Quantifying predictive accuracy in survival models

Seth T. Lirette, MS,[a,b] and Inmaculada Aban, PhD[b]

[a] Center of Biostatistics and Bioinformatics, University of Mississippi Medical Center, Jackson, MS
[b] Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL

Received Sep 19, 2015; accepted Sep 19, 2015
doi:10.1007/s12350-015-0296-z

For time-to-event outcomes in medical research, survival models are the most appropriate to use. Unlike logistic regression models, quantifying the predictive accuracy of these models is not a trivial task. We present the classes of concordance (C) statistics and R^2 statistics often used to assess the predictive ability of these models. The discussion focuses on Harrell's C, Kent and O'Quigley's R^2, and Royston and Sauerbrei's R^2. We present similarities and differences between the statistics, discuss the software options from the most widely used statistical analysis packages, and give a practical example using the Worcester Heart Attack Study dataset. (J Nucl Cardiol 2017;24:1998–2003.)

Key Words: Survival analysis • c statistic • R^2 statistics • discrimination • predictive accuracy

INTRODUCTION

Often there are two goals when building a statistical model: (1) describing associations between a set of predictors and an outcome and (2) building a model that predicts future outcomes with high accuracy. The focus of this review is the latter. Related to this, we can present ways to assess the validity of adding predictors to a set of established prognostically important variables. For example, would adding high-sensitivity C-reactive protein, fibrinogen, or left ventricular hypertrophy to model already containing age, blood pressure, and smoking status provide a substantial increase in accuracy of a predictive model for heart disease?

There are two distinct ways of evaluating predictive accuracy in a survival model. The first involves an extension of the concordance (C) statistics used for binary outcome. This class of statistics seeks to quantify the ability of the model to correctly classify subjects into one of two categories (event and non-events). This process is known as discrimination. The second method extends the idea of the coefficient of determination (R^2) in linear regression which quantifies the amount of variability in the response explained by the model to the survival context. Both of these ideas progress, perhaps intuitively, into survival data because we can think of survival data as an interplay between a binary outcome (event vs non-event) and a continuous outcome (time). We will explore the class of C statistics and R^2 statistics, provide comparisons between the two and introduce some caveats into predictive model building. We also discuss software options and illustrate a working example of evaluating predictive accuracy using these statistics.

CLASS OF C STATISTICS

Concordance, using the common English definition, is agreement amongst ideas. This definition is transferrable to the statistical definition: Do the model and the actual observed values agree that an observation is an event or non-event? The values of C statistics can range from 0.5 to 1 with 0.5 indicating that the predictor does not perform better than random chance and 1 signaling that the predictive model has classified all events and non-events correctly. Pecina et al[1] describe four of the most widely used and studied of the class of C statistics: (1) Chambless and Diao's C (2) Harrell's C (3) Uno's C and (4) Gonen and Heller's k. The simplest extension from the binary case to survival is Chambless

Reprint requests: Seth T. Lirette, MS, Center of Biostatistics and Bioinformatics, University of Mississippi Medical Center, 2500 North State Street, Jackson, MS, 39216; *slirette2@umc.edu*

1071-3581/$34.00

and Diao's C[2] (C_{CD}). Stated verbally, it is the average product of event probabilities multiplied by survival probabilities, standardized by the product of the average event and survival probabilities. Harrell's C[3] (C_H) incorporates actual survival times into the concordance statistic, whereas C_{CD} omits them. One major drawback of C_H is that it is highly dependent on the censoring mechanism. This is problematic in that it is usually regarded the censoring should not influence results. Uno's C[4] (C_U) seeks to alleviate the censoring concerns by incorporating the Kaplan-Meier estimator into the estimator. Gonen and Heller[5] propose a different concordance index (C_{GH}) that reverses the definition used in the previous three estimators, based on what the probability is conditioned. C_{GH} can be calculated with only the regression coefficients from a Cox proportional hazards model and risk factor levels. The differences between the two types of estimators are analogous to the differences between sensitivity and specificity (C_{CD}, C_H, and C_U) and positive and negative predictive value (C_{GH}).

Pecina et al[1] postulate that C_{GH} will always be the lowest, C_{CD} will always be the highest, and C_H & C_U will be roughly equivalent. Each of these concordance indices is similar to the binary case. They can take values from 0.5 to 1 with higher values indicating better discrimination. Given the previous definition for discrimination in the survival context, C_H is generally the recommended choice due to attractive nonparametric properties, i.e., it does not make distributional assumptions about the data, although certain situations may warrant use of the other indices. Due to its widespread and recommended use, all further illustrations of C statistics in this paper will focus on C_H.

CLASS OF R^2 STATISTICS

In the context of linear regression the coefficient of determination, R^2, is a universally adopted measure of predictive ability. It is the proportion of variability in the outcome that is explained through the model covariates. R^2 values range from 0 to 1 with higher values indicating better predictive ability. For example, an R^2 of 0.68 says that 68% of the variability in the outcome is explained by the model. For survival models, 17 different R^2-type measures are investigated by Choodari-Oskooei et al[6,7] and separated into four categories: explained variation, explained randomness, predictive accuracy, and other. Explained variation uses variance as a measure of precision, explained randomness uses entropy, and predictive accuracy compares model-based survival probabilities with survival status at a specific time. The explained variation category is the most attractive given the natural extension from linear regression, and, thus will be our focus. They are the most familiar and utilize the most common interpretation of R^2. In general, Choodari-Oskooei et al[6,7] recommend the estimators described by Kent and O'Quigley[8] $\left(R^2_{PM}\right)$ and Royston and Sauerbrei[9] $\left(R^2_D\right)$, except in specific research contexts.

Choodari-Oskooei et al[6,7] set forth four criteria for a good measure of explained variation. First, the measure needs to be uninfluenced by censoring. A measure that is unaffected by both the censoring mechanism and the amount of censoring is very desirable in survival analysis. R^2_{PM} is unaffected by censoring while R^2_D is unaffected provided the covariates of the model are symmetrically distributed, but this limitation is usually easily overcome. Second, a good measure needs to be monotone, i.e., the measure should increase as the effect of a covariate on the outcome becomes stronger. Third, a good measure should be interpretable and naturally extend the interpretation of R^2 from linear regression. Both R^2_{PM} and R^2_D satisfy the second and third criteria. Last, a good measure should be robust to influential (outlier) observations. Of the proposed measures of variation, R^2_D was the only one to demonstrate resistance to influential observations. The recommendations for practice are to compute both R^2_{PM} and R^2_D for any study, and hopefully the results will be similar.

COMPARISONS OF C AND R^2

Comparing C_H to R^2_{PM} and R^2_D is somewhat like comparing apples and oranges, in one sense, given the stark differences between the two; however, the comparison makes sense on the grounds that they are all trying to answer the same basic question: ''Does adding x to a model already containing y and z improve the predictive accuracy?''

As stated above, C_H classifies a model's ability to discriminate non-events from events. Other statistics measure calibration, how closely the predicted probabilities agree numerically with the observed outcomes. Good discrimination does not necessarily imply good calibration, nor vice-versa. C_H gives only discriminant ability. One drawback of C_H is that it is not easily interpreted. Its sister statistic the area under the curve (AUC) of the receiving operating characteristic curve (ROC) for binary outcomes without a time component, can be viewed as a combination of sensitivity and specificity. That does not extend to the survival context. Thus, we are left with a measure with a range between 0.5 and 1 that does not have a good interpretation. Another drawback is that C_H is highly dependent on the censoring mechanism in the user's dataset. Consequently, it is not generalizable to a given population.

Moreover, it cannot be calculated in a dataset with no censored observations because no censored observations implies the absence of non-events. As a final note, C_H is completely based on ranks and therefore nonparametric. Some would see this as a good thing, but we would not necessarily agree. In constructing predictive models, almost all are parametric—containing distributional assumptions (or at least semi-parametric like the Cox model—a model with parametric and non-parametric properties). In our opinion, parametric methods that are likelihood-based give better information given the context. More discussion will be given in the example section on plausible values.

Contrastingly, R^2_{PM} and R^2_D values have the same interpretation as in the linear regression context: the percentage of variation in the outcome is explained by set of predictor variables in the model. Both of these statistics can be extended to include survival models other than Cox models. Compared to C_H, R^2_{PM} and R^2_D are not as nearly affected by censoring in the data, although they cannot be completely ignored. Further, they are both likelihood-based, which is desirable for the reasons presented in the preceding paragraph. The previous section explained the practical difference R^2_{PM} and R^2_D and why they should be used in tandem. One drawback is that R^2_{PM} and R^2_D are affected by the distributions of the predictor variables. R^2_{PM} and R^2_D will increase as the variance of predictors increase. With the researcher being aware of this, he or she should be careful when seeking to make statements generalizable to the target population.

It should be noted that there are no statistical laws against computing each of the three statistics and comparing them. You will often find that they agree with each other.

SOFTWARE OPTIONS

Table 1 shows the options from the most common statistical software packages for computation of the three measures presented.

The availability of options for calculating C_H, R^2_{PM}, and R^2_D vary by the software package chosen for analysis. Of course, it is possible to write code to calculate each of these, but that is beyond the scope of this article. We will discuss the options among the four most popular statistical software packages for medical/epidemiological research: IBM's SPSS, SAS, R, and Stata.

As of Version 22, SPSS does not have any procedures for calculating C_H, R^2_{PM}, or R^2_D. A technical note[10] on the company's website gives the details on computation, coding, and calling of a macro used to calculate C_H, but only for Cox models. Nothing could be found documenting an approach for either of the R^2 measures.

SAS version 9.4 also excludes these three measures in its base package. Two papers give details on calculating C_H. One[11] gives the details and code for calculating C_H. The other[12] gives a macro that is used to calculate what they call ''Harrell's Optimism.'' Another paper[13] gives two macros, based on two separate algorithms, to calculate R^2_{PM}. All of these approaches are only described for Cox models. Nothing could be found for R^2_D.

For *R*, the only open source package in this list, nothing was discovered that could be used for calculating either of the R^2 statistics. The package, ''Hmisc,'' written by Harrell, is a comprehensive package containing many data analysis and graphics tools. In this package, the ''rcorr.cens'' is used to compute C_H for Cox models.

Starting with Stata version 9 and continuing through the current version, 14, C_H can be computed through the ''estat concordance'' command after a Cox regression. R^2_{PM} and R^2_D are calculated through the user-written command ''str2ph'' and ''str2d,'' respectively. The command ''str2ph'' is easily extendable beyond Cox models to other proportional hazards models. It will also give reasonable results for non-proportional hazards parametric models and flexible parametric survival models. The alternative ''str2d'' is valid for any proportional hazard, proportional odds, or probit model for survival data, except for models with a gamma survival distribution. Royston[14] gives an overview of both of these statistics and details an example of both.

In summary, Stata is the only package that can compute all three statistics. SAS, through user-written

Table 1. Statistical software and discrimination statistics

Software	C_H	R^2_{PM}	R^2_D
IBM SPSS	Macro from technical note	None	None
SAS	Two macros (user-written)	Two macros (user-written)	None
R	Package "Hmisc" (user-written)	None	None
Stata	"estat concordance" (base Stata)	"str2ph" (user-written)	"str2d" (user-written)

macros will give C_H and R^2_{PM}. Stata also provides confidence intervals for R^2_{PM} and R^2_D but not for C_H. Both SPPS, via a macro, and R, via an added package, with calculate C_H but neither of the R^2 values.

EXAMPLE

As an example, the Worcester Heart Attack Study[15] offers a subset of 500 subjects often cited[16] for teaching examples for survival data. The data contain 500 observations with 22 variables, of which we will use age at hospital admission, gender, BMI, heart rate, diastolic blood pressure, and congestive heart failure to build a survival Cox model. The event of interest is all-cause mortality, and in the data we have 215 events, giving 285 censored observations. The median survival time was 4.45 years. Table 2 shows univariate Cox models for each of the designated variables.

Looking at the hazard ratios, we see higher BMI and higher diastolic BP being protective and all other variables being hazardous, with congestive heart complications showing the largest hazard ratio. C statistics typically fall in the range from 0.5 to 1, with 0.5 indicating that the discriminant ability is no better than tossing a fair coin. R^2 values range from 0 to 1, higher indicating more explained variance and therefore better predictive accuracy. In spite of the P values being highly significant, none of these univariate models give a strikingly accurate predictive model based on C_H, R^2_{PM}, and R^2_D. This is not surprising as we expect mortality is predicted by more than a single risk factor. Taking one variable at a time, age at hospital admission is the most accurate with $C_H = 0.731$, $R^2_{PM} = 0.362$, and $R^2_D = 0.298$. C_H does not have a practical interpretation, but we can say that this model performs better than flipping a coin. The interpretation for R^2_{PM} (R^2_D) would be: ''36.2% (29.8%) of the variation in time to death can be explained by the variation in age at hospital admission.'' Although for this particular univariate model, R^2_{PM} is greater than R^2_D, this is not always the case—see for instance the univariate model for gender.

One may also investigate the contribution of the variables to the model using these indices in a sequential manner by adding one variable at a time. From the univariate models in Table 2, the order of variables from largest to lowest R^2_{PM} and R^2_D are as follows: AGE, CHC, BMI, HR, DBP, and Gender. Building on the models using this ordering, one can see the increment in the C_H, R^2_{PM} and R^2_D values when adding one more predictor variable. The largest incremental increase from the model with only AGE happens after adding CHC in the model with an increase of 0.026 in C_H, 0.077 in R^2_{PM}, and 0.059 in R^2_D. Adding BMI in a model with AGE and CHC only resulted in an increase of 0.008 in C_H, 0.017 in R^2_{PM}, and 0.01 in R^2_D. This full additive procedure is seen in Table 3.

Another aspect to mention is that it is much easier to improve a model that starts with a low accuracy. It is very difficult to significantly improve a C statistic that starts at 0.90 or an R^2 statistic that starts at a 0.85. The lower the starting value, the easier it is to improve. Table 4 displays this by showing the discrimination statistics in two separate Cox models. The columns labeled ''Before'' show the statistics with all other variables listed in this table, excluding the one on the current row. The columns labeled ''Full'' show the statistics for the full Cox model. Hazard ratios and P values are from the full model.

Adding age at hospital admission to a model with gender, BMI, HR, DBP, and CHC increases R^2 values from $R^2_{PM} = 0.365$ to 0.496 (from $R^2_D = 0.331$ to 0.414). Similarly, adding congestive heart complications to obtain the full model showed about 5% increase in the R^2 measures. The other variables do not add much in terms of predictive accuracy when all other variables are in the model. On the other hand, the maximum increase in the C_H values from fitting the full model was 0.74-0.777 attributed to age at hospital admission. As a final

Table 2. Univariate Cox models and their discriminative accuracy

	HR	P value	C_H	R^2_{PM}	R^2_D
Age at hospital admission (age)	1.068	<0.001	0.731	0.362	0.298
Female (gender)	1.464	0.006	0.542	0.021	0.033
Body Mass Index (BMI)	0.906	<0.001	0.648	0.132	0.141
Initial heart rate (HR)	1.015	<0.001	0.614	0.080	0.081
Initial diastolic BP (DBP)	0.984	<0.001	0.611	0.066	0.067
Congestive heart complications (CHC)	3.314	<0.001	0.642	0.195	0.244

Shown are hazard ratios (HR) and P values from the univariate models, along with the three discrimination statistics for each

Table 3. Cox models and their predictive accuracy statistics

Predictors in the model	C_H	R^2_{PM}	R^2_D
AGE	0.731	0.362	0.298
AGE + CHC	0.757	0.439	0.357
AGE + CHC + BMI	0.765	0.456	0.367
AGE + CHC + BMI + HR	0.769	0.473	0.373
AGE + CHC + BMI + HR + DBP	0.776	0.490	0.397
AGE + CHC + BMI + HR + DBP + Gender	0.777	0.496	0.414

Table 4. Full Cox models and semi-full Cox models and their predictive accuracy

	HR full	*P* value full	C_H before	C_H full	R^2_{PM} before	R^2_{PM} full	R^2_D before	R^2_D full
Age at hospital admission (age)	1.051	<0.001	0.740	0.777	0.365	0.496	0.331	0.414
Female (gender)	0.763	0.060	0.776		0.490		0.397	
Body Mass Index (BMI)	0.956	0.006	0.771		0.481		0.403	
Initial heart rate (HR)	1.011	<0.001	0.770		0.469		0.380	
Initial diastolic BP (DBP)	0.989	0.003	0.771		0.478		0.389	
Congestive heart complications (CHC)	2.176	<0.001	0.759		0.442		0.361	

Hazard ratios (HR) and p-values are from the full model. Columns labeled "Before" indicate that the model included all variables on this table except the one on the current row

note, a model with all of the 6 explanatory variables considered explains about 49.6% using R^2_{PM} (or 41.4% using R^2_D) of the observed variability in the time to mortality.

SUMMARY/CONCLUSIONS

We discussed available methods to quantify predictive accuracy of survival models. Of the class of *C* statistics Harrell's *C* is the most highly recommended for general use. The R^2 values of Kent and O'Quigley and Royston and Sauerbrei are recommended in that class of statistics. We prefer R^2 measures due to their interpretability and robustness to censoring. Stata is the only statistical software package currently able to provide all three, SAS has macros for two, and both R and SPSS can only calculate C_H. Computing some or all three values for any model is quick and relatively simple, given the proper computer program. If the researcher is able, he or she should calculate all three to give a robust overall summary of predictive ability and subsequently increase his or her knowledge of the value of the new predictor.

We should also emphasize the importance of model diagnostic measures, such as how well the model fits and validating the assumptions involved. These three measures, by themselves, are not sufficient to give all of the information about the usefulness of the model. They should be used in conjunction with model diagnostic tools.

Disclosure

Neither author has financial interest to disclose.

References

1. Pencina MJ, D'Agostino RB, Song L. Quantifying discrimination of Framingham risk functions with different survival C statistics. Stat Med. 2012;31:1543-53.
2. Chambless LE, Diao G. Estimation of time-dependent area under the ROC curve for long-term risk prediction. Stat Med. 2006;25:3474-86.
3. Harrell FE, Lee KL, Mark DB. Tutorial in biostatistics: Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15:361-87.
4. Uno H, Cai T, Pencina MJ, D'Agostino R, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med. 2011;30:1105-17.
5. Gönen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. Biometrika. 2005;92(4):965-70.

6. Choodari-Oskooei B, Royston P, Parmar MKB. A simulation study of predictive ability measures in a survival model I: Explained variation measures. Stat Med. 2012;31:2627-43.
7. Choodari-Oskooei B, Royston P, Parmar MKB. A simulation study of predictive ability measures in a survival model II: Explained variation measures. Stat Med. 2012;31:2644-59.
8. Kent J, O'Quigley J. Measures of dependence for censored survival data. Biometrika. 1988;75(3):525-34.
9. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. Stat Med. 2004;23:723-48.
10. Can SPSS Statistics produce Harrell's C or Somer's D following a Cox regression? IBM, Armonk, NY, Tech. Note Reference #1478383, July 11, 2014.
11. Liu L, Forman S, Barton B. Fitting Cox model using PROC PHREG and beyond in SAS. SAS Global Forum, Vol. 2009.
12. Miao Y, Cenzer IS, Kirby KA, Boscardin WJ. Estimating Harrell's optimism on predictive indices using bootstrap samples. SAS Global Forum, Vol. 2013.
13. Heinzl H. Using SAS to calculate the Kent and O'Quigley measure of dependence for Cox proportional hazards regression model. Comput Methods Programs Biomed. 2000;63:71-6.
14. Royston P. Explained variation for survival models. Stata J. 2006;6:83-96.
15. Goldberg RJ, Gore JM, Alpert JS, Dalen JE. Recent changes in attack and survival rates of acute myocardial infarction (1975 through 1981): the Worcester Heart Attack Study. JAMA. 1986;255(20):2774-9.
16. Hosmer DW, Lemeshow S, May S. Applied survival analysis: Regression modeling of time-to-event data. 2nd ed. Hoboken: Wiley; 2008. p. 191.