. *Technical Report P-229, SAS/STAT software.* Release 6.07. Carey, 1992.

[11] Curry B; Rumelhart DE. MSnet: A neural network that classifies mass spectra. *Tetrahedron Computer Methodology*, 1990, 3:213–37.

[12] Cox, DR. *The Analysis of Binary Data.* Chapman and Hall, London, 1970.

[13] Hanson DL; Horsburgh CR Jr; Fann SA; Havlik JA; Thompson SE 3d. Survival prognosis of HIV-infected patients. *Journal of AIDS*, 1993 Jun, 6(6):624–9.

[14] Ingram DD; Kleinman JC. Empirical comparisons of proportional hazards and logistic regression models. *Statistics in Medicine*, 1989 May, 8(5):525–38.

[15] Ohno-Machado L. Identification of Low Frequency Patterns in Backpropagation Neural Networks. *JAMIA Symposium Supplement,* 1994.

[16] Ohno-Machado L; Walker MG; Musen MA. Hierarchical Neural Networks for Survival Analysis. Proceedings of the *MEDINFO*, 1995.