Paper 244-26

Survival Analysis And The Application Of Cox's Proportional Hazards Modeling Using SAS

Tyler Smith, and Besa Smith, Department of Defense Center for Deployment Health Research, Naval Health Research Center, San Diego, CA

Abstract

In recent papers published in the American Journal of Epidemiology, the authors used Cox's proportional hazards regression modeling to model the time until an event of interest and compare the cumulative probability of hospitalization over time for two or more cohorts while adjusting for other influential covariates. In this presentation these statistical procedures will be looked at more closely by using SAS. The usefulness of the baseline option in PROC PHREG will be demonstrated with the creation and output of survival function estimates, which are a function of the cumulative probability estimates over time.

This real data example using Cox modeling will show what increased risk for hospitalization for an event of interest might look like graphically and in a risk of event type ratio. Further analyses using the survivor function estimates will look for graphical representation of a temporal bias during the observation period. The SAS system's PROC PHREG with baseline option was instrumental in searching for possible temporal biases and dealing with attrition of subjects over our study period.

Introduction to Survival Analysis

The term "survival analysis" pertains to a statistical approach designed to take into account the amount of time an experimental unit contributes to a study. That is, it is the study of time between entry into observation and a subsequent event. Originally, the event of interest was death hence the term, "survival analysis." The analysis consisted of following the subject until death. The uses in the survival analysis of today vary quite a bit. Applications now include time until onset of disease, time until stockmarket crash, time until equipment failure, time until earthquake, and so on. The best way to define such events is simply to realize that these events are a transition from one discrete state to another at an instantaneous moment in time. Of course, the term "instantaneous", which may be years, months, days, minutes, or seconds, is relative and has only the boundaries set by the researcher.

The History of Survival Analysis

The origin of survival analysis goes back to mortality tables from centuries ago. However, it was not until World War II that a new era of survival analysis emerged. This new era was stimulated by interest in reliability (or failure time) of military equipment. At the end of the war these newly developed statistical methods emerging from strict mortality data research to failure time research, quickly spread through private industry as customers became more demanding of safer, more reliable products. As the uses of survival analysis grew, parametric models gave way to nonparametric and semiparametric approaches for their appeal in dealing with the ever-growing field of clinical trials in medical research. Survival analysis was well suited for such work because medical intervention follow-up studies could start without all experimental units enrolled at start of observation time and could end before all experimental units had experienced an event. This is extremely important because even in the best-developed studies, there will be subjects who choose to quit participating, who move too far away to follow, or who will die from some unrelated event. The researcher was no longer forced to withdraw the experimental unit and all associating data from the study, instead techniques called censoring enabled researchers to analyze incomplete data due to delayed entry or withdrawal from the study. This was important in allowing each experimental unit to contribute all of the information possible to the model for the amount of time the researcher was able to observe the unit.

The last great strides in the application of survival analysis techniques has been a direct result of the

availability of software packages and high performance computers which are now able to run these difficult and computationally intensive algorithms relatively efficiently.

Some Tools Used in Survival Analysis

First, recall that time is continuous, which results in the probability of an event at a single point of a continuous distribution being zero. We are challenged to define the probability of these events over distribution. This is best described by graphing the distribution of event times. To ensure the readers will start with the same fundamental tools of survival analysis, a brief descriptive section of these important concepts will follow. A more detailed description of the probability density function, the cumulative distribution function, the hazard function, and the survivor function, can be found in any intermediate level statistical textbook.

So that the reader will be able to look for certain relationships while reading, it is important to note before the brief descriptions the one-to-one relationship that these four functions possess. The pdf can be obtained by taking the derivative of the cdf and likewise, the cdf can be obtained by taking the integral of the pdf. The survivor function is simply 1 minus the cdf. Which leaves the hazard function as simply being the pdf over the survivor function. It will be these relationships later that will allow us to calculate the cdf from the survivor function estimates that the SAS procedure PROC PHREG will output.

The Cumulative Distribution Function

The cumulative distribution function (cdf) is very useful in describing the continuous probability distribution of a random variable, such as time, in a survival analysis. The cdf of a random variable T, denoted $F_T(t)$, is defined by $F_T(t) = P_T(T \leq t)$. This is interpreted as a function that will give the probability that the variable T will be less than or equal to any value t that we choose. Several properties of a distribution function F(t) can be listed as a consequence of the knowledge of probabilities. Because F(t) has the probability $0 \leq F(t) \leq 1$, then F(t) is a nondecreasing function of t, and as t approaches ∞, F(t) approaches 1.

The Probability Density Function

The probability density function (pdf) is also very useful in describing the continuous probability distribution of a random variable. The pdf of a random variable T, denoted $f_T(t)$, is defined by $f_T(t) = d F_T(t) / dt$. That is, the pdf is the derivative or slope of the cdf. Every continuous random variable has its own density function, the probability $P(a \leq T \leq b)$ is the area under the curve between times a and b.

The Survival Function

Let $T \geq 0$ have a pdf f(t) and cdf F(t). Then the survival function takes on the following form:

$$S(t) = P\{T > t\}$$

$$= 1 - F(t)$$

That is, the survival function gives the probability of surviving or being event-free beyond time t. Because S(t) is a probability, it is positive and ranges from 0 to 1. It is defined as S(0) = 1 and as t approaches ∞, S(t) approaches 0. The Kaplan-Meier estimator, or product limit estimator, is the estimator used by most software packages because of the simplistic step idea. The Kaplan-Meier estimator incorporates information from all of the observations available, both censored and uncensored, by considering any point in time as a series of steps defined by the observed survival and censored times. The survival curve describes the relationship between the probability of survival and time.

The Hazard Function

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

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

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

$$= f(t) / S(t)$$

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 somewhere between t and $t + \Delta$, divided by the probability that the individual survived beyond time t. The hazard function seems to be more intuitive to use in

survival analysis than the pdf because it attempts to quantify the instantaneous risk that an event will take place at time t given that the subject survived to time t.

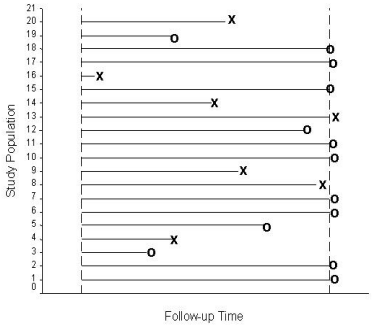
Incomplete Data

Observation time has two components that must be carefully defined in the beginning of any survival analysis. There is a beginning point of the study where time=0 and a reason or cause for the observation of time to end. For example, in a complete observation cancer study, observation of survival time may begin on the day a subject is diagnosed with the cancer and end when that subject dies as a result of the cancer. This subject is what is called an uncensored subject, resulting from the event occurring within the time period of observation. Complete observation time data like this example are desired but not realistic in most studies. There is always a good possibility that the patient might recover completely or the patient might die due to an entirely unrelated cause. In other words, the study cannot go on indefinitely, waiting for an event from a participant and unforeseen things happen to study participants that make them unavailable for observation. The censoring of study participants therefore deals with the problems of incomplete observations of time due to assumed ***random*** factors not related to the study design. **Note:** This differs from truncation where observations of time are incomplete due to a selection process inherent to the study design.

Left and Right Censoring

The most common form of incomplete data is right censoring. This occurs when there is a defined time (t=0) where the observation of time is started for all subjects involved in the study. A right censored subject's time terminates before the outcome of interest is observed. For example, a subject could move out of town, die of an unexpected cause, or could simply choose not to participate in the study any longer. Right censoring techniques allow subjects to contribute to the model until they are no longer able to contribute (end of the study, or withdrawal), or they have an event. Conversely, an observation is left censored if the event of interest has already occurred when observation of time begins. For the purposes of this study we focused on right censoring.

The following graph shows a simple study design where the observation times start at a consistent point in time (t=0). The X's represent events and the O's represent censored observations. Notice that all observations are classified with an event, or they are censored at time of separation or at the end of the study period. Some subjects have events early in the study period and others have events at the end of the study period. Likewise some subjects leave early, but most do not have an event during the entire study and are simply right censored at the end. There is no need for left censoring or truncation techniques in this simple example.



Time Dependencies

In some situations the researcher may find that the dynamical nature of a variable causes changes in value over the observation time. In other instances the researcher may find that certain trends effect the probability of the event of interest over time. There are easy ways to test and account for these temporal biases within PROC PHREG but be careful if you have a large number of observations as the computation of the subsequent partial likelihood is very taxing and time consuming. An easier way to see if there are existing temporal biases is to look at the plots of the cumulative distributions of the probability of event. If there is a steep increase or decrease in the cumulative probability, it may suggest more investigation is needed. It is important to note here that when a time dependent variable is introduced into the model, the ratios of the hazards will not remain steady. This only effects the model structure. We will still be doing a Cox regression but instead the model used is called the extended Cox model.

The Studies

The underlining purpose of these studies was to investigate the effect of a specific exposure on an outcome of interest. Then we sought to identify 2 or more cohorts who might have had this exposure or some degree of exposure and compare the hospitalization experiences for certain outcomes of interest to another similar cohort without that particular exposure.

Demographic data

Demographic data available for analysis included social security numbers (for linking purposes only), gender, date of birth, race, ethnicity, home of record, marital status, military occupational status, military pay grade, length of military service, deployment status, salary, date of separation from military service, military service branch, and exposure status.

Hospitalization Data

Data describing hospitalization experiences were captured from all United States Department of Defense military treatment facilities for the period of October 1, 1988, through December 31, 1999. The actual observation period varied by study. Removal of personnel with diagnoses of interest prior to the start of the study follow-up period was completed. These data included date of admission in a hospital and up to eight discharge diagnoses associated with the admission to the hospital. Additionally, a preexposure period covariate (coded as yes or no) was used to reflect a hospital admission during the 12 months prior to the start of the exposure period. Note: the exposure period was the year from August 1, 1990, to August 1, 1991. Diagnoses were coded according to the *International Classification of Diseases, Ninth Revision* (ICD-9). For these analyses, we scanned for the specific 3-,4-, or 5-digit component of the ICD-9 diagnoses.

Observation Time

The focus of each study was to see if a certain exposure or lack of exposure had any influence on the targeted disease outcomes. For each subject, hospitalizations (if any) were scanned in chronological order and diagnostic fields were scanned in numerical order for the ICD-9 codes of interest. Only the first hospitalization meeting the outcome criteria was counted for each subject. Subjects were classified as having an event if they were hospitalized in any Department of Defense hospital facility worldwide with the targeted diagnoses, and as censored otherwise. Observation time varied with the dates we chose to start and end observation but was calculated from the start of follow-up until event, separation from military service, or the end of the study period, whichever occurred first. Subjects were allowed to leave the study and assumed a random early departure distribution. Delayed entry and events occurring before the start date of the study were not a concern, therefore only right censoring was needed to allow for the random early departure of subjects (see previous graph).

Cox's Proportional Hazards Regression

There are several reasons Cox's proportional hazards modeling was chosen to explain the effect of covariates on time until event. They are discussed below and include: the relative risk, no parametric assumptions, the use of the partial likelihood function, and the creation of survivor function estimates.

Relative Risk

The simple interpretation given by the Cox model as "relative risk" type ratio is very desirable in explaining the risk of event for a certain covariate. For example, when we have a two-level covariate with a value of 0 or 1, the hazard ratio becomes e^{β}. If the value of the coefficient is $\beta = \ln(3)$ then it is simply saying that the subjects labeled with a 1 are three times more likely to have an event than the subjects labeled with a 0. In this way we had a measure of difference between our exposure cohorts instead of simply knowing whether they were different.

No Parametric Assumptions

Another attractive feature of Cox regression is not having to choose the density function of a parametric distribution. This means that Cox's semiparametric 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 only 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 survival time,

thereby not allowing a temporal bias to become an influential player on the endpoint.

Use of the Partial Likelihood Function

The Cox model has the flexibility to introduce time-dependent explanatory variables and handle censoring of survival times due to its use of the partial likelihood function. This was important to our study in that any temporal biases due to differences in hospitalization practices for different strata of the significant covariates over the years of study needed to be handled correctly. This ensured that any differences in hospitalization experiences between the exposed and nonexposed would not be coming from these temporal differences.

Survivor Function Estimates

With the SAS option BASELINE, a SAS dataset containing survival function estimates can be created and output. These estimates correspond to the means of the explanatory variables for each stratum.

Analysis

Univariate Analyses

Using PROC FREQ, and PROC UNIVARIATE, an initial univariate analysis of the demographic variables crossed with hospitalization experience was carried out to determine possible significant explanatory variables to be included in the model runs. All variables with a chi-square value or t statistic of .15 or less were considered possibly significant and were therefore retained for the model analysis. Additionally the distributions of attrition were checked to see if the cohorts separated from active duty military service equally.

Modeling Approach

Using PROC PHREG, a saturated Cox model was run after creating dummy variables, necessary for the output of hazard ratios for the categorical explanatory variables. A manual backward stepwise analysis was carried out to create a model with statistically significant effects of explanatory variables on survival times.

Programming

```
PROC PHREG DATA=ANALYDAT;
  MODEL INHOSP*CENSOR(0)= expose1 pwhsp
  status1 sex1 age1-age3 ms1 paygr1-paygr2
  oc_cat1-oc_cat9 ccep
            /RL TIES=EFRON ;
TITLE1 'Cox Regression With Exposure Status In
the Model ';
RUN;
```

The options used in this survival analysis procedure are described below:

DATA=ANALYDAT names the input data set for the survival analysis.

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

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.

Stratification By Exposure Status

These data were then stratified by exposure and the models were run with the exposure flag covariate withdrawn from the model. This allowed for inspection of interaction between exposure status and covariates. Running these separate models also allowed for the computation of survival function estimates using the BASELINE function in PROC PHREG. The survival curves (which are really step functions for such numerous events that they appear continuous) were now available to compute the cumulative distribution function for the separate cohorts.

Time Dependent Covariates

After the final model of significant explanatory variables was created, it was necessary to validate the proportional hazards assumption. If the researcher believes that there may be a time dependency from a certain variable then simply add

x1time to the list of independent variables and the following below the model statement.

x1time=x1*(t)

Where t is the time variable and x1 is the suspected time dependent variable.

If the interaction term is found to be insignificant we can conclude that the proportional hazards assumption holds. This is necessary to ensure that there was no adverse effect from time-dependent covariates creating different rates for different subjects, thus making the ratios of their hazards nonconstant.

Survivor Function Estimates By Exposure

The following is the code used after the ANALYDAT was stratified into exposed or nonexposed. This produces the survivor function estimates by exposure while simultaneously checking to see if there were any interactions between the covariates and the exposure status.

```
PROC PHREG DATA=EXPOSE1;
  MODEL INHOSP*CENSOR(0)=pwhsp status1
  sex1 age1-age3 ms1 paygr1-paygr2 oc_cat1-
  oc_cat9 ccep
            /RL TIES=EFRON ;
  BASELINE OUT=SURVS SURVIVAL=S;
RUN;
```

The new options used in this survival analysis procedure are described below:

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

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

SURVIVAL=S tells SAS to produce the survivor function estimates in the output data set.

A simple calculation of 1-Survivor function estimates in SURVS, obtained from running the BASELINE option, produced the cumulative distribution functions. We could now see the cumulative probability estimates of hospitalization over time. We were then able to visually scan for differences in hospitalization experiences of between the cohorts and look for insight as to whether or not the proportional hazards assumption had been violated.

The Plots

Figure 1 is what the cumulative distribution function would look like if there were a violation of the proportional hazards assumption. Note the sharp increase in probability of hospitalization beginning right before the third year and lasting for approximately 1 year. After this one year period the top curve then levels off and becomes parallel again with the bottom curve.

Figure 1.

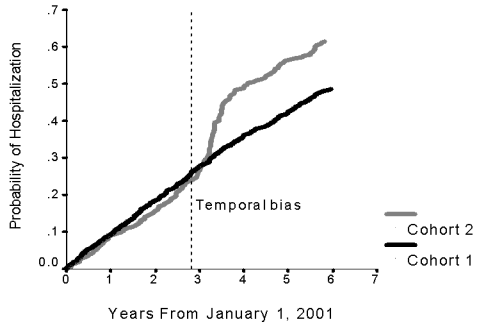


Figure 2 shows what the cumulative distribution function would look like if there were no violation of the assumption of proportional hazards but there did happen to be an observed significant difference in the disease experience between the two cohorts over the length of the study period.

Figure 2.

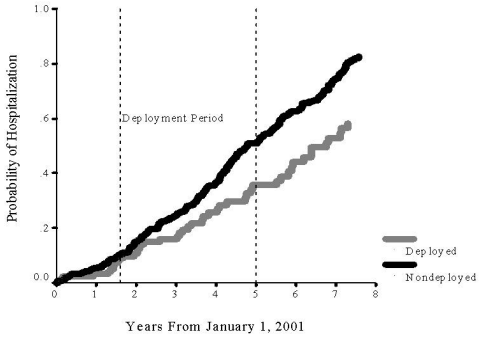
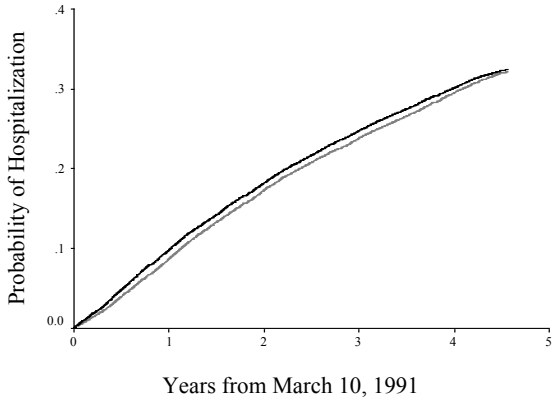


Figure 3 shows what the cumulative distribution function would look like if there were no problem with the assumption of proportional hazards. The figure also shows what the curves would look like it there were not a significant difference observed between the diagnosis experience of the two cohorts.

Figure 3.



Computing the Generalized R^2

Recently I was asked whether SAS computed the $\mathbf{R^2}$ value and what it was for that particular model. If the researcher desires, the $\mathbf{R^2}$ value can be computed easily from the output of your regression, although it is not an option of PROC PHREG. Simply compute

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

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 (ie; more variables does not equal a better model necessarily). Also, the above computation does not give you the proportion of variance of the dependent variable explained by the independent variables as it would in linear regression, but does give you a measure of how associated the independent variables are with the dependent variable.

Residual Analysis

A residual analysis is very important especially if the sample size is relatively small. Add the following after your model statement to output the martingale and deviance residuals:

```
BASELINE OUT=SURVS SURVIVAL=S
XBETA=XBET RESMART=MARTING
RESDEV=RDEV;
```

Then a simple plot of the residuals against the linear predictor scores will give the researcher an idea of the fit or lack of fit of the model to individual observations.

```
PROC GPLOT DATA=SURVS;
  PLOT (MARTING RDEV) * XBET / VREF=0;
  SYMBOL1 VALUE=CIRCLE;
```

Results

Using the initial univariate comparisons for events occurring during the study, the following variables were selected for the subsequent model analyses: gender, age group, marital status, race/ethnicity, military occupational category, military pay grade, salary, service branch, pre-exposure period hospitalization, and exposure status. Home of record was not shown to be significantly affecting the endpoints in any either of the studies' models and was dropped. Salary and length of service were dropped from analyses due to colinearity with age.

The subjects in cohort 1 had similar risks for two of the three diseases during the August 1, 1991, to July 31, 1997 study period compared with subjects in cohort 2. The corresponding cumulative probability plots (above) were nearly parallel for the follow-up period. However, the Cox model did reveal some consistently better predictors of hospitalization with the two diseases, which included female gender, preexposure period hospitalization, enlisted pay grade, and US Reserve service type.

The subjects in cohort 1 had significantly different risks for the third disease during the August 1, 1991, to July 31, 1997 study period compared with subjects in cohort 2. The corresponding cumulative probability plots (above) were nearly parallel for the first three years of follow-up, then there was a drastic increase in hospitalization for a period of about 1 year and then once again the curves became nearly parallel again. Time-dependent

variables included in the modeling also confirmed this result. The hazards ratio was not significantly greater than 1 for the first 3 years of follow-up and the last 3 years as well. However, during the 1 year in question, though, the risk of hospitalization with that particular disease was almost 3 times that of cohort 1. Further investigation found that certain treatment facilities had adopted an approach of administratively hospitalizing these subjects for extensive clinical evaluations. This approach was later dropped almost 1 year to the date of the start.

Conclusions

The Cox proportional hazard model's robust nature allows us to closely approximate the results for the correct parametric model when the parametric is unknown or in question. Using the SAS® system procedure PROC PHREG, Cox's proportional hazards modeling was used to compare the hospitalization experiences of two or more cohorts. The two studies produced models suggesting no increase in risk among the exposed which were later confirmed by producing the cumulative distribution function. The study also found at least one model which initially suggested an increase in risk among the exposed. Further analysis revealed that this sharp increase in risk for approximately 1 year was likely due to an outside factor affecting the process of hospitalizing personnel for this particular disease event.

The SAS® system's PROC PHREG with censoring and the baseline option is a powerful tool for handling early departure of subjects during the study period. It is also useful for producing data sets, including survival function estimates, which can be used in a simple equation to produce estimates of probability of events. When graphed, these show cumulative probability of event curves as a function of time. If it were not for the graphs of the cumulative distribution functions, which showed a sharp temporal bias, results may have been reported and interpreted with much different results.

References

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

Kleinbaum DG, *Survival Analysis: A self-Learning Text.* New York: Springer-Verlag; 1996

SAS Institute Inc., SAS/STAT® *User's Guide, Version 6, Fourth Edition, Volume 1,* Cary, NC: SAS Institute Inc., 1989. 943 pp.

SAS Institute Inc., SAS/STAT® *User's Guide, Version 6, Fourth Edition, Volume 2,* Cary, NC: SAS Institute Inc., 1989. 846 pp.

SAS Institute Inc. SAS/STAT® *Software: Changes and Enhancements through Release 6.11*. Cary, NC: SAS Institute Inc., 1996. 1104 pp.

Allison, Paul D., *Survival Analysis Using the SAS® system: A Practical Guide*, Cary, NC: SAS Institute Inc., 1995. 292 pp.

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.* Amer J of Epidemiol; June 1, 2000, Vol 151.

Gray GC, Smith TC, Knoke JD, Heller JM. *The postwar hospitalization experience among Gulf War veterans exposed to chemical munitions destruction at Khamisiyah, Iraq.* Amer J of Epidemiol; September 1, 1999 - Volume 150 No 5.

Acknowledgments

Thank you to Navy CAPT Greg Gray, Director of the Department of Defense Center for Deployment Health Research at the Naval Health Research Center, San Diego. CAPT Gray's support and encouragement of learning new concepts and devising better ways to tackle statistical problems in pursuit of the best possible research answers.

Approved for public release: distribution unlimited.

This research was supported by the Department of Defense, Health Affairs, under work unit no. 60002.

About The Authors

Besa Smith has used SAS for 4 years including work as an undergraduate in Biology and Chemistry at CSU, Chico, a graduate at the SDSU Graduate School of Public Health in Biostatistics, and currently as a biostatistician with the DoD Center for Deployment Health Research at NHRC.

Responsibilities include management of large military and demographic data, mathematical modeling and statistical analysis. She has been an invited presenter for WUSS 2000, and the San Diego Area's SAS users Group fall 2000 meeting.

Besa Smith, MPH
Biostatistician, Henry Jackson Foundation
Department of Defense Center for Deployment Health Research, at the Naval Health Research Center, San Diego
(619) 553-7603
besa@nhrc.navy.mil

Tyler Smith has used SAS for 9 years, including work as a student in Math and Statistics, as a graduate student at the University of Kentucky Department of Statistics, and currently as a senior statistician and data analyst with the Naval Health Research Center. His responsibilities include mathematical modeling, analysis, management, and documentation of large hospitalization and demographic data sets. He has been invited to speak at the International Biometrics Society meetings, WUSS 99, SUGI 2000, WUSS 2000, and the San Diego SAS users group 1999 and 2000 fall meetings.

Tyler C. Smith, MS
Senior Statistician, Henry Jackson Foundation
Department of Defense Center for Deployment Health Research, at the Naval Health Research Center, San Diego
(619) 553-7593
SMITH@nhrc.navy.mil