Validating a Cox Proportional-Hazards Model

Tristan Grogan MS, David Elashoff PhD

Department of Medicine Statistics Core, David Geffen School of Medicine at UCLA

Abstract

While several methods have been published on validating standard logistic or linear models, much less material exists on validating time-to-event models, such as the Cox proportional-hazards model. During the course of this research, an investigation of four different strategies for validating the Cox model was carried out, utilizing data from a prostate cancer study. These validation techniques are especially important for biomarker studies to aid in combating the effect of selection bias, and would strengthen the results and credibility of any study.

This paper will present four performance measures for assessing whether a model has evidence for being validated or not—comparing model coefficients between training and test data sets, assessing Harrell's c-index between the training and test models, running a cross-validation technique, and comparing recurrence risk predictions between the training and test models.

Key Words: Cox proportional-hazards, Validation, Survival data analysis, biomarkers

1. Introduction

In biomarker studies, often times, several potential markers are pre-screened with statistical analyses, which can drastically increase the chances of making a type I error (false positive). One way to address this issue is to restrict the familywise error rate by implementing a technique such as the Bonferroni correction. However, techniques such as Bonferroni can be too conservative (especially with a small sample size) and important markers may slide under the statistical detection threshold. Another way to address this quagmire is through model validation techniques; which set aside a portion of the data (testing set) to be reserved for assessing the multivariable prognostic model (training model). Having ways to externally validate the final model would help protect against over optimistic p-values and, at times, erroneous conclusions.

Tumor-associated macrophages (TAMs) have been associated with worse pathological characteristics and prognosis in several cancers such as colon, breast, endometrial, Non-Hodgkin's lymphoma, and bladder.[123456] However, other reports suggest that increased TAMs are associated with improved prognosis whereas others report no importance.[78] The significance of TAMs in men with prostate cancer has not been studied comprehensively, and the studies that do exist are inconsistent.[91011]

A specific tumor-associated macrophage, CD68, has been thought to be associated with prostate cancer biochemical recurrence after radical prostatectomy. Biochemical recurrence was defined as a single PSA >0.2, 2 values at 0.2, or secondary treatment for an elevated post-operative PSA. Patients were monitored for biochemical recurrence with a median follow-up time of 5 years post prostatectomy.

Our cohort consisted of 330 patients who underwent radical prostatectomy. A 2/3 to 1/3 training and test set was determined by randomly splitting the full dataset while stratifying by the outcome (recurrence) variable, ensuring that each data set had a similar baseline hazard estimate. This allowed for an easier comparison of model coefficients between the two models.

Each patient had multiple cores sampled from their prostate (See Figure 1); which were categorized as normal, pre-cancerous, or cancerous. The number of cells with CD68 staining was assessed for each core and various summary measures for each type of core and patient were computed—the maximum CD68 count across all core samples, difference in CD68 count from cancer to normal cores, mean/median CD68 count for each type of core, and max CD68 count for each core.

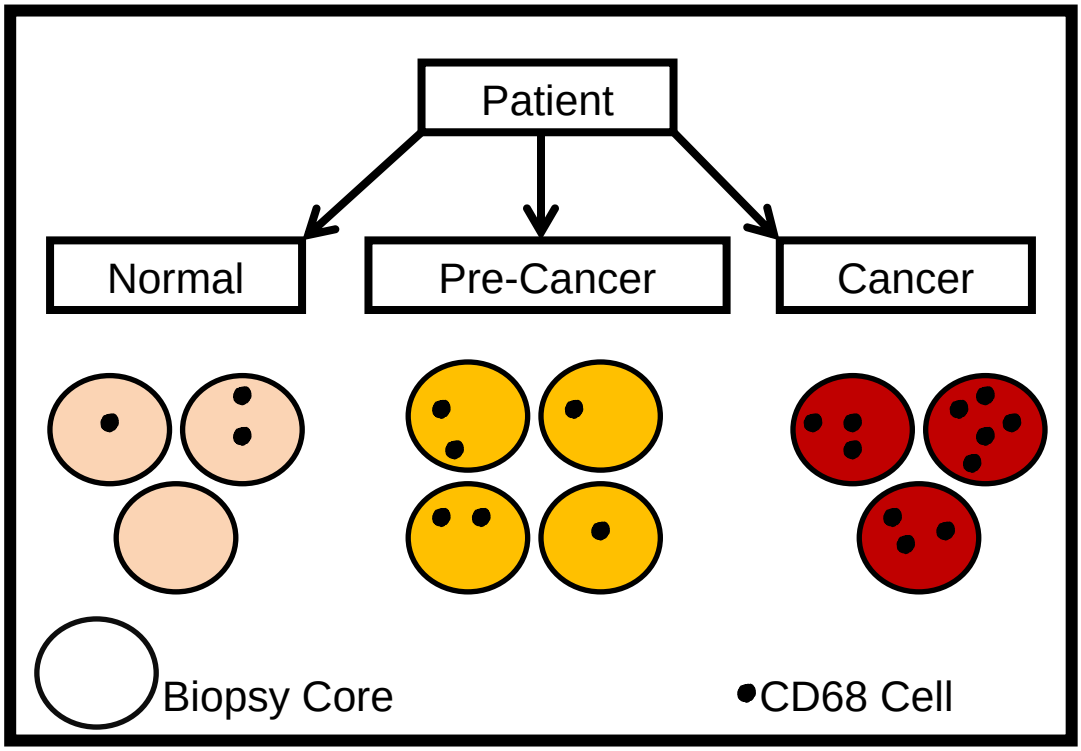

Figure 1: Graphical description of biopsy cores and CD68 count

2. Methods

Univariate Cox proportional-hazards models predicting biochemical recurrence were constructed for each formulation of the CD68 marker (Table 1).

Table 1: Univariate CD68 screening Cox PH models on the training data set

CD68 Formulation	P-Value	Hazard Ratio (95% CI)
mean nl cores	0.879	0.99 (0.91, 1.09)
mean pin cores	0.556	1.03 (0.94, 1.11)
mean ca cores	0.043	1.07 (1.00, 1.14)
median nl cores	0.678	0.98 (0.90, 1.07)
median pin cores	0.551	1.03 (0.94, 1.11)
median ca cores	0.125	1.05 (0.99, 1.12)
max nl cores	0.862	1.01 (0.94, 1.07)
max pin cores	0.909	1.00 (0.95, 1.05)
max ca cores	0.005	1.05 (1.01, 1.09)
max all cores	0.022	1.03 (1.01, 1.07)
avg all cores	0.169	1.06 (0.98, 1.15)
avg diff nl-pin cores	0.886	1.01 (0.92, 1.10)
avg diff nl-ca cores	0.031	0.93 (0.87, 0.99)
avg diff pin-ca cores	0.168	0.95 (0.89, 1.02)

Out of the 14 formulations of the CD68 variable, four were associated ($p<0.05$) with biochemical recurrence from the univariate Cox models (Table 1). The mean CD68 count from the cancer cores (within each patient) was decided as the formulation to represent CD68 quantity over the other three significant ($p<0.05$) formulations, because it had the highest effect size estimate (HR = 1.07) and made more clinical sense than the others.

2.1 Coefficient Comparison

This strategy involves running two independent models on both the training and test data sets, allowing for a comparison of the magnitude and direction of the coefficients to see if they are similar. Comparing p-values is also possible, although not the main focus of this method. If coefficient estimates are widely different between the training and test models, then either the covariates may not have been balanced in each data set, or perhaps the estimates from the training model aren't very reliable.

2.2 Harrell's Concordance Index

The Harrell's c-index is a metric to compare the strength of predictive performance for survival models. Moreover, the index has similar interpretation to the AUC for logistic models (and is a rank based statistic).

Model based risk scores are calculated for each subject. Then, for each eligible pair of patients, the algorithm counts how often the patient with the higher risk score experienced an event before the patient with the lower risk score experienced their event. Pairs of subject that either are both censored or have one subject censored prior to a subject with an event are omitted from the calculation. The other two scenarios; both two

patients have events, and a patient who was censored later than a patient who had an event are included in the calculation. For example, a patient may have had a recurrence at 1 year and another patient was censored after 2 years of follow-up. It could be determined that if the model assigned a higher risk score to the patient who recurred, that would contribute to a concordant case. The resulting proportion of all possible pairs is the probability of concordance[12]. Comparing the percentage of concordant points in the test data by applying the test model and training model will give an idea for how similar the two models are classifying the patients.

2.3 Cross-validation

Cross-validation was performed on the entire data set, utilizing the penalized package in R (optL1 function). Instead of having one training data set and one test data set as before, the cross-validation technique creates several partitions of the full data set. The specific cross-validation technique applied to this data set implemented the lasso penalty using 10-fold likelihood. The package deliberately does not provide standard errors for p-values/confidence intervals since they are not meaningful for biased estimates[13]. However, by comparing these cross-validated coefficients, a sense for the reliability and stability of coefficient estimates from the training model can be attained.

2.4 Recurrence Risk Predictions

This method allows a comparison of the survival predictions on the test data set using both the training and test set models. The goal is to show that predicting recurrence using the training model is no different than using the test model. Tertile cut-offs of the linear fitted value are created by the training set model on the training set data. Then, each patient in the test data set will have a linear predictor score created from the test and training models. Thereafter, the data is split into tertiles based on the cut-off values from the training data. Kaplan-Meier curves are created and the log-rank test can be used to assess differences between the two curves[14].

3. Results

Table 2: Descriptive statistics by training and test datasets

Variable	Training Data (n=219)	Test Data (n=110)
	Mean (SD)	**Mean (SD)**
Follow-up time (months)	73.5 (53.7)	71.8 (56.0)
CD68 Mean Cancer cores	6.5 (3.4)	6.8 (3.9)
Pathological Gleason Sum	6.3 (0.96)	6.3 (1.0)
Pre-operative serum PSA (ng/ml)	10.3 (7.7)	9.6 (6.2)
	Frequency	**Frequency**
Biochemical Recurrence (PSA>0.2)	86 (39.3%)	43 (39.1%)
Extracapsular Extension	19 (8.7)%	16 (14.5)%

3.1 Coefficient Comparison

Table 3: Training and test dataset models

Training Data Set	**HR (95% CI)**	**P-Val**
CD68 Mean CA	1.06 (0.99, 1.13)	0.082
Path. Gleason Sum	1.73 (1.37, 2.19)	<0.001
Extracapsular E	2.99 (1.62, 5.54)	<0.001
PSA Level	1.03 (1.00, 1.05)	0.036
Test Data Set	**HR (95% CI)**	**P-Val**
CD68 Mean CA	1.00 (0.92, 1.09)	0.974
Path. Gleason Sum	1.63 (1.15, 2.29)	0.005
Extracapsular E	2.69 (1.30, 5.57)	0.008
PSA Level	1.05 (1.01, 1.10)	0.021
Full Set (Training + Test)	**HR (95% CI)**	**P-Val**
CD68 Mean CA	1.03 (0.98, 1.09)	0.219
Path. Gleason Sum	1.70 (1.41, 2.05)	<0.001
Extracapsular E	2.77 (1.74, 4.39)	<0.001
PSA Level	1.03 (1.01, 1.06)	0.002

3.2 Harrell's Concordance-index

Model	**Training Data Concordance (standard error)**	**Test Data Concordance (standard error)**
Training Model	0.708 (0.03)	0.702 (0.024)
Test Model	0.681 (0.029)	0.728 (0.047)

3.3 Cross-validation

Variable	**Hazard Ratios Cross-validation 10 fold**	**Training model HR (95% CI)**	**P-Val**
CD68 Mean CA	1.04	1.06 (0.99, 1.13)	0.082
Path. Gleason Sum	1.86	1.73 (1.37, 2.19)	<0.001
Extracapsular E	3.57	2.99 (1.62, 5.54)	<0.001
PSA Level	1.04	1.03 (1.00, 1.05)	0.036

3.4 Recurrence Risk Predictions

PSA, Pathological Gleason Sum, and the percent of patients with extracapsular extension are higher in tertile 3 (Figure 2). This trend is not seen in the CD68 variable. The curves are very similar for each plot indicating that the training and test models are not predicting recurrence substantially different utilizing the test data for patients with varying tertiles of risk.

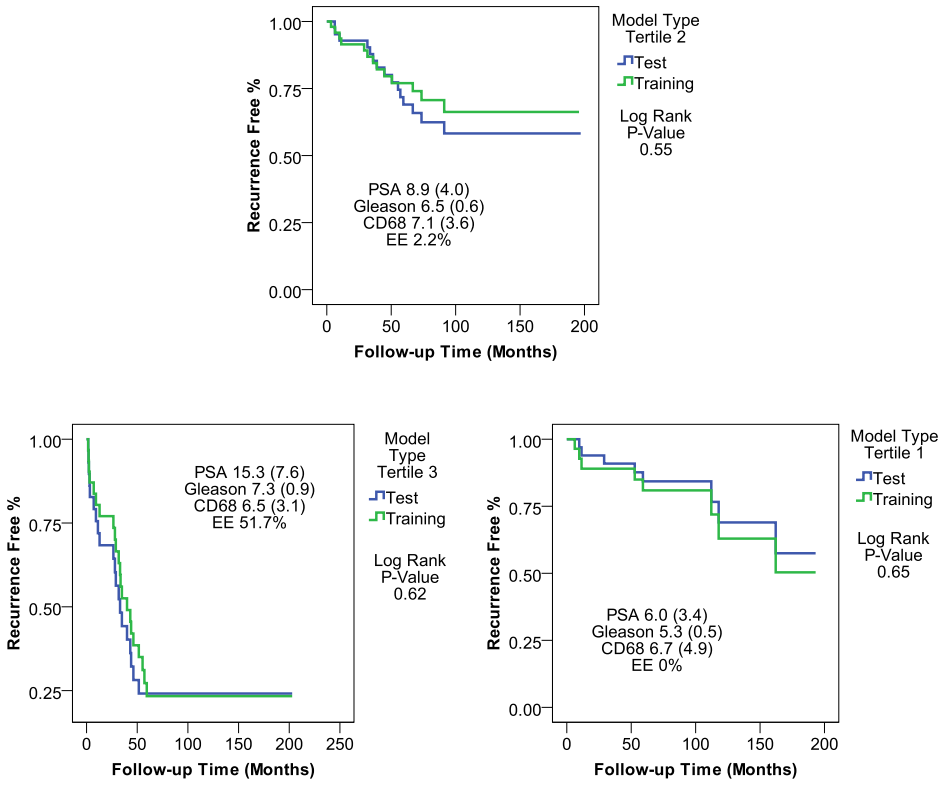

Figure 2: Recurrence risk predictions

4. Findings

In conclusion, all four of the methods tested gave fairly consistent results in terms of validating the overall training model and the results were as follows. The training and test model coefficients are consistent in terms of magnitude and direction, with the exception of the CD68 variable (Table 2). There was no noticeable drop-off in Harrell's c-index when the training model was applied to the test data set (Table 4). The cross-validation technique on the full data set produced similar coefficients to the training set model (Table 3). By using the tertile split method we saw that both models seemed to predict recurrence similarly on the test data set (Figure 2).

In the training set model, the CD68 variable was marginally significant (p=0.08); however, it does not appear to have any effect in the testing cohort and drops out of the full multivariate model. The clinical result here is that the CD68 variable doesn't seem to be a very strong predictor of recurrence after we account for the other clinically relevant covariates. This could be an indication that the original apparent utility of the CD68 variable for recurrence prediction was the result of selection bias due to the multiple formulations of the variable that were considered in the training set.

This research has shown four relatively easy to implement techniques for validating a Cox proportional-hazards model. The benefit of using these validation techniques will help deter overfitting and increase credibility of your final model. Although these

techniques were applied internally to the same data set, they are justifiable to use for an external data set validation as well.

Acknowledgements

This research is supported by the National Center for Advancing Translational Sciences, Grant UL1TR000124. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

[1] Leek RD, Lewis CE, Whitehouse R, Greenall M, Clarke J, Harris AL. Association of macrophage infiltration with angiogenesis and prognosis in invasive breast carcinoma. Cancer Res. 1996;56(20):4625-9. Epub 1996/10/15.

[2] Bingle L, Brown NJ, Lewis CE. The role of tumour-associated macrophages in tumour progression: implications for new anticancer therapies. J Pathol. 2002;196(3):254-65. Epub 2002/02/22.

[3] Bacman D, Merkel S, Croner R, Papadopoulos T, Brueckl W, Dimmler A. TGF-beta receptor 2 downregulation in tumour-associated stroma worsens prognosis and high-grade tumours show more tumour-associated macrophages and lower TGF-beta1 expression in colon carcinoma: a retrospective study. BMC Cancer. 2007;7:156. Epub 2007/08/19.

[4] Salvesen HB, Akslen LA. Significance of tumour-associated macrophages, vascular endothelial growth factor and thrombospondin-1 expression for tumour angiogenesis and prognosis in endometrial carcinomas. Int J Cancer. 1999;84(5):538-43. Epub 1999/09/30.

[5] Takayama H, Nishimura K, Tsujimura A, et al. Increased infiltration of tumor associated macrophages is associated with poor prognosis of bladder carcinoma in situ after intravesical bacillus Calmette-Guerin instillation. J Urol. 2009;181(4):1894-900. Epub 2009/02/25.

[6] Canioni D, Salles G, Mounier N, et al. High numbers of tumor-associated macrophages have an adverse prognostic value that can be circumvented by rituximab in patients with follicular lymphoma enrolled onto the GELA-GOELAMS FL-2000 trial. J Clin Oncol. 2008;26(3):440-6. Epub 2007/12/19.

[7] Zhou Q, Peng RQ, Wu XJ, et al. The density of macrophages in the invasive front is inversely correlated to liver metastasis in colon cancer. J Transl Med. 2010;8:13. Epub 2010/02/10.

[8] Mahmoud SM, Lee AH, Paish EC, Macmillan RD, Ellis IO, Green AR. Tumour-infiltrating macrophages and clinical outcome in breast cancer. J Clin Pathol. 2012;65(2):159-63. Epub 2011/11/04.

[9] Lissbrant IF, Stattin P, Wikstrom P, Damber JE, Egevad L, Bergh A. Tumor associated macrophages in human prostate cancer: relation to clinicopathological variables and survival. Int J Oncol. 2000;17(3):445-51. Epub 2000/08/12.

[10] Nonomura N, Takayama H, Nakayama M, et al. Infiltration of tumour-associated macrophages in prostate biopsy specimens is predictive of disease progression after hormonal therapy for prostate cancer. BJU Int. 2011;107(12):1918-22. Epub 2010/11/04.

[11] Shimura S, Yang G, Ebara S, Wheeler TM, Frolov A, Thompson TC. Reduced infiltration of tumor-associated macrophages in human prostate cancer: association with cancer progression. Cancer Res. 2000;60(20):5857-61. Epub 2000/11/04.

[12] Harrell, Jr. F. (2001). Regression Modeling Strategies. New York: Springer.

[13] Goeman J, Meijer R, Chaturvedi N. Penalized: L1 and L2 penalized estimation in GLMs and in the Cox model (Version 0.9-41) [R Version 2.15.0]. Leiden, Netherlands: Leiden University Medical Center. Retrieved June 26, 2012. Available from: http://cran.r-project.org/web/packages/penalized/index.html.

[14] Royston P, Altman DG (2011) External validation of a Cox prognostic model: principles and methods. Statistics in Medicine, submitted.