Institute for Health and Society, Medical College of Wisconsin

Datum

Biostatistics NEWSLETTER
Key Function of the CTSI
& MCW Cancer Center Biostatistics Unit

ROC Curves and the C statistic

Brent Logan, PhD, Professor, Division of Biostatistics

Receiver Operating Characteristic (ROC) curves and the Concordance (C) statistic are often used to assess the ability of a risk factor to predict outcome. We will focus on prediction models for a binary outcome using logistic regression, although extensions for censored time to event data are also available. For example, often a biomarker or risk factor is included in a logistic regression model to predict the likelihood a patient will develop a disease of interest. These predictive probabilities or risks can be examined to see how accurate they are at identifying patients who will develop the disease or not. Predictive accuracy is usually considered in two domains: calibration, which refers to a measure of how well the predicted probabilities agree with the observed probabilities, and discrimination, which refers to how well the model can separate the diseased individuals from the healthy individuals. For example, if the predicted probabilities for the diseased individuals are all higher than the predicted probabilities for the healthy individuals, then we say that the model has perfect discrimination. Discrimination is commonly measured using ROC curves. To construct an ROC curve, the predicted probabilities of the outcome of interest are repeatedly dichotomized into above vs. below a cutpoint. For each cutpoint, one can estimate the sensitivity (probability that the predicted risk is above the cutpoint among patients with the disease) and the specificity (probability that the predicted risk is below the cutpoint among patients without the disease). One can vary the cutpoint to show a range of sensitivities vs. specificities, and an ROC curve is a plot of Sensitivity vs. (1-Specificity) over the range of possible cutpoints. If the model has perfect discrimination, the ROC curve should hit the upper left corner of the plot (100% sensitivity and 100% specificity).

The area under the ROC curve is a useful measure for summarizing the ROC curve. If a curve is close to the upper left corner (Sensitivity=100%, Specificity=100%), then the area under the ROC curve should be close to 1. The area under the ROC curve is equivalent to another statistic commonly used to summarize model discrimination, the C statistic or Concordance statistic. The C statistic is interpreted as the probability that a randomly selected subject who experienced the outcome will have a higher predicted probability of having the outcome occur than a randomly selected subject who did not experience the outcome. In addition to computing the area under the ROC curve, this probability can be estimated by taking all pairs of observations where one patient experienced the event and the other did not, and computing the proportion of those pairs where the patient experiencing the event had the higher predicted risk. The C statistic can also be interpreted as the rank correlation between predicted probabilities of the outcome occurring and the observed response.

As an example, consider a recent study by Levine et al. (2012) looking at the ability of a biomarker panel with 6 components to improve prediction of 6 month mortality after development of acute Graft-versus-host disease (aGVHD) in patients receiving

(continued on page 2)

Volume 19, Number 4
November/December 2013

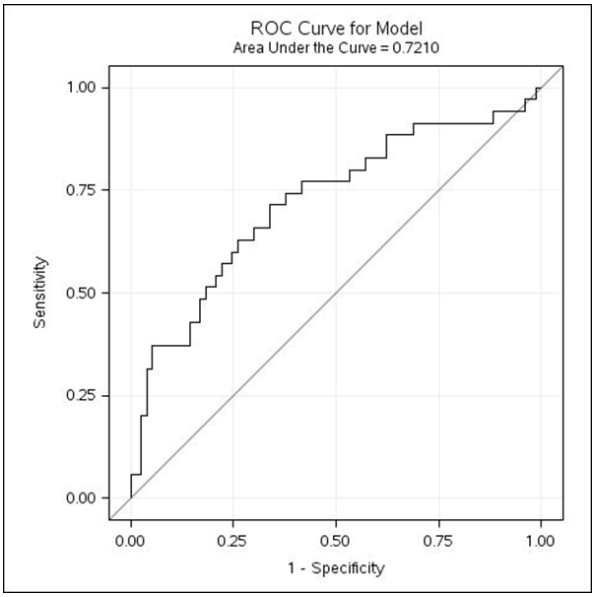
In this issue:

a bone marrow transplant. For this illustration we focus on the biomarker panel measured at diagnosis of aGVHD, although the authors also considered subsequent assessments. A logistic regression model was constructed including the biomarkers. This model produces predicted probabilities of response for each patient, and can be examined to see how accurate it is at identifying patients who are alive or not at 6 months. A total of 35 patients died within 6 months, while 77 were still alive. An illustrative set of thresholds for the predicted probability of 6 month mortality are shown in Table 1. For each threshold, one can determine the sensitivity and specificity, where sensitivity is the % of deceased patients with predicted mortality above the threshold, and specificity is the % of surviving patients with predicted mortality below the threshold. As expected, as the threshold for predicted 6 month mortality rises, the sensitivity decreases and the specificity increases. The ROC curve is shown in Figure 1, and is a plot of the sensitivity vs. 1 minus the specificity across the entire set of potential thresholds, not just the ones listed in the table.

Table 1: Predicted 6 month mortality thresholds, sensitivity, and specificity for the acute GVHD example.

Predicted 6 mo. Mortality Threshold	# of deceased patients above threshold, out of 35 (% Sensitivity)	# of surviving patients below threshold out of 77 (% Specificity)
10%	33 (94%)	6 (8%)
20%	31 (89%)	29 (36%)
30%	25 (71%)	50 (65%)
40%	18 (51%)	62 (81%)
50%	13 (37%)	73 (95%)
60%	5 (14%)	75 (97%)

Figure 1: ROC Curve for acute GVHD example



The area under this ROC curve is C=0.721. The authors also considered a prediction model which included clinical characteristics alone (C=0.623) and a model with both clinical characteristics and the biomarker panel (C=0.734). The biomarker panel does better than clinical characteristics alone at discriminating between patients who survive to 180 days and those who do not, and use of the biomarker panel as well as clinical characteristics results in even further improvement in the C statistic.

While the C statistic can be a useful tool for describing the discrimination ability of a predictive model, it is important to keep in mind its limitations (Cook, 2007). One concern is that the scale of the C statistic is condensed so that even important predictive variables with a substantial odds ratio may have only a small impact on the C statistic. The concept of calibration is also important in assessing the importance of a risk factor, and calibration can be assessed for example through a Hosmer-Lemeshow test. This test compares the observed and predicted risk probabilities within groups, to assess how closely the predicted risk matches the observed risk. More global measures of model fit such as likelihood statistics or Brier scores, which combine calibration and discrimination, may also be useful.

References:

1. Levine JE, Logan BR, Wu J, Alousi AM, Meade JB, Ferrara JLM, Ho VT, Weisdorf DJ, Paczesny S. Acute Graft-Versus-Host Disease Biomarkers Measured During Therapy Can Predict Treatment Outcomes: A Blood and Marrow Transplant Clinical Trials Network Study, Blood, 119:3854-60, 2012.
2. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation. 115: 928-935, 2007.
3. Harrell FE Jr. *Regression modeling strategies.* New York: Springer, 2001.

Longitudinal Studies: Design and Analysis Considerations

Jennifer Le-Rademacher, PhD, Assistant Professor, Division of Biostatistics

Longitudinal studies are routinely conducted in many areas such as medicine, psychology, sociology, and public health. However, longitudinal studies require special considerations in study design as well as in data analysis. This article provides a brief discussion of design and analysis considerations to help optimize the benefits of longitudinal studies. Unlike cross-sectional studies where outcomes for each subject are measured at a single time point, outcomes in longitudinal studies are measured at multiple time points per subject. Subjects in a longitudinal study are expected to have the same number of responses measured at a set of pre-specified time points. Responses measured over a period of time allow investigators to evaluate changes in outcomes over time. Longitudinal responses further allow investigators to adjust for variability in responses among individuals when evaluating the changes in response over time.

As an example, suppose data for ten individuals were collected in a longitudinal study where the outcome was measured at six monthly visits for each patient after they were given a new treatment. Figure 1 shows the outcomes for each patient plotted against time. The response pattern can be explored by connecting the outcomes measured at the monthly visits for each individual. Figure 1 suggests an increase in response over time for all patients. It also suggests that there is variability in outcomes among subjects. To contrast longitudinal data with cross-sectional data, Figure 2 shows data from a cross-sectional study for the same treatment where the same outcome was measured but only once per subject. Sixty patients would be needed to provide the same number of responses (ten at each time point) as in the longitudinal study. Furthermore, since the responses at different time points came from different individuals, it is impossible to explore patients' response pattern over time. It is also impossible to separate the effect of time on outcome from inherent variability among subjects.

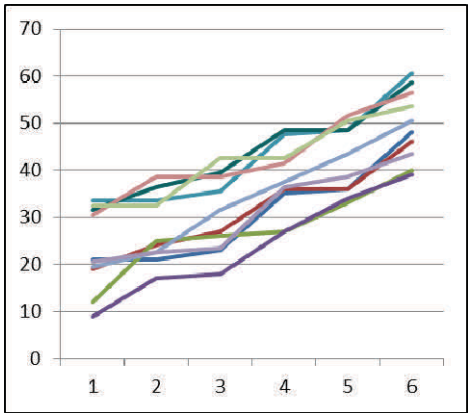


Figure 1: Responses from longitudinal study

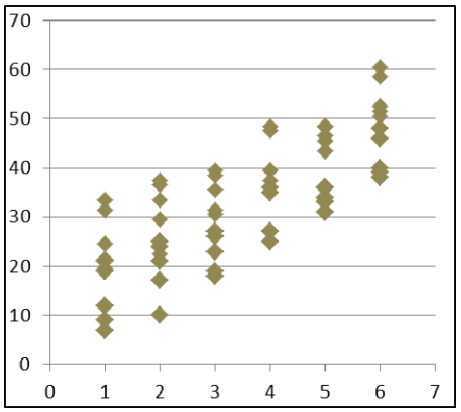


Figure 2: Responses from cross-sectional study

Although there are many advantages to longitudinal studies, designing longitudinal studies require special considerations and analysis of longitudinal data can be complex. To measure outcomes at multiple time points, subjects in a longitudinal study are typically followed for a longer period of time than in a cross-sectional study evaluating the same outcome. During the study period, some subjects may drop out of the study completely and some may miss one or more follow-up visits resulting in unobserved response at these time points. Missing data is a common concern in longitudinal studies. If not handled correctly, missing data can lead to biased conclusions at the end of study. Although it may not be completely eliminated, missing data can be minimized with careful planning in the design stage. For instance, the number of responses measured should be the minimum necessary to answer the study questions. Responses should be taken during time periods where patients are more likely to follow up. A well-planned follow-up process to ensure responses are collected at these time points will also help minimize missing data.

Besides missing data, another reason analysis of longitudinal data is more complex than analysis of cross-sectional data is due to the correlation between multiple responses from the same subject. Analysis ignoring this correlation may lead to biased conclusions. Various analysis methods are available for longitudinal data. These methods model quantities with different interpretations of treatment effect. They also assume different mechanism of missing data. Common analysis methods include evaluating the cumulative effect of treatment at each time point (repeated analyses, example Table 1), evaluating the effect of treatment from one time point given the response from a previous time point (transition models, example Table2), or more flexible approach that model the conditional effect of treatment as well as the effect of time, or an interaction between these two effects (mixed models, example Tables 3 – 5) on outcome. The appropriate method for a study depends on the study's objectives as well as the assumption of why responses are missing.

(continued on page 4)

Suppose a longitudinal study was conducted to evaluate the effect of a new treatment compared to a control. The response of interest was measured at the time of randomization and monthly for 4 months after randomization. For illustration, three different methods were used to analyze the data. Results from these analyses are shown in Tables 1 through 5. In the first analysis (Table 1), the response at each time point was modeled separately. Baseline measure was included as a covariate in all four models. Table 1 shows that, given the same measure at baseline, response at one month after randomization in treatment group was about 8 units higher compared to the control group whereas response at four months was only 4 units higher in the treatment group compared to the control group. This analysis estimates the cumulative effect of treatment separately for each of the four time points after adjusting for the baseline. It does not compare responses between post-randomization periods. This analysis includes subjects with responses observed at all time points (these subjects are called complete cases). Inference from this analysis is valid if the reason for missing is unrelated to the outcome of interest where the complete cases are a random sample of the population.

Table 1: Cumulative effect of treatment adjusted for baseline response

Response	1 month	2 months	3 months	4 months
Intercept	10.24 (<.001)	10.92 (<.001)	7.69 (<.001)	8.20 (<.001)
Baseline	0.42 (<.001)	0.42 (<.001)	0.49 (<.001)	0.47 (<.001)
Treatment	7.97 (<.001)	6.13 (<.001)	5.68 (<.001)	3.76 (<.001)

In the second analysis (Table 2), the response at each time point was also modeled separately but unlike the first analysis which adjusts for the response at baseline, this analysis adjusts for the response measured at the previous month. As expected, the effect of treatment on month 1 response in this analysis is the same as in the first analysis. However, the effects of treatment on response in subsequent months are different in this analysis because these models estimate the effect of treatment on response at the current time point conditional on response from the previous month. Specifically, the average increase in response at 2 months is 2.5 units higher for subjects in the active group compared to subjects with the same response at 1 month in the control group. This analysis estimates the conditional effect of treatment at each of the four time points after adjusting for response from the previous month. However, it is not possible to estimate the cumulative effect of treatment on response from this analysis. Similar to the previous analysis, this method only includes complete cases and is valid only if missing data are unrelated to response.

Table 2: Transitional effect of treatment adjusted for response in previous month

Response	1 month	2 months	3 months	4 months
Intercept	10.24 (<.001)	9.67 (<.001)	4.20 (<.001)	5.40 (<.001)
Prior month response	0.42 (<.001)	0.45 (<.001)	0.63 (<.001)	0.66 (<.001)
Treatment	7.97 (<.001)	2.53 (0.02)	1.83 (0.10)	0.03 (0.98)

The third analysis uses a linear mixed model which is a flexible approach to model the effect of treatment as well as the effect of time on response. Table 3 shows that baseline measure, treatment, time, and time by treatment interaction are all highly associated with response (p-values < 0.05). Treatment effect at various time points can be estimated from this model as shown in Table 4. Note that these estimates are similar to the results from the first analysis (Table 1). In addition to treatment effect, time effect can also be obtained from this model. Due to the interaction between treatment and time in the model, the effect of time can be estimated separately for the control and the treatment group. Table 5 shows comparisons of responses between various time points. There is no difference in responses over time in the control group whereas there is a time effect in the treatment group. The mixed model allows estimation of the treatment effect at various time points as well as estimation of the effect of time on response. More importantly, this analysis includes all available data including responses from subjects with some missing observations. Inferences from the mixed models are valid under a less restrictive assumption of missing data where the reason for

Table 3: Overall tests from linear mixed model

Effect	Num DF	Den DF	F value	P-value
Baseline	1	349	183.55	<.0001
Treatment	1	349	45.98	<.0001
Time	3	1050	13.01	<.0001
Treatment*Time	3	1050	3.49	0.0152

Table 4: Treatment effect at various time points adjusted for baseline from mixed model

Time Point	Estimate (95% CI)	P-value
1 month	7.96 (5.65, 10.28)	<.0001
2 months	6.13 (3.81, 8.44)	<.0001
3 months	5.68 (3.37, 8.00)	<.0001
4 months	3.76 (1.45, 6.08)	0.0015

(continued on page 5)

missing responses can be related to the observed responses but not related to the unobserved responses. This example illustrates some common methods available for longitudinal data analysis. However, these methods provide estimates with different interpretation. In general, the mixed models approach is flexible and can be used to estimate various combinations of treatment effect as well as time effect on response. These models also provide valid inferences under a less restrictive assumption about the missed observations. It is important to select the appropriate analysis method to address the study objectives and the assumption about missing data.

Table 5: Time effect by treatment from mixed model

Time Comparison	Control		Treatment	
	Estimate (95% CI)	P-value	Estimate (95% CI)	P-value
2 m vs. 1 m	0.66 (-1.12, 2.44)	0.47	-0.26 (-1.12, 2.44)	0.69
3 m vs. 1 m	-1.48 (-3.26, 0.30)	0.10	-2.62 (-3.26, 0.30)	<.0001
4 m vs. 1 m	-1.21 (-2.99, 0.57)	0.18	-3.31 (-2.99, 0.57)	<.0001
3 m vs. 2 m	-2.14 (-3.92, -0.36)	0.02	-2.37 (-3.65, -1.08)	0.0003
4 m vs. 3 m	0.27 (-1.51, 2.06)	0.76	-0.69 (-1.97, 0.60)	0.29

Upcoming Events:

Bridging the sciences BIOSTATISTICS

International Conference on Survival Analysis in Memory of John P. Klein
June 26-27th, 2014
Medical College of Wisconsin, Milwaukee, WI
More information can be found here: www.mcw.edu/biostatistics/JPKconference.htm

Biostatistics Lecture Series 2014:
Analysis of Count Data
Kwang Woo Ahn, PhD
Friday, March 7, 2014
12:00-1:00 pm
Medical Education Building– M2050

Simple Statistics Using "R"
Jessica Pruszynski, PhD
Friday, April 4, 2014
12:00-1:00 pm
Medical Education Building– M2050

Datum

Editors:
Jennifer Le-Rademacher, PhD
Ruta Brazauskas, PhD
Brent Logan, PhD
Alexis Visotcky, MS
Haley Montsma, BBA

Datum is published by the Division of Biostatistics. It is available online at: www.mcw.edu/biostatistics/datum.htm.
We welcome your newsletter questions, comments, and suggestions. Please contact us with address updates or to be added/removed from our mailing list.

Please direct all newsletter correspondence to:
Datum Newsletter ● Medical College of Wisconsin ● Division of Biostatistics
8701 Watertown Plank Road ● Milwaukee, WI 53226—0509
Phone: 414-955-7439 ● E-mail: hmontsma@mcw.edu