PharmaSUG 2014 - Paper HA09

Using SAS® to Calculate and Compare Adjusted Relative Risks, Odds Ratios, and Hazard Ratios

Besa Smith, MPH, PhD, ANALYDATA, San Diego, CA

Tyler Smith, MS, PhD, National University, School of Health and Human Services,
Department of Community Health, San Diego, CA

ABSTRACT

In the past decade, health outcomes research has gained in popularity as increasing focus has been given to improving patient outcomes. Whether upon refining screening strategies for earlier detection of disease, reducing readmission or nosocomial infection rates, improving patient satisfaction, or evaluating new patient therapies and treatments, the underlying focus is to improve prevention, control, and/or treatment. The analysis of the occurrence of health outcome events is dependent upon differing risk sets among those with and without the event of interest and is often modeled using one of three approaches. Using regression methods, we often see relative risk estimates, odds ratios, or hazards ratios presented after adjusting for a list of covariates that may be distorting our view. This paper will use SAS® to compare the process and results of a log-binomial regression, logistic regression, and Cox regression in the context of several covariates and including a temporal element. Discussion of why a researcher would use a certain approach in a specific situation will be discussed. Health outcome researchers strive to identify at-risk populations by providing quantitative evidence that allows for more informed decisions by practitioners and policy makers. This paper presents the code and results of three frequently used approaches in the evolving environment of health analytics.

INTRODUCTION

In the Institute of Medicine's 2000 report Crossing the Quality Chasm they state, "Between the health care that we now have and the health care we could have lies not just a gap, but a chasm." This sentiment is echoed by many who feel our US health care system is broken. Fueled by the discussion of cost and quality, there are many factors responsible for the growing attention to the quality of healthcare in the US including a desire to do better. Health outcomes research is applied clinical and population-based research that seeks to study and optimize the end results of healthcare by identifying shortfalls in practice in order to develop strategies to improve care. The spectrum of health outcomes studied is thus broad and ranges from morbidity due to disease or injury, to patient satisfaction. Health outcomes research has become the assimilation of all aspects of the healthcare process that are made up by the interrelated nature of clinical, administrative, and policy processes and their impact on populations.

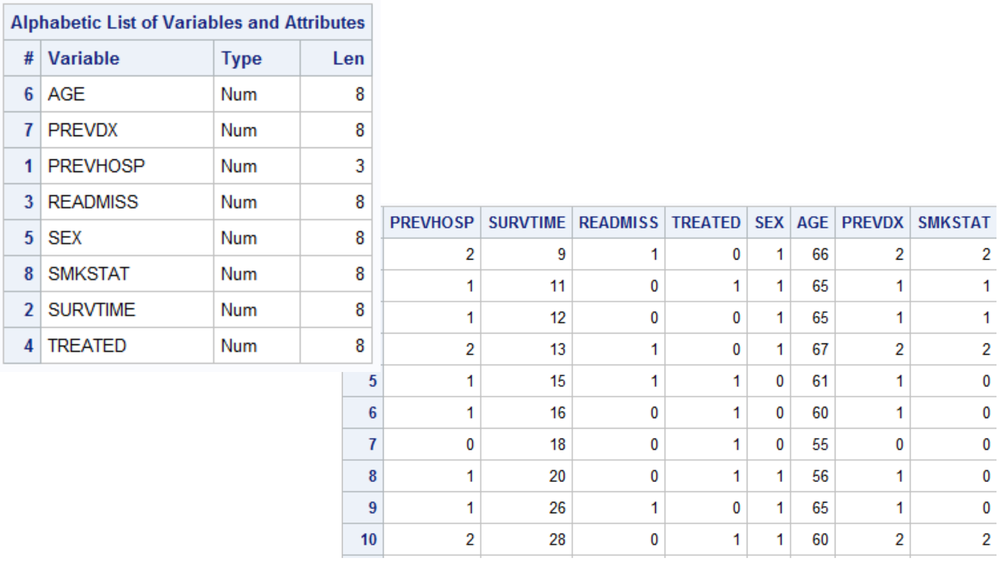
Health outcomes research requires the joining of many disciplines to manage the complex aggregation of interventions, implementation of disease management or prevention programs, and creation of clinical and business decisions to aid in controlling costs and allocating resources more efficiently through the examination of clinical, economic, medical, and quality-of-life outcomes with a common goal of improving patient health.

Health informatics solutions in the past decade have enabled massive amounts of clinical and administrative data to become managed and warehoused with increasing integration and interoperability. There is now a significant gap for talent capable of taking these data, framing objectives relative to the enterprise, conducting an analysis, interpreting the results, and disseminating the findings. In this paper we focus on three measures of effect that are often used in health outcomes research. This paper is designed to be an intermediate level review of the three measures (the odds ratio, relative risk, and hazard ratio) and should be a starting point for the theory, programming, and interpretation. Suggested readings are presented at the end of this paper and should be considered as important complementary resources.

DATA FOR THIS PAPER

In this paper we will focus on a dichotomous outcome variable. We will apply all analyses to a set of data consisting of 61 observations and 8 variables. The outcome of interest will be hospital readmission (yes/no). Based on observational data, the objective of this analysis is to investigate the effect of a remote monitoring intervention post heart failure readmission.

The data look like this:



Alphabetic List of Variables and Attributes

#	Variable	Type	Len
6	AGE	Num	8
7	PREVDX	Num	8
1	PREVHOSP	Num	3
3	READMISS	Num	8
5	SEX	Num	8
8	SMKSTAT	Num	8
2	SURVTIME	Num	8
4	TREATED	Num	8

	PREVHOSP	SURVTIME	READMISS	TREATED	SEX	AGE	PREVDX	SMKSTAT
	2	9	1	0	1	66	2	2
	1	11	0	1	1	65	1	1
	1	12	0	0	1	65	1	1
	2	13	1	0	1	67	2	2
5	1	15	1	1	0	61	1	0
6	1	16	0	1	0	60	1	0
7	0	18	0	1	0	55	0	0
8	1	20	0	1	1	56	1	0
9	1	26	1	0	1	65	1	0
10	2	28	0	1	1	60	2	2

THE ODDS RATIO

The odds ratio (OR) is a measure of association used to quantify the relationship between the dependent variable, Y, and the primary independent variable of interest, X_1. It is primarily used in research where estimates of incidence, or risk, are unattainable in either or both the exposed and unexposed populations. In these situations where risk cannot be measured, we use the odds ratio as a measure of the relative odds of disease.

$$\text{OR} = \frac{\text{Odds of exposure among those with the disease}}{\text{Odds of exposure among those without the disease}}$$

We can break this down further where:

$$\text{Odds of exposure among the diseased} = \frac{\text{Proportion of diseased who were exposed}}{\text{Proportion of diseased who were not exposed}}$$

$$\text{Odds of exposure among the nondiseased} = \frac{\text{Proportion of nondiseased who were exposed}}{\text{Proportion of nondiseased who were not exposed}}$$

Consider the following 2x2 table:

	Diseased	Not diseased	
Exposed	a	b	
Not exposed	c	d	
			N

The "a" cell represents the number of diseased who were exposed, while the "c" cell represents the number of diseased who were not exposed. Likewise, "b" represents the number of nondiseased who were exposed, and "d" represents the number of nondiseased who were not exposed. Knowing this, we can calculate the proportions of individuals in each disease state who were either exposed or not exposed:

Proportion of diseased who were exposed = a/(a+c)
Proportion of diseased who were not exposed = c/(a+c)
Proportion of nondiseased who were exposed = b/(b+d)
Proportion of nondiseased who were not exposed = d/(b+d)

With these individual proportions, we can then return to calculating the odds:

$$\text{Odds of exposure among the diseased} = \frac{a/(a+c)}{c/(a+c)}$$

$$\text{Odds of exposure among the nondiseased} = \frac{b/(b+d)}{d/(b+d)}$$

And the odds ratio would then be:

$$OR = \frac{\left(\frac{a/(a+c)}{c/(a+c)}\right)}{\left(\frac{b/(b+d)}{d/(b+d)}\right)}$$

Mathematically, this reduces to:

$$OR = \frac{ad}{bc}$$

Using our example in this paper, we would interpret an OR = 0.5 in the following manner: Patients who had the remote monitoring intervention were half as likely to be readmitted for heart failure when compared with patients who did not have the remote monitoring intervention.

THE RELATIVE RISK

Similar to the odds ratio, the relative risk (RR) is a measure of association used to quantify the relationship between the dependent variable and the primary independent variable of interest. The relative risk, however, is a direct comparison between the risk of disease in the exposed persons and the risk of disease in the unexposed persons.

$$RR = \frac{\text{Risk of disease among exposed}}{\text{Risk of disease among unexposed}}$$

For this we need a measure of risk which can be estimated using the incidence rate of disease during a specified period of time:

$$\text{Incidence rate of disease} = \frac{\text{Number of new cases of disease in the population}}{\text{Number of persons at risk for developing the disease}}$$

Therefore the relative risk becomes:

$$RR = \frac{\text{Incidence rate of disease among exposed}}{\text{Incidence rate of disease among unexposed}}$$

And using our 2x2 above:

$$RR = \frac{a/(a+b)}{c/(c+d)}$$

Using our example in this paper, we would interpret an RR = 0.5 in the following manner: Patients who had the remote monitoring intervention were at half the risk for readmission due to heart failure when compared with patients who did not have the remote monitoring intervention.

THE HAZARD FUNCTION

In many studies the single largest limitation to the odds ratio or relative risk is the inability to incorporate a time element into the estimation. It stands to reason that if a patient has twice the amount of observation time for an event, their probability of event would be greater. The hazard function describes the concept of the risk of an outcome (e.g., death, failure, hospitalization) in an interval after time t, conditional on the subject having "survived" to time t. It is the probability that an individual dies (has an event) somewhere between t and (t + Δ), divided by the probability that the individual survived beyond time t.

The hazard function may be more intuitive to use in survival analysis than the pdf because it quantifies the instantaneous risk that an event will take place at time t given that the subject survived to time t. Cox recognized this appeal in his 1972 paper where he outlines a robust regression method that does not require the identification of a probability distribution to represent survival times.

The hazard function h(t) is given by the following:

$$h(t) = P\{ t < T < (t + \Delta) \mid T > t\}$$

$$= f(t) / (1 - F(t)) = f(t) / S(t)$$

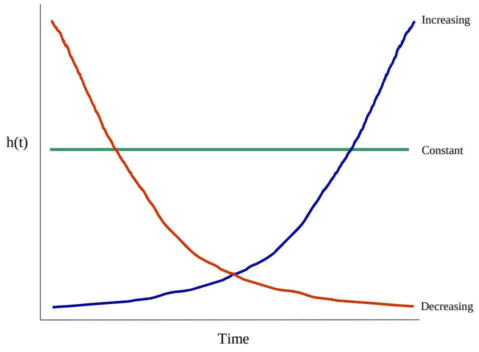


Figure 1. The plot of a constant hazard such as seen with accidents, an increasing hazard such as seen with the aging process of a mechanical engine, and a decreasing hazard such as seen with risk of dying after surgery.

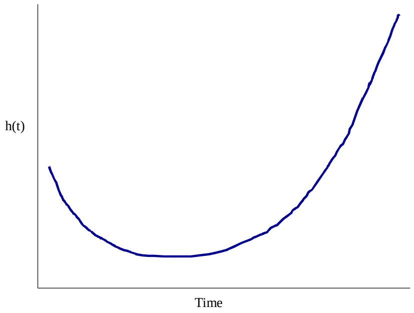


Figure 2. The plot of the hazard of death during a lifetime begins high at birth then goes down for many years before beginning to steadily increase through the aging process.

Incomplete Data

Observational and experimental studies involving follow-up over time often experience late arrival along with loss to follow-up of subjects during the observation period. Through censoring and truncation techniques, survival analysis allows for a study to start without all experimental units yet enrolled and to end before all experimental units have experienced an event. This is important because even in well-developed studies there will be subjects who choose to quit participating, move too far away to be followed, die from some unrelated event, or will simply not have an event before the end of the observation period. Using censoring techniques, the researcher can allow each experimental unit to contribute to the model all of the information possible for the amount of time the researcher is able to observe the unit.

Right and Left Censoring

The most common form of censoring for incomplete data is right censoring where a subject's follow-up time terminates before the outcome of interest is observed. Assumed non-informative, Type I right censoring occurs when the observation time reaches the end of a defined study period and the subject has not had an event, while Type II right censoring occurs when the researcher ends the follow-up period based on a pre-specified number of events occurring. The term right censoring also includes censored subjects who are lost to follow-up. Right censoring techniques allow subjects to contribute to the model until they are no longer able to contribute (end of the study, or withdrawal).

An observation is left censored if the event of interest has already occurred when observation time began. For example, in a study of myocardial infarction we begin following a group of people at age 50. However, some may have already had an event prior to the start of follow-up and unless you gain information as to the time of the events, the myocardial infarction may be left censored at age 50. In this paper we focus on the more typical right censoring.

Figure 3 presents a study design where the observation times start at differing points after the beginning of the study period. After t=0 is established, there is a fixed follow-up period. The X's represent events and the O's represent censored observations. Some subjects have events early in the study period and others have events at the end of the study period. Likewise some subjects enter the study period late and/or leave the study period early, but most do not have an event during the entire study and are simply right censored at the end. In this example there is no need for truncation techniques and we assume the censoring to be non-informative.

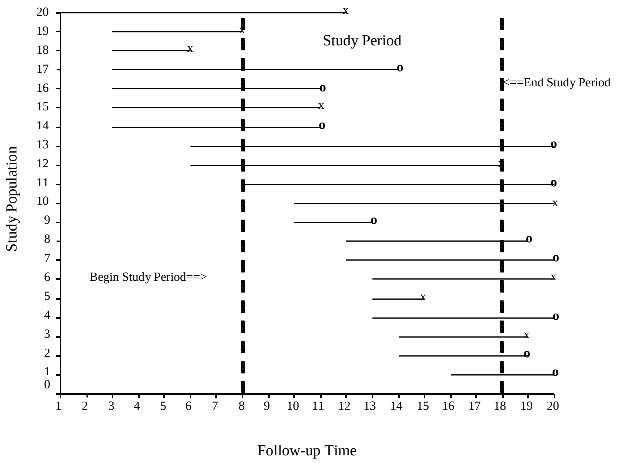


Figure 3. Follow-up time with delayed entry and censoring. (x denotes event; o denotes censored)

PROC FREQ

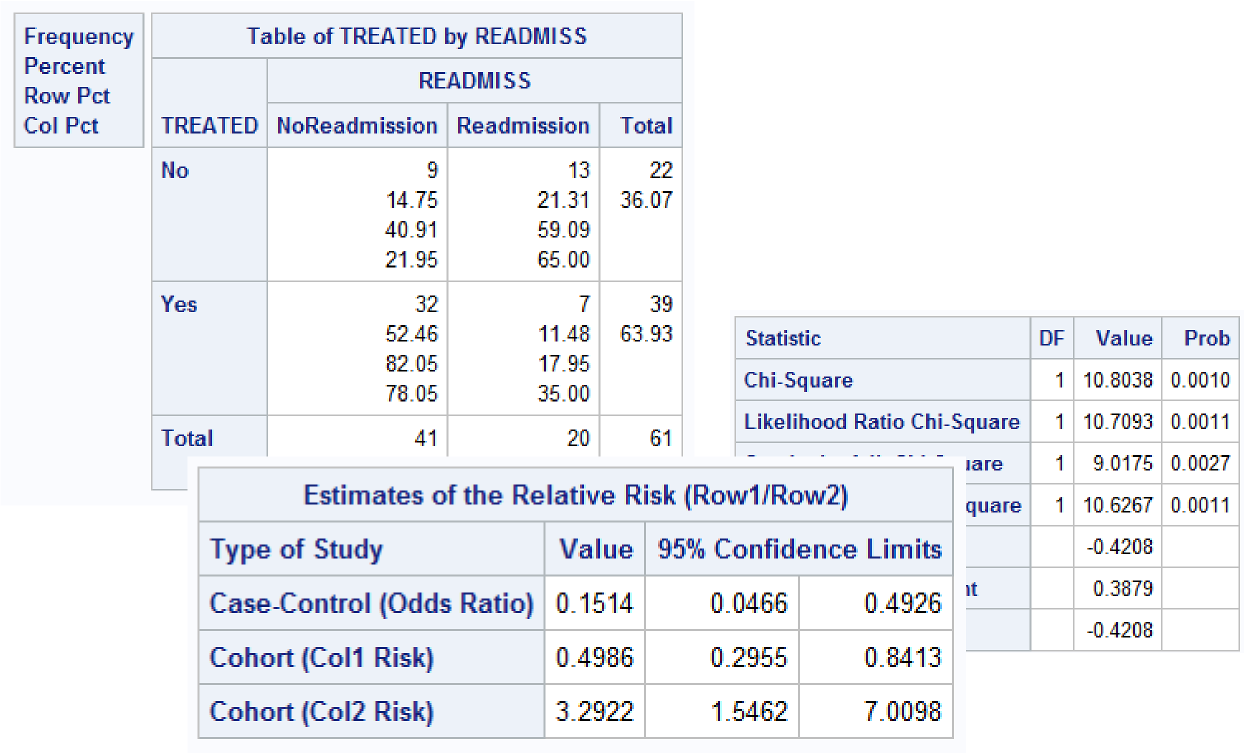
To get unadjusted measures of effect, start with PROC FREQ and investigate the measures as below.

```
proc freq data=temp;
  tables (sex treated prevhosp smkstat)*readmiss / chisq measures CL;
  format readmiss adm_fmt. treated trt_fmt. sex sex_fmt. prevhosp hosp_fmt.  smkstat smk_fmt.;
  title1 'Examine Unadjusted Associations By Readmission';
run;
```

CHISQ option in the tables statement computes several statistics including a Chi-square (known as Pearson chi-square test). It compares the observed frequencies with the expected frequencies collectively (considering the degree of freedom for each of the variables).

MEASURES option in the TABLES statement computes several statistics that describe the association between the row and column variables of the contingency table. In our case, we are interested in the odds ratio and relative risks.

CL option in the TABLES statement computes asymptotic confidence limits for all MEASURES



Frequency Percent Row Pct Col Pct	Table of TREATED by READMISS		
	READMISS		
TREATED	NoReadmission	Readmission	Total
No	9 14.75 40.91 21.95	13 21.31 59.09 65.00	22 36.07
Yes	32 52.46 82.05 78.05	7 11.48 17.95 35.00	39 63.93
Total	41	20	61

Statistic	DF	Value	Prob
Chi-Square	1	10.8038	0.0010
Likelihood Ratio Chi-Square	1	10.7093	0.0011
uare	1	9.0175	0.0027
quare	1	10.6267	0.0011
		-0.4208	
nt		0.3879	
		-0.4208	

Estimates of the Relative Risk (Row1/Row2)			
Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	0.1514	0.0466	0.4926
Cohort (Col1 Risk)	0.4986	0.2955	0.8413
Cohort (Col2 Risk)	3.2922	1.5462	7.0098

Note: because our 2x2 is set up with 0 in the top (no outcome) and left (no treatment), make sure you have the correct numbers in the formula places for the correct interpretation.

$$OR = \frac{ad}{bc} = \frac{7*9}{32*13} = 0.15$$ (which is the same as the case-control OR above)

$$RR = \frac{a/(a+b)}{c/(c+d)} = \frac{7/39}{13/22} = 0.30$$ (invert the cohort (col2risk = 3.29) to get the expected 0.30)

PROC LOGISTIC

Logistic regression is a statistical method used to evaluate many independent variables (X_1, X_2, …, X_p) in order to predict a dichotomous outcome. Generally this outcome is denoted as Y = 1 or Y = 0 for the two possibilities.

In logistic regression the probability of an occurrence of the outcome being investigated is defined as:

$$P(Y=1) = \frac{1}{1+\exp\left[-\beta_0 + \left(\sum_{k=1}^{p} \beta_k X_k\right)\right]}$$

SAS offers several procedures to estimate the binary logit model using ML estimation which include PROC LOGISTIC, PROC GENMOD, PROC PROBIT, and PROC CATMOD.

PROC LOGISTIC is a procedure for fitting linear regression models for binary or ordinal outcomes. The following is sample code for this procedure:

```
ods html path = 'c:\YourPath' body='Name.html';
proc logistic data=temp descending;
      class readmiss (ref = '0') treated (ref='0') sex (ref='0') prevhosp (ref='0') smkstat (ref='0')
              / param=reference;
      model readmiss=treated sex prevhosp smkstat / cl rl lackfit;
       title1 ‘Adjusted Odds of Readmission of Treated Compared to Non Treated’;
run;
ods html close;
```

ods html with the close after the run will send your output to an HTML file.

Data=temp names the input data set for the logistic regression.

Descending: The default in SAS is to model the probability that the dependent variable outcome of MI is equal to 0. The descending option allows us to model the probability that MI is equal to 1 and compares the probability of outcome to probability of no outcome for the odds ratio.

Class statement allows us to establish the reference category in the categorical variables without first making “dummy” variables in a data step. In this case, we are using reference cell coding.

Param=reference requests that the parameter estimates, odds ratios, and confidence intervals be calculated using reference cell coding. The default parameter estimates would be computed using the effect coding scheme which estimates the difference in the effect of each non-reference level compared to the average effect over the other levels of the variable.

CL= requests for each explanatory variable, the 95% (the default alpha level because the ALPHA= option is not invoked) Wald or profile likelihood confidence intervals for the odds ratios.

Lackfit requests the Hosmer-Lemeshow goodness of fit test for the model. The null hypothesis is that there is a good fit of the model to the observed data across the risk groups (we wish to fail to reject the null).

There are **MANY** options that are not discussed here and can be found at: https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_logistic_sect016.htm

PROC LOGISTIC output:

The LOGISTIC Procedure

Model Information	
Data Set	WORK.TEMP
Response Variable	READMISS
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	61
Number of Observations Used	61

Number of observations read and number of observations used is important to check to confirm the regression is running on the numbers you expect.

Note the probability modeled is your outcome = to 1.

Response Profile

Ordered Value	READMISS	Total Frequency
1	0	41
2	1	20

Probability modeled is READMISS=1.

Class Level Information

Class	Value	Design Variables	
TREATED	**0**	0	
	1	1	
SEX	**0**	0	
	1	1	
PREVHOSP	**0**	0	0
	1	1	0
	2	0	1

Confirm the reference category for the odds ratios are correct.

<u>Output not shown</u>: Not included in this paper is the AIC (Akaikes information criterion, lower is generally better), SC (Schwartz criterion which penalizes for more parameters then the AIC, lower is generally better), and the -2 log likelihood for the model fit statistics; the likelihood ratio, score, and Wald tests for testing whether all of the parameters taken together in the fitted model are equal to 0 when compared to the model with only the intercept; significance of each variable in its entirety (not categories of the variable) as well as the different categories.

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
10.7838	7	0.1483

Fail to reject the null and conclude that there is a good fit of the model to the observed data across the risk groups.

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
TREATED	1	5.9524	0.0147
SEX	1	0.3036	0.5817
PREVHOSP	2	2.1409	0.3428
SMKSTAT	2	1.2478	0.5359

Investigate the overall variable significance. In this case, we will include the non-statistically significant variables to control for possible confounding.

Odds Ratio Estimates and Wald Confidence Intervals

Effect	Unit	Estimate	95% Confidence Limits	
TREATED 1 vs 0	1.0000	**0.186**	**0.048**	**0.718**
SEX 1 vs 0	1.0000	1.564	0.318	7.684
PREVHOSP 1 vs 0	1.0000	2.842	0.360	22.446
PREVHOSP 2 vs 0	1.0000	3.878	0.617	24.384
SMKSTAT 1 vs 0	1.0000	0.850	0.123	5.853
SMKSTAT 2 vs 0	1.0000	2.299	0.351	15.046

Interpretation: After controlling for sex, previous hospitalization, and smoking status, those receiving the remote monitoring intervention were at 0.19 times the odds of being readmitted when compared to those who did not receive the intervention. This finding was statistically significant at the alpha=0.05 level (95% CI = 0.05, 0.72) because the confidence interval does not include 1.00.

PROC GENMOD

Use PROC GENMOD to produce adjusted Relative Risks with the Log Binomial:

```
proc genmod data=temp descending;
    class   treated sex prevhosp  smkstat  / param=reference ;
    model readmiss=treated sex prevhosp smkstat /  dist=binomial link=log lrci waldci aggregate ;
    estimate "treated" treated -1 1 / exp alpha=0.05;
    title1 'Binomial Regression for Adjusted RR for Readmission';
run;
```

Descending A very important point since version 8.1 came out is that when fitting a logistic regression using PROC GENMOD, the default now models the probability that the dependent variable readmiss is equal to 0. Versions prior to 8.1 modeled the higher level of the binary outcome variable (i.e. disease is present). Therefore, like PROC LOGISTIC, we use the descending option to model the probability that Y=1.

Class statement in GENMOD is the same as with PROC GLM and PROC ANOVA for determining which variables in the model will define categorical (classification) levels. These should be variables which code for terms such as replication id (in GEE), exposure level, etc. They can be character or numeric in value.

Dist=binomial option identifies the appropriate distribution for the data, in this case binomial. Other potential choices include Gaussian, Poisson, normal, gamma, inverse Gaussian, negative binomial (negbin), and multinomial (mult). If the DIST = option is omitted, SAS will assume the Gaussian distribution. Note: in this example we also used **Dist=Poisson** to calculate adjusted relative risks.

Link=log option refers to a transformation which is carried out on the responses prior to analysis, in this case the log. Other potential choices include identity, logit, power, probit, and complementary log log links. When this option is omitted, SAS will assume the identity link function resulting in no transformation.

Estimate will produce the estimated odds ratio for the exposure effect along with its associated standard error and confidence limits. The syntax for the ESTIMATE statement is exactly the same as that for the CONTRAST statement although the CONTRAST statement tests whether a linear combination of means is significantly different from 0. It should be mentioned that including the statement “lrci” and “waldci” after the link=log will produce wald and likelihood ratio confidence intervals about the parameter estimates.

The word between the quotes will label the output, and the variable name that comes after the label in quotes will call on the variable you wish to investigate. The contrasts in the input statement (–1 1 for treatment) are the same as the column of the class level information output from the logistic regression above. Therefore, the output relative risk will reflect the same comparisons as what was seen in PROC LOGISTIC.

Aggregate specifies the subpopulations on which the Pearson chi-square and the deviance are calculated and applies only to the multinomial distribution or the binomial distribution with binary (single trial syntax) response.

Exp after the backslash requests that the parameter estimates, standard error, and the confidence limits be computed and output.

There are **MANY** options that are not discussed here and can be found at: http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_genmod_sect022.htm

PROC GENMOD output:

```
          The GENMOD Procedure
          Model Information
Data Set                 WORK.TEMP
Distribution              Binomial
Link Function                  Log
Dependent Variable        READMISS

Number of Observations Read   61
```

The GENMOD output looks and feels very similar to the LOGISTIC output. Again, the number of observations read and number of observations used is important to check to confirm the regression is running on the numbers you expect.

Number of Observations Used	61
Number of Events	20
Number of Trials	61

Class Level Information

Class	Value	Design Variables	
TREATED	**0**	1	
	1	0	
SEX	**0**	1	
	1	0	
PREVHOSP	**0**	0	0
	1	1	0
	2	0	1
SMKSTAT	**0**	1	0
	1	0	1
	2	0	0

Class levels and the response profile are important to review to make sure they are what you expect for reference coding.

Note the probability modeled is your outcome = to 1.

Response Profile

Ordered Value	READMISS	Total Frequency
1	1	20
2	0	41

Additionally, the output shows the parameter information, criteria for assessing goodness of fit (including deviance, and log likelihood), and analysis of parameter estimates.

Contrast Estimate Results

Label	Mean Estimate	Mean Confidence Limits		L'Beta Estimate	Standard Error	Alpha	L'Beta Confidence Limits		Chi-Square	Pr > ChiSq
treated	0.3830	0.1541	0.9518	-0.9597	0.4644	0.05	-1.8700	-0.0494	4.27	0.0388
Exp(treated)				0.3830	0.1779	0.05	0.1541	0.9518		

Interpretation: After controlling for sex, previous hospitalization, and smoking status, those receiving the remote monitoring intervention were at 0.38 times the risk of being readmitted when compared to those who did not receive the intervention. This finding was statistically significant at the alpha=0.05 level (95% CI = 0.15, 0.95) because the confidence interval does not include 1.00.

Unlike the LOGISTIC procedure, the GENMOD procedure will not give the global test of the null hypothesis that all of the parameters taken together in the fitted model are equal to 0 when compared to the model with only the intercept. To calculate the likelihood ratio chi-square test, take the deviance (in output) from the reduced model (or null model if you remove all variables) and minus the deviance in the full model. This will give you a chi-square statistic with the degrees of freedom equal to the number of variables removed. PROC GENMOD does include an LSMEANS statement that provides an extension of least squares means to the generalized linear model.

PROC PHREG

The measure that is appropriate to use when we have differences in observed time comes from PHREG and Cox's Proportional Hazards Modeling. Cox introduced a new way of analyzing time-to-event data by making no assumptions about the baseline hazard of individuals and only assuming that the hazard functions of different individuals remained proportional and constant over time.

When there are several independent variables, and in particular when some of these are continuous, it is much more useful to use a regression method such as Cox rather than a Kaplan Meier approach.

Here, the hazard function for individual i is modeled as:

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

where $h_O(t)$ is the baseline hazard function, β's are regression coefficients, and x_i denote covariates.

The underlying or baseline hazard is the hazard when all covariates equal zero.

$$h(t,x) = h(t,0)e^{\beta' x}$$

$h(t,0)$ is the baseline hazard rate at time t for covariate vector 0. A subject's hazard at time t is proportional to the baseline hazard $h_O(t)$. The proportionality factor depends on the covariate vector for an individual. If all covariate values are homogenous, then it gets subsumed into the baseline hazard function.

The probability that an individual dies, leaves, etc., at time T_i, is given by:

$$\frac{e^{\beta' x_j}}{\sum e^{\beta' x_j}}$$

The conditioning eliminates the baseline hazard function.

Researchers favor Cox's proportional hazards modeling because of the robust semi-parametric method of calculating the probabilities of survival while simultaneously adjusting for other possibly influential variables. Other attractive features of Cox modeling include: relative risk type measure of association, no parametric assumptions, use of the partial likelihood function, and creation of survival function estimates.

Cox's semi-parametric modeling allows for no assumptions to be made about the parametric distribution of the survival times, making the method considerably more robust. Instead, the researcher must validate the assumption that the hazards are proportional over time. The proportional hazards assumption refers to the fact that the hazard functions are multiplicatively related. That is, their ratio is assumed constant over the survival time, thereby not allowing a temporal bias to become influential on the endpoint. In other words, the Cox proportional hazards model assumes that changes in the hazard of any subject over time will always be proportional to changes in the hazard of any other subject and to changes in the underlying hazard over time.

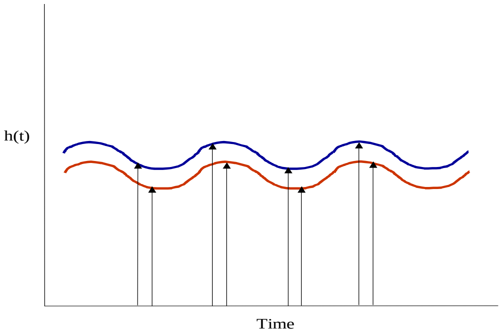


Figure 4. Graphical representation of proportional hazards over the follow-up period.

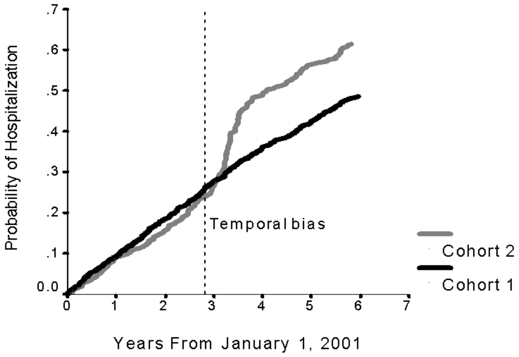


Figure 5. A cumulative distribution function that violates the proportional hazards assumption. Note the sharp increase in probability of hospitalization beginning right before the third year and lasting for approximately one year. After this one-year period the top curve then levels off and becomes parallel with the bottom curve once again (Smith AJE 2001).

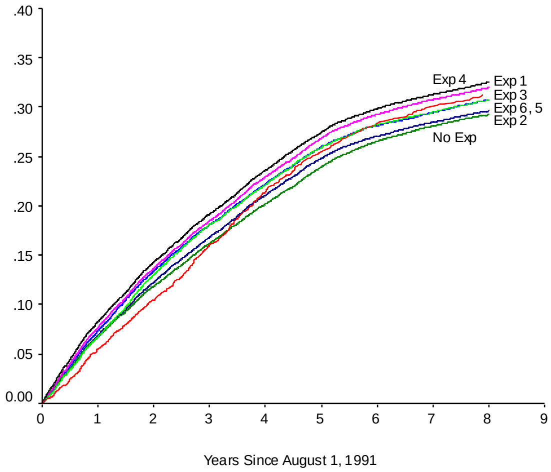


Figure 6: Stratified cumulative distribution functions of event by exposure status. Here, no violation of the assumption of proportional hazards occurs but a significant difference in the probability of event between the 7 exposure groups was observed (Smith AJE 2003).

```
proc phreg data=temp;
    class treated (ref='0') sex (ref='0') prevhosp (ref='0') smkstat (ref='0') /  param=ref;
    model survtime* readmiss(0) =  treated  x  sex  prevhosp smkstat / rl ties=efron;
    title1 'Cox Proportional Hazard Model Survival Differences by Treatment';
run;
```

Class statement in PHREG is the same as LOGISTIC and GENMOD for determining which variables in the model will define categorical (classification) levels.

RL requests the 95% (the default alpha level because the ALPHA= option is not invoked) confidence limits for the hazard ratios for each explanatory variable.

TIES=efron gives the researcher the approximations to the EXACT method without using the tremendous CPU it takes to run the EXACT method. Both the EFRON and the BRESLOW methods do reasonably well at approximating the EXACT when there are not a lot of ties. If there are a lot of ties, then the BRESLOW approximation of the EXACT will be very poor. If the time scale is not continuous and is therefore discrete, the option TIES=DISCRETE should be used.

The PHREG Procedure

Model Information	
Data Set	WORK.TEMP
Dependent Variable	SURVTIME
Censoring Variable	READMISS
Censoring Value(s)	0

The PHREG output looks and feels very similar to the previous GENMOD and LOGISTIC output. Again, the number of observations read and number of observations used is important to check to confirm the regression is running on the numbers you expect.

Model Information	
Ties Handling	EFRON

Number of Observations Read	61
Number of Observations Used	61

Class Level Information

Class	Value	Design Variables	
TREATED	**0**	0	
	1	1	
SEX	**0**	0	
	1	1	
PREVHOSP	**0**	0	0
	1	1	0
	2	0	1
SMKSTAT	**0**	0	0
	1	1	0
	2	0	1

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
61	20	41	67.21

Check the number of events and censored values are correct.

Type 3 Tests

Effect	DF	Wald Chi-Square	Pr > ChiSq
TREATED	1	6.7046	0.0096
SEX	1	0.2641	0.6073
PREVHOSP	2	5.0159	0.0814
SMKSTAT	2	7.9201	0.0191

Investigate the overall p-values.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
TREATED	**1**	1	-1.57255	0.60732	6.7046	0.0096	0.208	0.063	0.682	TREATED 1
SEX	**1**	1	0.30129	0.58630	0.2641	0.6073	1.352	0.428	4.265	SEX 1

Analysis of Maximum Likelihood Estimates										
Parameter		**DF**	**Parameter Estimate**	**Standard Error**	**Chi-Square**	**Pr > ChiSq**	**Hazard Ratio**	**95% Hazard Ratio Confidence Limits**		**Label**
PREVHOSP	**1**	1	2.19612	0.99507	4.8709	0.0273	8.990	1.279	63.207	PREVHOSP 1
PREVHOSP	**2**	1	1.01766	0.67331	2.2844	0.1307	2.767	0.739	10.354	PREVHOSP 2
SMKSTAT	**1**	1	-0.54530	0.80930	0.4540	0.5004	0.580	0.119	2.832	SMKSTAT 1
SMKSTAT	**2**	1	1.53051	0.68084	5.0533	0.0246	4.621	1.217	17.548	SMKSTAT 2

Interpretation: After controlling for sex, previous hospitalization, and smoking status, AND taking into account time, those receiving the remote monitoring intervention were at 0.21 times the risk of being readmitted when compared to those who did not receive the intervention. This finding was statistically significant at the alpha=0.05 level (95% CI = 0.06, 0.68) because the confidence interval does not include 1.00.

***If you wanted to test the time interaction you could run the following code. This is often important to do first in order to include a time dependent variable and then extend the Cox model to use these types of variables.

```
proc phreg data=temp;
class treated (ref='0') sex (ref='0') prevhosp (ref='0') smkstat (ref='0') / param=ref;
     model survtime*readmission(0) =  treated  x  sex  prevhosp smkstat / rl ties=efron;
     x=treated*(log(survtime) - (log(mean survival)));
     title1 ‘Cox Regression of Treatment Status, Investigate Proportional Hazards Assumption by
    Testing for Interaction’;
  run;
```

x=treated*(log(survtime)-(log(mean survival))) tests the interaction of treatment with time to determine if the proportional hazards assumption is met. You can get the mean survival from KM. If x is not significant, you can conclude that the proportional hazards assumption is met and remove the variable from the model.

Graphical investigation of proportional hazards can be accomplished after data are stratified by treatment status in order to compute the survivor function estimates for the two treatment arms. Using the BASELINE function in PROC PHREG, you can output the survivor function estimates. The survival curves can then be displayed or tone can compute the cumulative distribution function for the separate treatment arms over the study period.

Graphing using PROC GPLOT

```
proc sort data=temp;
    by treated;
```

```
 proc phreg data=temp;
     by treated;
     class treated (ref='0') sex (ref='0') prevhosp (ref='0') smkstat (ref='0') /  param=ref;
     model survtime*censor(0) =  treated  x  sex  prevhosp smkstat / rl ties=efron;
    baseline out=surv1 survival=s ;
     title1 'Cox Proportional Hazard Model Survival Differences by Treatment';
run;


   options ps=52;
   goptions device=win;

   symbol1 line=1 color=blue value=square i=join;
   symbol2 line=2 color=red value=star i=join;

   proc gplot data= surv1;
     plot survtime*s=treated;
     title1 font=swissb 'Cox Proportional Hazard Model' ;
     title2 font=swissb h=1.5 'Survival Differences by Treatment';
   run;
```

BY stratifies the analysis by the categories in the by variable, after data are sorted in that manner.

BASELINE without the COVARIATES= option produces the survival function estimates corresponding to the means of the explanatory variables for each stratum.

OUT=surv1 names the data set output by the BASELINE option.

SURVIVAL=s tells SAS to produce the survival function estimates in the output data set.

TEST statement allows testing of subgroups of regression coefficients. This statement is not shown above but can be done with “test age, occupation;” after the model statement. This test statement will test the null hypothesis that age and occupation taken together are not related to probability of event after adjusting for the other variables in the model. This statement is also useful when testing the global significance of a categorical variable in which the model statement expresses only the dummy variables.

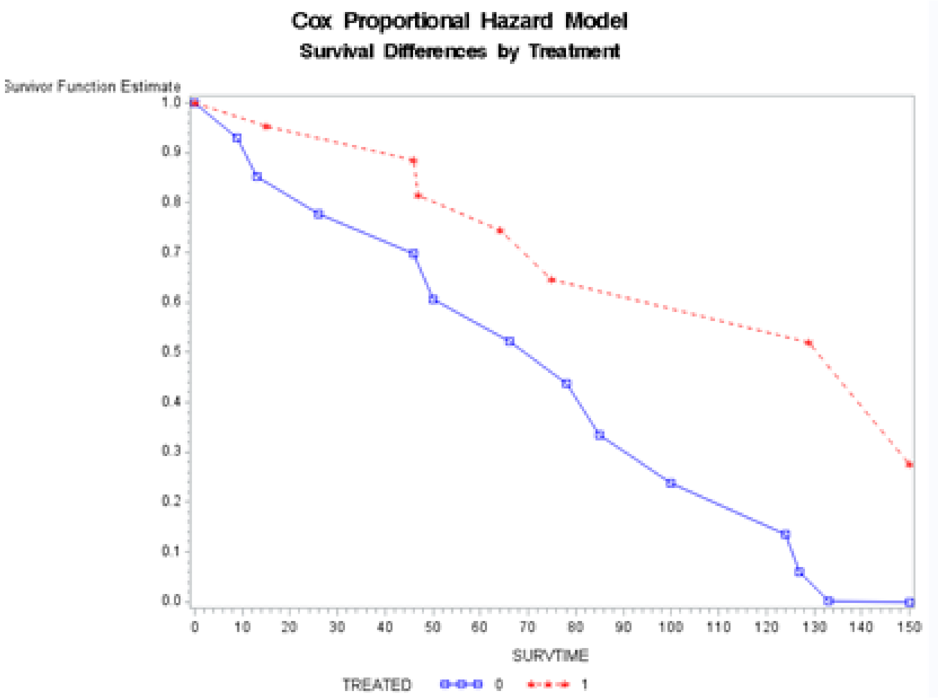


Figure 7. The stratified treatment arm survival curves over the follow-up period.

Likelihood Ratio Test

This test makes use of the log likelihood value given by the –2logL in the SAS output. If the researcher would like to see the importance of a variable or a group of variables in the model they should run a full and a reduced model. The full model includes all the variables and the reduced model removes the variable or variables you would like to inspect. Taking the difference of the two values will yield a test statistic having a chi-square distribution under the null hypothesis with the number of degrees of freedom equal to the number of variables removed from the model.

Computing The Generalized R^2

It may be helpful to compute the $\mathbf{R}^2$ value for the Cox model. Although it is not an option of PROC PHREG, the $\mathbf{R}^2$ value can be computed from the output of the regression.

$R^2 = 1 - \exp(-LR / n)$

Where LR is the likelihood-ratio chi-square statistic for testing the null hypothesis that all variables included in the model have coefficients of 0, and n is the number of observations. The researcher needs to take extreme caution when comparing the R^2 values of Cox regression models. Remember from linear regression analysis, R^2 can be artificially increased by simply adding explanatory variables to the regression model (i.e.; more variables do not equal a better model necessarily). Also, the above computation does not give the proportion of variance of the dependent variable explained by the independent variables as it would in linear regression, but it does give a measure of how associated the independent variables are with the dependent variable.

SUMMARY OF THE MEASURES

Unadjusted OR	**= 0.151**
Unadjusted RR	**= 0.304**
Adjusted OR Logistic	**= 0.186**
Adjusted RR Binomial	**= 0.382**
Adjusted RR Poisson	**= 0.383**
Adjusted HR Cox	**= 0.208**

SUMMARY

In summary, this paper was designed to be an intermediate level review of three measures of association commonly used in health outcomes research (the odds ratio, relative risk, and hazard ratio) and should be a starting point for the theory, programming, and interpretation. The SAS statistical procedures were presented, however, in each case, additional model diagnostics, collinearity investigations, and tests need to be run to validate the approach taken by the researcher. The measures were similar and ranged from an OR of .15 unadjusted to a RR from the Poisson that was 0.38. In each case the intervention appeared to significantly impact readmissions though the magnitude of effects were very different. Significant care should be given to the correct measure to be used based on the data and study design.

RECOMMENDED READING

Allison, Paul D., *Survival Analysis Using the SAS® system: A Practical Guide*, Cary, NC: SAS Institute Inc., 1995.

Altman,D.G., Deeks,J.J. & Sackett,D.L. 1998. Odds ratios should be avoided when events are common. *BMJ* **317,** 1318.

Cox DR. *Regression models and life tables* (with discussion). J R Stat Soc [Ser B] 1972;B34:187-220.

Cox, D. R. & Snell, E. J. *The Analysis of Binary Data,* Second Edition, London: Chapman and Hall; 1989.

Deddens,J.A. & Petersen,M.R. 2004. Re: "Estimating the relative risk in cohort studies and clinical trials of common outcomes". *Am J Epidemiol* 159**,** 213-214.

Deeks,J. 1998. When can odds ratios mislead? Odds ratios should be used only in case-control studies and logistic regression analyses. *BMJ* 317**,** 1155-1156.

Hosmer JR. DW, Lemeshow S. *Applied Survival Analysis; Regression Modeling of Time to Event Data.* New York: John Wiley & Sons; 1999.

McNutt,L.A., Wu,C., Xue,X. & Hafner,J.P. 2003. Estimating the relative risk in cohort studies and clinical trials of common outcomes. *Am J Epidemiol* 157**,** 940-943.

SAS Institute Inc. SAS/STAT® Software version 9.0. Cary, NC: SAS Institute Inc., 2002.

Smith TC, Gray GC, Knoke JD. Is systemic lupus erythematosus, amyotrophic lateral sclerosis, or fibromyalgia associated with Persian Gulf War service? An examination of Department of Defense hospitalization data. Amer J Epidemiol, 2000 Jun; 151(11): 1053-1059.

Smith TC, Gray GC, Weir JC, Heller JM, Ryan MAK. *Gulf War Veterans and Iraqi Nerve Agents at Khamisiyah. Postwar Hospitalization Data Revisited*. Am J Epidemiol, 2003; 158: 457-467.

Therneau TM, Grambsch PM. Modeling survival data: extending the Cox model. New York: Springer-Verlag; 2000.

Zhang,J. & Yu,K.F. 1998. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *JAMA* 280**,** 1690-1691.

Zou,G. 2004. A modified poisson regression approach to prospective studies with binary data. *Am J Epidemiol* 159**,** 702-706.

CONTACT INFORMATION

Dr. Besa Smith has worked in government, academic, and private industries and has served as a senior epidemiologist, senior biostatistician, and head of analytics for a 35-40 member multi-disciplinary research team. She is currently a senior scientist and founder of the health analytics consulting business, Analydata. Additionally, Dr. Smith has joint appointments with National University and the University of California, San Diego. She is an adjunct professor in the Department of Community Health in the School of Health and Human Services at NU and an assistant adjunct professor in the Department of Family and Preventive Medicine in the School of Medicine at UCSD. She teaches epidemiology and biostatistics courses to undergraduate, graduate, and medical students. Dr. Smith has a BS in biology; MPH in biometry, and PhD in epidemiology. With over 15 years leveraging health analytics in longitudinal studies and medical health outcomes research, she has >70 peer-reviewed publications in scientific journals and >100 scientific presentations.

Besa Smith, MPH, PhD
Epidemiologist and Biostatistician
Analydata
San Diego, CA 92107
besasmith@analydata.com

Dr. Tyler Smith is professor of biostatistics, epidemiology, public health and health informatics; and program lead for the Health Analytics master's degree. Dr. Smith received a BS in mathematics/statistics from California State University, Chico; MS in statistics from the University of Kentucky; and PhD in epidemiology from the University of California, San Diego. With ~20 years of experience in health research leading large longitudinal studies, infant health registries, and medical health outcomes research, he has 120 peer-reviewed publications in scientific journals, >250 scientific presentations and has been PI/COI on grants totaling >$20,000,000. Currently Dr. Smith serves the SAS community through his efforts as Content Area Lead for SAS Global Forum 2014; 2015 SAS Global Forum Conference Chair; Junior Professional Award co-Chair for Western User's of SAS Software, and as part of the Executive Board for the San Diego SAS User's Group.

Tyler C Smith, MS, PhD
Associate Professor and Chair
Program Lead MS Health and Life Science Analytics
Director Health Science Research Center
Department of Community Health
School of Health and Human Services
National University
San Diego, CA 92123
tsmith@nu.edu