Creating Complex Graphics for Survival Analyses with the SAS® System

Steven E. Elkin, MBH Consulting, Inc., New York, NY
William Mietlowski, Sandoz Pharmaceuticals Corp., East Hanover, NJ
Kevin McCague, Sandoz Pharmaceuticals Corp., East Hanover, NJ
Andrea Kay, Sandoz Pharmaceuticals Corp., East Hanover, NJ

ABSTRACT

It is frequently of interest to evaluate a therapy across several studies (e.g., meta-analysis) or across several subgroups within a study (e.g. studies with great patient heterogeneity like cancer trials). Comprehensively displaying data from such analyses in a succinct fashion can be difficult; two types of unique graphics are presented here with the SAS system. Simulated clinical trial data from an oncology study is analyzed by stratified Log-Rank and standard survival analyses with the PHREG[i] and LIFETEST[ii] procedures. The primary graph (see figure 1) includes a plot of hazard ratios and their 95% confidence intervals on a logarithmic scale, observed minus expected values and their variance, event and patient frequencies, stratification categories, and relevant p-values. The Annotate facility and the GSLIDE[iii] procedure from SAS/GRAPH® software enables the presentation of all statistical data to appear along side of the corresponding plot on a single page. Data from the appropriate statistical procedures above is extracted with PROC PRINTTO[iv] and data step programming. The proportion of events within each strata is used to control the symbol size of the plotted hazard ratios. Additionally, a stacked vertical bar chart (see figure 2), enhanced with similar features of the annotate facility, is presented to examine patient enrollment and survival trends.

INTRODUCTION

In the January 4, 1992 issue of the Lancet[v], the Oxford Early Breast Cancer Trials Group presented a meta-analysis of 133 adjuvant breast cancer studies. In it, an elegant graphical presentation effectively displayed the consistency of the results for tamoxifen and/or chemotherapy. The plots showed the direction of the effect for each study, the relative importance of each study, and the magnitude and precision of the combined estimate and tests for interaction. These figures were used as a model for this analysis. The primary differences with this presentation are that, hazard ratios were plotted in lieu of odds ratios, and the scope of the analysis is much smaller; comparing a few hundred patients between two treatments controlling by a limited strata here versus combining thousands of patients across 133 trials and many categories in the Lancet paper.

STATISTICAL METHODS

Patient data was simulated such that treatment, survival, demographic variables, and disease history were randomized according to expectations of an upcoming clinical trial. Statistical analyses were performed with PROC's PHREG and LIFETEST.

Patient survival was analyzed by treatment and each of the following stratification factors, sex, disease stage (II/III, IV), the Karnofsky Performance Status (KPS) (50-60, 70-100), country (US, non US), and a combined stratification factor (potentially 8 levels combining sex, disease stage, and KPS). Treatment significance was determined by the stratified log-rank analysis controlling for the combined stratification variable. The interaction effect between treatment and the various stratification factors, and all necessary descriptive statistics were determined in all of the analyses.

PROC PHREG generated the hazard ratios, their 95% confidence limits, and the interaction effect between treatment and the given stratification factor. PROC LIFETEST generated the number of events (deaths), and the observed minus expected events and its variance.

DISPLAY METHODS

Relevant data from the PHREG and LIFETEST procedures, which was output using the PRINTTO procedure, was extracted with typical data step programming. Although several possible methods to generate the desired display existed, the GSLIDE procedure was used to produce an entirely annotated graph.

One alternative method was to create the graph by importing output from PROC TIMEPLOT[vi] into SAS/GRAPH with PROC GPRINT[vii]. This method was found to be difficult to control, especially with respect to aligning the numerical data with a desirable font. Another possibility would have been the use of PROC GANNO[viii].

The main piece in the display methodology was the creation of an annotation data set which could be incorporated into a SAS/GRAPH procedure. As previously mentioned, the GSLIDE procedure was chosen to display the annotations. Appropriate GOPTIONS were used which allowed for on-screen display of the graphs and permanent storage in SAS catalogs. Although Windows version 6.08 was used for this analysis, this methodology is platform independent. Minimal changes could, but not necessarily, be required to the GOPTIONS to run the program on other platforms.

Annotation Data Set Creation: Definitions and Text Display

The Annotate facility is detailed in Volume 1 of the SAS/GRAPH Software manual, chapters 18 and 19. Basically, the facility uses a data set of commands, represented by observations in the data set, to produce graphical output. Pre-defined named variables control the action that is to take place, such as what graphics element to draw (e.g. a line), where to place it (e.g. in the middle of the page), and how to do it (e.g. a solid line). Details of the data set used to produce this graphic follows.

These pre-defined variable names were primarily used in this analysis, FUNCTION, TEXT, XSYS, YSYS, POSITION, X, Y, SIZE, STYLE, LINE. Each variable serves a unique purpose, and works inter-dependently with the values assigned to the other variables in a row of data to produce a desired effect.

Typically, it is desirable to keep XSYS and YSYS constant throughout the entire data set. The assignment of these variables determine behavior of the X, Y coordinate variables. Three factors are controlled by XSYS and YSYS; definition of the coordinate system (absolute or relative), units for the coordinates (percent, axis values, cell values), and drawing area for the output (data, graphics, procedural output). Retaining '3' as values for XSYS and YSYS meant that X and Y coordinate values throughout the data set would define an absolute percentage of the entire graphics output area.

The POSITION variable was used to control the justification of each elements output. It was desirable to align the numerical output, observed-expected and standard error data, by their decimal points. Since the data was rounded to the same number of significant digits, the desired affect was achieved by assigning POSITION to '4' which right justified these elements about the chosen X, Y coordinate. All other data were centered about their given X, Y coordinate by assigning POSITION to '5'. It is important to note the potential power of the POSITION variable. Although all elements were primarily centered in this display, it is possible to have text left justified, graphics symbols centered, and numeric values right justified on the same graph.

The FUNCTION variable may be the most critical variable in the Annotate facility. All of its values are pre-defined and determine the action that is to be performed. The value of FUNCTION for displaying text is 'label'; then the TEXT variable is used to define the literal text. Other values for FUNCTION used in this analysis are 'move' and 'draw', which when defined in consecutive rows of data are used to display graphics such as lines and boxes. Defining FUNCTION as 'symbol' enables the display of all standard symbols and mimics the behavior one would achieve by creating a typical SYMBOL statement for presentation in a SAS/GRAPH procedure.

Most of this graphic was text. The display of the column headers in this analysis is defined near the beginning of this annotate data set. The text 'Deaths / Patients', 'Deaths in A', and 'Relative Hazard' appear on the same row or X coordinate. This information was therefore created by defining FUNCTION as 'label', TEXT with the given text above, and an identical X coordinate. Numeric values throughout the display were also treated as text. The death ratios were calculated with the SUMMARY procedure, and the observed minus expected data and standard errors were extracted from the appropriate statistical procedures as previously described. All of this data was made available by the set statement when first creating the annotate data set. Text variables *a* and *b* were previously created and contained the number of patients and deaths for each treatment by strata. To display this information, 'label' was of course used for the FUNCTION variable, and the TEXT variable was simply set to equal *a* or *b*. The observed minus expected data was accessed initially as numeric data in the annotate data set, and a PUT function was used to display it in the TEXT variable assignment.

Positioning this information, briefly mentioned above, was mainly handled as follows. All column headers were positioned with explicit X and Y coordinates. Explicit X coordinates were used for the stratum categories, descriptive data, and statistics. A row factor variable, *rowfact*, was calculated in a previous data set. It defined Y coordinates based on the number of strata in the analysis, and changed relative to a fixed point on the plot (Y=70).

Annotation Data Set Creation - Producing the Plot

The most difficult part of this analysis was producing the plot of hazards and the 95% confidence intervals on a logarithmic scale. Since the plot was on a single horizontal axis, the difficulty was limited to determining the horizontal placement, the X coordinate, of symbols and lines which plotted the data. A log scale was used because hazard ratio models are log linear (e.g. Cox proportional hazards model).

The hazard ratios and their 95% confidence limits were extracted from the appropriate statistical procedures in a previous data step and set into the annotate data set. Given the remaining space available on the display page, horizontal positioning of data was made relative to a starting point, X coordinate, of 78.5. Multiplying the natural logs of the hazard ratios, lower, and upper confidence limits, variables *rr, l95,* and *u95*, respectively, by another scaling factor, *sfctr*, and adding the relative starting point, an appropriate X coordinate was calculated (e.g., x=sp+(log(rr)*sfctr)). The scaling factor consistently applied a percentage value to increase the separation of the horizontal points on the plot. Without it, the X coordinates would only be determined by the actual values of the data points. Since the plot was to only cover a range of values from 0.25 to 4.00, the plot would only horizontally span 1.32% (i.e., ln[4-0.25]) of the graphics page of output. Applying a scaling factor of 14.5 meant that the plot actually spanned 19.2% (i.e., 1.32*14.5) of the graphics page of output. Only 21.5% of the page was actually available for the plot, since an *x* coordinate starting position of 78.5 had already been chosen. The various positional factors were chosen to make optimum use of the available space to present the required data and plot. Vertical positioning of the plotted data was handled similarly to that of the text data previously described.

Hazard ratios were represented in the plot by an 'x'. The size of this symbol was varied according to the relative number of events in each strata. The percentage of events in each strata was calculated and output with the FREQ procedure for further application in the annotate data set. This percentage was multiplied by an arbitrary value (3.5) to create the size of the plotted symbol (e.g., size=3.5*percent/100). In order to choose the appropriate symbol, the font was derived by defining the STYLE variable as 'marker', and the TEXT variable as 'X'. All other text and data on the graph was displayed in the Swiss font including the overall hazard ratio which was displayed by a circled plus sign (i.e. text='+').

The upper and lower confidence limits were connected by a solid line. The FUNCTION variable was set to 'move' and 'draw' in succeeding observations in the annotate data set for the lower and upper limits, respectively. Assigning the LINE variable to a value of 1, a solid line was drawn connecting the two points whose *x* coordinates were calculated similarly to the hazard ratios (e.g., x=sp+(log(*l95 or u95*)*sfctr)). Occasionally the limits fell outside the plotted range of values. In these cases, an arrow was drawn to indicate the continuing confidence interval. Left and right arrows were defined by setting TEXT equal to 'I' and 'J', respectively, and STYLE equal to 'MARKER'.

Sample Code

The following sample code illustrates the creation of the annotation data step. Only a portion of the required code is provided to describe the main ideas in constructing the graph. Strata values of 9999 indicate totals across all strata information.

```
***Determine Proportion of Total Events in Each Strata;
proc freq data=effdata; where censor=1;
     tables strat. / out=prop noprint;  run;

***Create Annotate data, initially set data with stats
***(e.g. observed-expected, std err., strata, etc.);
data anno; length function $8. text $100. style $6.;
retain xsys ysys '3'; set allstat; by strat.;

*** Enter the Overall Left and Right End Points ;
*** an Overall Scaling Factor to Resize the Plot ;
*** and Overall Horizontal Starting Point for the Plot;
lend=0.25; rend=4; sfctr=14.5; sp=78.5;

***Label Column Headings and draw dashed underline;
function='label';
  text='Deaths / Patients';
  position='5'; size=1.1;
  x=32; y=75;
output;
function='move';
```

```
  x=24; y=72;
  position='5'; size=.;
output;
function='draw';
  x=40; y=72;
  line=3;
  position='5'; size=1;
output;
***Write Obs. - Expected Data***;
function='label';
  style='swiss';
  text=compress(put(o_e,6.1));
  x=47; y=65-rowfact;
  position='4'; size=1;
output;
***Draw CI Lines***;
if l95 ge lend then do;
 function='move';
  x=sp+(log(l95)*sfctr); y=65-rowfact;
  position='5'; size=.;
 output;
end;
else do;
 function='symbol';
  x=sp+(log(lend)*sfctr); y=65-rowfact;
  text='I'; style='marker'; position='5'; size=1;
 output;
 function='move';
  x=sp+(log(lend)*sfctr); y=65-rowfact;
  position='5'; size=.;
 output;
end;
***Draw Hazard Ratio Points***;
if lend le rr le rend then do;
 function='symbol';
  x=sp+(log(rr)*sfctr); y=65-rowfact;
    if strat. = 9999 then do;
      text='+'; position='5'; size=1.5; style='';
    end;
    else do;
      text='X'; style='marker';
      position='5'; size=3.5*percent/100;
    end;
 output;
end;
***Axis ticks are created similarly to Hazard Ratios;
***Axis/reference lines created by move/draw functions;
*** Label sections of the Plot;
function='label';
  text="A is Worse";
  x=sp-9; y=65-(&maxsfact.+11);
  line=1;
  position='5'; size=1;
output;

*** Display with PROC GSLIDE;
proc gslide annotate=anno; quit;
```

ALTERNATIVE GRAPHICS

The annotate facility in SAS/GRAPH can enhance many other types of survival analysis graphics. Standard survival curves would be one, and another, discussed here, is a stacked vertical bar chart showing enrollment and survival trends.

This graph is particularly useful for interim analyses or as a monitoring tool to check enrollment and survival patterns. Unexpected changes in either enrollment or deaths can be easily viewed from this chart, which could point out changes in selection criteria for new patients or other types of protocol deviations.

Three main factors are controlled in producing this graph. One is creation of an appropriate annotation data set. Another is the PROC GCHART* syntax necessary for a vertical, stacked, bar chart. Finally, there is the labeling of axis tick marks.

In this graph, the only function of the annotation data set is to display the number alive over the number dead above the appropriate bar. The LABEL function, POSITION, XSYS, YSYS, X, Y, and TEXT variables are used in a similar manner as discussed earlier (see Sample Code section below). Appropriate placement of this data is simply controlled by assigning X to the value of the month, and Y to a value slightly more than the total number of patients. The coordinate system is set to use data values (i.e. XSYS=YSYS='2'), so values of X and Y control placement according to the actual axis definitions.

Since the data set to be graphed contains the number of patients classified by the month and survival status, a stacked vertical bar chart can be produced in PROC GCHART by defining VBAR as the month variable, MONTHNUM; the SUMVAR option as the total number of patients variable, PATSTAT; and the SUBGROUP option as the survival status class variable, START, with the PATTERNID option set to SUBGROUP. Appropriate axis and legend definitions are further detailed in the sample code section below.

Unfortunately, axis ticks may only be labeled by literally specifying a tick value and its corresponding label in the VALUE option of an AXIS statement. Since each situation different time spans may be graphed requiring different tick labels for the changing months, it was ideal to have the tick labeling process be data driven. A macro is presented in the sample code section which performs a do loop and writes out the required VALUE option to display each tick with their corresponding labels in a single macro variable, &ALLTICK. Data driven formats, created earlier in the program, are used here to display an abbreviated month and year as labels for the ticks.

Sample Code

Only a portion of the required code is provided to describe the main ideas in constructing the bar chart.

```
proc format; value censor 0='Alive' 1='Dead'; run;

*** Create Last Month macro variable;
proc sql noprint;
     select max(monthnum) into :lastmnth
     from enrdata; quit;

***Create Month/Year format for Chart;
data monyrfmt; length label $8.;
 set enrdata2 dummy(where=(monthnum=9999));
by monthnum;
if first.monthnum;
start=monthnum;
fmtname='monthyr';
if label='  ' then
  label=substr(put(enrdate,monyy5.),1,3) || ' ' ||
  substr(put(enrdate,monyy5.),4); run;
proc format cntlin=monyrfmt; run;

***Calculate Patient Status Totals***;
proc summary data=enrdata2; class monthnum censor;
     var pat; output out=status n=patstat; run;
data stat; set status;
if monthnum=. then monthnum=9999; run;
proc transpose data=stat out=statx; where censor ne .;
     by monthnum;
     var patstat;
     id censor;
     format censor censor.; run;

*Create Annotate data to place status totals over bars;
data anno; length function $8. text $100.;
 retain xsys '2' ysys '2' position '5'; set statx;
if dead=. then dead=0;
if alive=. then alive=0;
totpat=dead+alive;
if monthnum=9999 then do;
```

```
  call symput('alive',put(alive,3.));
  call symput('dead',put(dead,3.));
end;
else do;
***Write No. Alive;
 function='label';
   text=put(alive,3.);
   x=monthnum;
   y=totpat+5;
   size=1;
 if totpat ne 0 then output;
***Draw Line Between No. Alive and No. Dead;
 function='label';
   text='_____';
   x=monthnum; y=totpat+4;
   size=.5;
 if totpat ne 0 then output;
end; run;

***Determine Label Values for X-Axis Tick Marks;
data statx2; length tick $200.; set statx;
tick=compress('tick=' || put(monthnum,2.)) ||
" j=c h=1.55 '" || put(monthnum,monthyr3.) || "'" ||
" j=c h=1.55 '" || substr(put(monthnum,monthyr.),5) ||
"'";  run;
%macro tick; %global tick&t;
%do t=1 %to &lastmnth.;
  data _null_; set statx2(where=(monthnum=&t.));
  call symput("tick&t.",left(trim(tick))); run;
%end;
%mend tick; %tick

%macro ticklbl; %global alltick;
%let alltick= ;
%do i=1 %to &lastmnth;
  %let alltick= &alltick &&tick&i.;
%end;
%mend ticklbl; %ticklbl

***Create graph data set;
data anly; length label $5.;
 set stat(where=(censor ne .)); by monthnum;
label=put(censor,censor.); fmtname='patcat';
if upcase(label)='DEAD' then start=1;
else start=2; run;

***Create Survival status format;
data anlyfmt; set anly; by start;
if first.start; run;
proc format cntlin=anlyfmt; run;

***Create Chart with SAS/GRAPH Code;
legend1 frame label=(h=2.5 'Patient Status:')
        value = (h=2.5);
pattern1 value=solid;
pattern2 value=x1;
axis1 label = (h=3 a=90 "No. of Patients Entered")
      order = 0 to 35 by 5  value = (h=2.5);
axis2 label = (h=3 "Month")  value = (&alltick.);

proc gchart data=anly;
   where monthnum ne 9999;
   vbar monthnum / anno=anno discrete frame
     sumvar=patstat subgroup=start patternid=subgroup
     legend=legend1 raxis=axis1 maxis=axis2;
     format start patcat.; quit;
```

CONCLUSION

SAS/GRAPH has been shown here to be an effective tool in displaying survival analysis data. The annotate facility is an important feature which enables statistics to appear adjacent to corresponding graphics to create highly informative displays. The hazard ratio plot shows that the SAS system can be used to graphically display the robustness of the efficacy of a treatment across studies or subgroups within a study. Although shown here for mortality data, it could be easily modified for odds ratio comparisons, for binary data (e.g. tumor response), or even response rates in a single arm study. The bar chart shows that descriptive data can be elegantly displayed to check enrollment and survival status consistency during the course of a study.

REFERENCES

[i] SAS Institute Inc. (1992), SAS Technical Report P-229, SAS/STAT® Software: Changes and Enhancements, Release 6.07, Cary, NC: SAS Institute Inc., pp 433-479.

[ii] SAS Institute Inc. (1989), SAS/STAT Users Guide, Version 6, Fourth Edition, Vol. 2, Cary, NC:SAS Institute Inc., pp 1027-1069.

[iii] SAS Institute Inc. (1990), SAS/GRAPH Software: Reference, Version 6, First Edition, Volume 2, Cary, NC: SAS Institute Inc., pp 1259-1268.

[iv] SAS Institute Inc. (1990), SAS Procedures Guide, Version 6, Third Edition, Cary, NC: SAS Institute Inc., pp 483-492.

[v] The Lancet, Volume 339, January 11, 1992, "Systemic Treatment of Early Breast Cancer by Hormonal, Cytotoxic, or Immune Therapy".

[vi] SAS Institute Inc., SAS Procedures Guide, Version 6, Third Edition, Cary, NC: SAS Institute Inc., 1990, pp 579-593.

[vii] SAS Institute Inc. (1990), SAS/GRAPH Software: Reference, Version 6, First Edition, Volume 2, Cary, NC: SAS Institute Inc., pp 1131-1145.

[viii] SAS Institute Inc. (1990), SAS/GRAPH Software: Reference, Version 6, First Edition, Volume 2, Cary, NC: SAS Institute Inc., pp 739-750.

[ix] SAS Institute Inc. (1990), SAS/GRAPH Software: Reference, Version 6, First Edition, Volume 1, Cary, NC: SAS Institute Inc., pp 467-596.

[x] SAS Institute Inc. (1990), SAS/GRAPH Software: Reference, Version 6, First Edition, Volume 2, Cary, NC: SAS Institute Inc., pp 751-866.

ACKNOWLEDGMENTS

The authors would like to thank Dominick Scapati, Harvey Monder, Beate Kiese, and Sandra Williams of Sandoz Pharmaceuticals for their support in producing this paper.

Steven E. Elkin
251 West 92nd St., Apt. 2F
New York, NY 10025
Daytime: (201) 503-6868, e-mail: steveeelk@aol.com

William Mietlowski, Kevin McCague, Andrea Kay
Sandoz Pharmaceuticals Corp.
59 State Route 10, Bldg. 419
East Hanover, NJ 07936

FIGURE 1

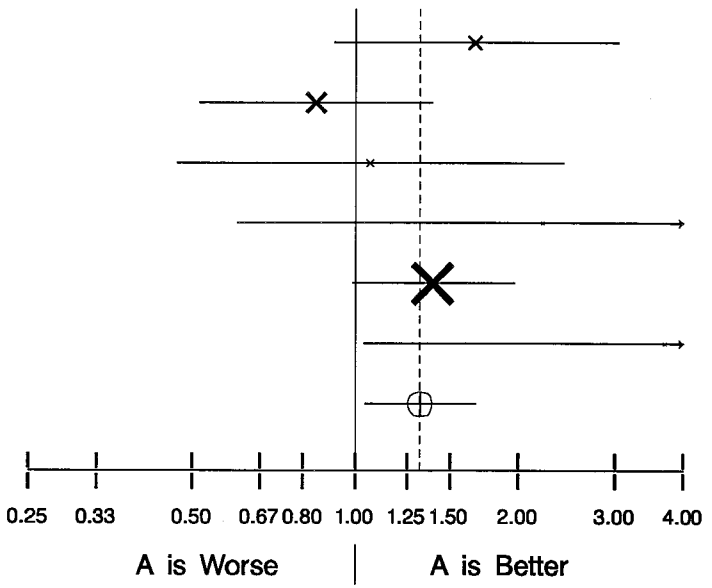
Simulated Oncology Trial Data

Hazard Ratios of Overall Survival by Strata and Treatment Group

Strata	Deaths / Patients A	Deaths / Patients B	Deaths in A O−E	Deaths in A var(O−E)
F, St. II/III, KPS 70−100	20 / 30	24 / 31	−5.4	10.4
F, St. IV, KPS 70−100	32 / 45	31 / 46	2.6	15.5
F, St. IV, KPS 50−60	11 / 15	12 / 16	−0.4	5.7
M, St. II/III, KPS 70−100	4 / 7	6 / 7	−1.9	2.3
M, St. IV, KPS 70−100	62 / 86	67 / 87	−10.6	31.6
M, St. IV, KPS 50−60	4 / 11	7 / 10	−3.2	2.3
Total	133 / 194	147 / 197	−18.9	67.8

Relative Hazard

Ratio and 95% CI



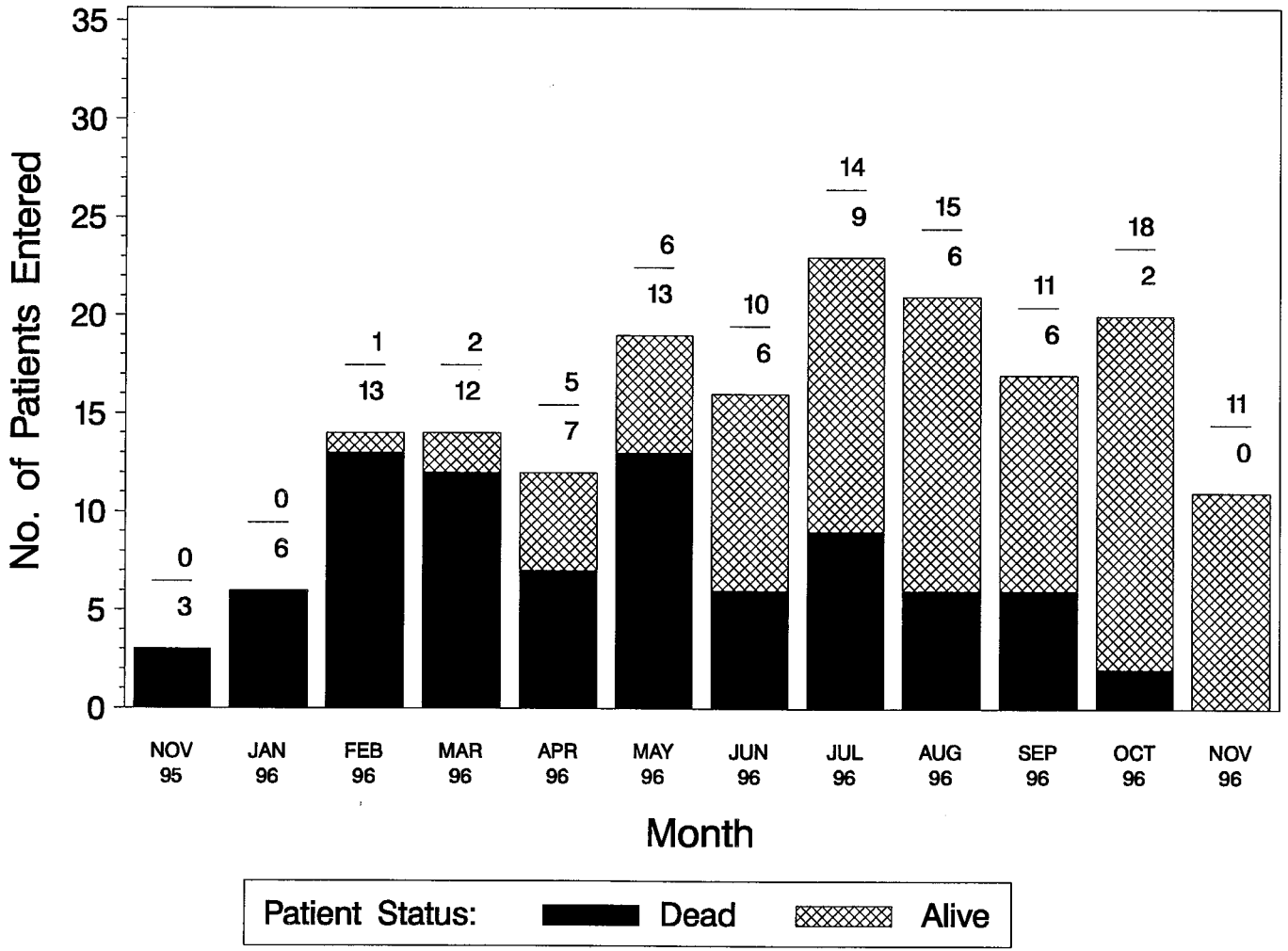
Stratified Log−Rank test p−value=0.0218.

Treatment by Strata interaction p−value=0.188.

Notes: The size of the symbols used in the plot is proportional to the number of deaths.
The X−axis is presented as a log−scale.

FIGURE 2

Simulated Oncology Trial Data (Interim Analysis)
Recruitment and Mortality Status by Calendar Month



Patient status at each monthnum is presented as the number alive above the number dead. Cumulative totals for the study period are, alive= 93, dead= 83.
Calendar month refers to month of randomization.
Interim Analysis Date: 15NOV96