

PharmaSUG – H002

7 Steps to Progression Free Survival Insights Using SAS®

Karen Walker, Walker Consulting LLC, Chandler, AZ

ABSTRACT

7 steps to Progression Free Survival Insights using SAS® will show how to take any oncology source data, render it to tumor domains, create both subject level and parameter level analysis data then subsequently produce stunning insightful documents with SAS that can be used to find a cure for cancer. A growing number of people are affected by cancer, as thus we need as many persons as possible to understand it. Because I feel so strongly about this, I'm giving up the goods on the best work I know. You'll find in this paper a collection of methods, and programs that are so intuitive, and eloquently assembled that anyone who can read will understand my process. Perhaps there's a talented doctor who has no time to tinker with a computer program, and yet is close to finding the right treatment. This paper is for that person. My aim is to make CDISC SDTM and ADaM data rendering follow easier. So this paper illustrates the natural progression from: data gathering, to analysis, to insights.

INTRODUCTION

"We classify disease as error, which nothing but Truth or Mind can heal." – Mary Baker Eddy, Science and Health

This paper is written for everyone with hope of advancing studies, particularly oncology studies. While SAS is the software solution of choice, this paper has been developed broad enough to adapt to any software solution. Because the most important part of doing a study is the set up. When study data is properly arranged, the outcomes present themselves. The reader will learn SQL concepts; SAS code examples, data tables, graphs and how to render data tables using CDISC SDTM and ADaM implementation guides. This style of programming lends to portability, discovery, and methods that work with quality assurance every time.

STEP 7: KAPLAN MEIER PLOTS

Consider the process of choosing a cancer therapy having the results be the best outcome for survival. This is similar to any standard optimization by where one must select from list of choices. Using SAS OR we meet our objective when we make the optimal choice. The choices for the Kaplan Meier Plot on progression free survival are refined from a subject level ADaM dataset. Simply put, they are: 1 or 0. So outcomes are mined from a collection of oncology source data by placing a limit on choices from the analysis to 0 or 1 in PFS.

Figure 1 best represents the outcomes from this data table with a Kaplan Meier plot.

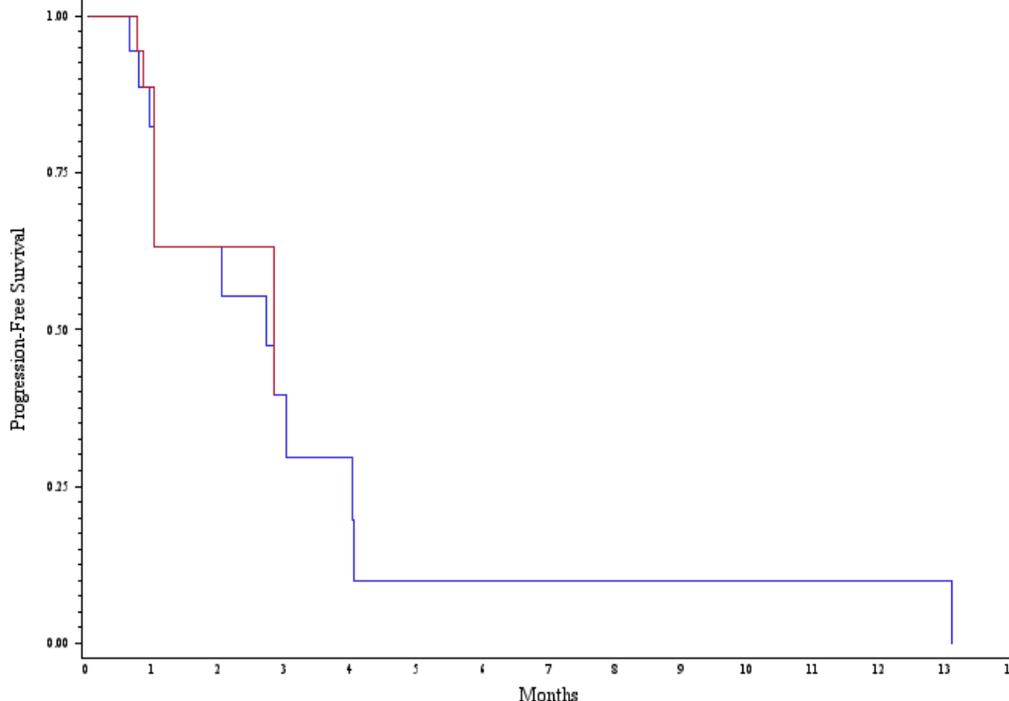


Figure 1.
DATA: PFS refined from ADRS_subject
Progression Free Survival by Months on Study

```

/**** Compute Kaplan Meier Statistics ****/

ods listing;
proc lifetest data = PFS
  alphaqt = 0.05 timelist=1 2 3 4 5 6 7 8 9 10 11 12 13 14
  plots = (survival)
  outsurv = KMPLLOT ;
  time months * aval(1);
  id subjid;
run;
ods listing close;
proc sort data=KMPLLOT; by months _censor_ survival; run;

/**** ANNOTATE DATASET *****/
data Annotate;
  length function $8;
  retain probability;
  set KMPLLOT;
  by months _censor_ survival;
  if first.survival;
  function = 'LABEL';
  if _CENSOR_ = 1 then GROUP = 1; else GROUP = 0;
  if .Z < SURVIVAL then probability = SURVIVAL;
  pvalue = probability;
run;

proc sort data=Annotate; by ady descending pvalue; run;

goptions reset=(axis, legend, pattern, symbol, title, footnote) rotate=landscape
hpos=90 vpos=90 vsize=4.2 in hsize= 7 in targetdevice=winprtc
ctext=black htext=1.6 ftext="Times New Roman" ftitle="Times New Roman" noprompt
device=png;
ods rtf close;
ods listing;
ods rtf file = "C:\My Documents\Figures\&prognm..RTF" bodytitle style=rtf;
axis1 major=none minor=none order=(0 to 10 by 1) label=(height=2.5 'Months');
axis2 offset=(2) order=(0.00 to 1.00 by 0.25) label=(angle=90 rotate=0 height=2.5
color=black 'Progression-Free Survival');
symbol i=STEPLJ;
proc gplot data = Annotate;
  plot pvalue*months = GROUP / haxis = axis1
  vaxis = axis2 nolegend noframe;
run;
quit;
ods rtf close;
ods listing;

```

months	_CENSOR_	SURVIVAL	SDF_LCL	SDF_UCL
	0		1	1
0.642857143	0	0.944444444	0.795656669	
0.785714286	0	0.885416667	0.722610655	
0.928571429	0	0.822172619	0.645407215	
	1	0	0.443350541	
2.035714286	0	0.553385417	0.360948029	
2.714285714	0	0.474330357	0.285971033	
2.821428571	0	0.395275298	0.217305686	
	3	0	0.134040455	
	4	0	0.066557011	
13.07142857	0	0		

STEP 6: MODEL TIME TO EVENT OR PFS END POINT

For our response analysis data, we have collected decisions for therapy as follows: CR complete response, SD sustained response and PD partial response on each day the subject's cancer was observed. We'll collapse those responses down to subject level by using the **SQL** code below keeping the best overall response (RECIST). The result PFS table is used as input for **SAS PROC LIFETEST**.

```
* Here AVAL is used like a data censoring flag ;
proc sql;
  create table PFS as
  select trtlpn, trtlp, subjid, avisit, avisitn, ady, avalc,
         case when bor = 'PD' then 1 else 0
              end as AVAL,
         ady/30.5 as months
  from ADRS_Subject
  order by trtlpn, trtlp, subjid, ady
;
quit;
```

STEP 5: WATER FALL PLOTS

Tumor Marker data is best understood by a series of test. The best way to show a series of test over time is the Waterfall plot. Here we evaluate the **CA-125** "Carbohydrate Antigen 125" levels during treatment. Often the CA-125 tumor marker is found on the rim of cancer cells and shows elevated levels (i.e ≥ 35 u/mL) in the blood of persons with ovarian cancer. The Waterfall plot will show us how CA-125 levels are affected as cancer subjects are exposed to therapy.

The level of CA-125 is less than 35 units per milliliter of blood for normal subjects. The percent change value is captured in the analysis variable "**AVAL**" on the response analysis dataset **ADRS_Param**. This Parameter level analysis dataset contains a series of measurements over time "**ADY**" for each subject, and each immunoassay test measurement of CA-125 is captured in "PARAM".

Figure 2 shows 25% of ovarian cancer samples had an increase in CA-125 Antigen levels, while 75% had decreased levels as treatment days **ADY** increased.

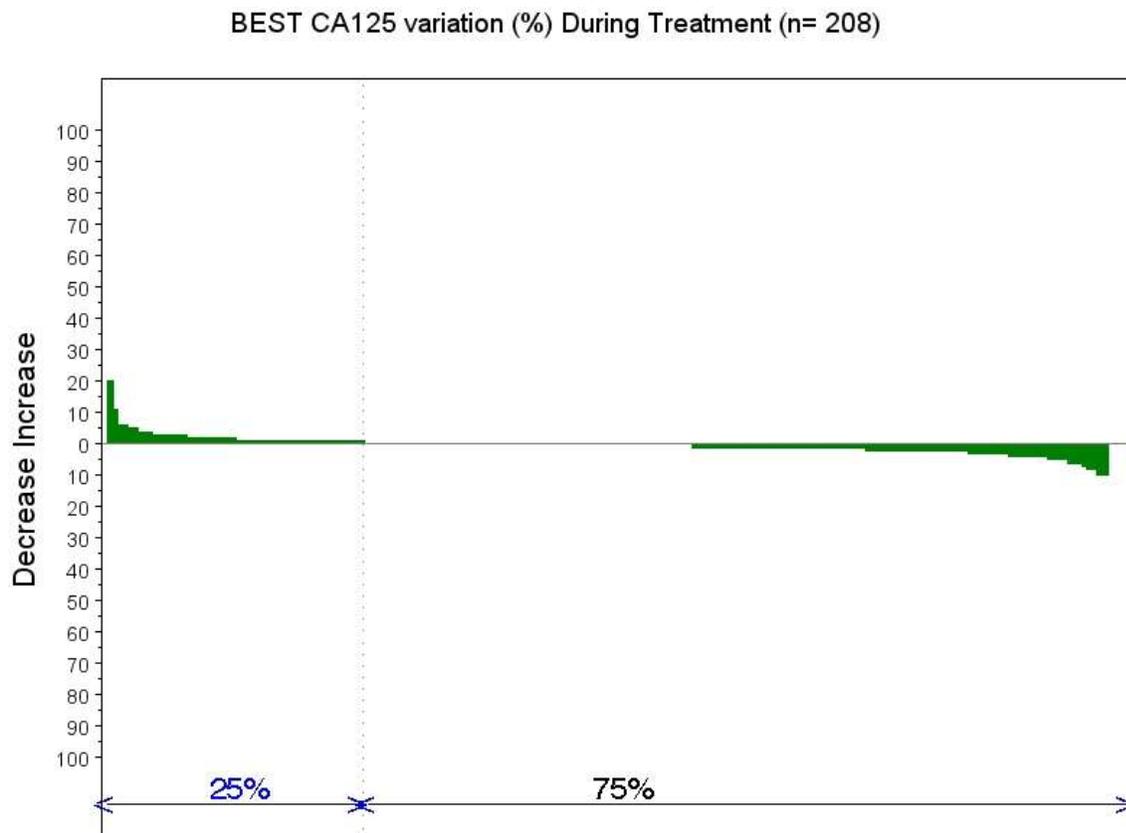


Figure 2.
Dataset ADRS_PARAM
Percent Change from Baseline CA125 U/ML 208 Subjects

WATERFALL CODE

Using data from the Parameter Level Analysis dataset **ADRS_Param**, we categorize the variation of the percent change relative to the baseline. This insight considers only the CA-125 level as it compares to samples taken on treatment. When the CA-125 measurement drops below baseline a value “**AVAL**” is assigned a 0 otherwise it's 1. Although we count the number of distinct subjects, the subject count is less relevant than the individual sample.

```
Data PCT_Change;
    set ADRS_param;
    flag=1 ;
    count= n ;
    if AVAL>0 then variation=0;
    else If AVAL<=0 then variation=1;
run;

* Count number of Subjects;
PROC SQL NOPRINT;
    SELECT n(DISTINCT subjid) INTO: subcnt FROM PCT_Change;
quit;

Proc freq data=PCT_Change;
    table variation/missing out=CA125F sparse;
run;

* CA125 Histogram;
Data CA125;
length title $55.;
set CA125F;
title='BEST CA125 variation (%) During Treatment' || trim(compbl(" (n=" || "&subcnt" || ")"));
PERCENT1=compbl(round(PERCENT,.1)||"%");
If variation=0 then do;
    call symput('n_increase',COUNT);
    call symput('n_increaseb',COUNT+0.5) ;
    call symput('increase_percent',PERCENT1);
end;
If variation=1 then do;
    call symput('n_decrease',COUNT);
    call symput('decrease_percent',PERCENT1);
end;
run;

proc sort data=CA125; by count descending percent; run;
```

STEP 4: RENDER ANALYSIS DATA MODEL

Analysis datasets are designed to support statistical analysis. They are created using the **SDTM** datasets. Good Programming Practices suggest analysis datasets that are built to be 1 procedure short of an output table, listing or figure. Analysis datasets and associated metadata are among the four types of data that can be submitted to the FDA. These also include study tabulation datasets, subject profiles, and listings. Two kinds of analysis datasets are discussed here: subject, and parameter level. The subject level contains the data specific to the study endpoints. The parameter level supports the subject level by charting a trail from the analyzed data, to the endpoint. For more information on **ADaM** data table definitions and types refer to the **ADaM** Implementation Guide.

SUBJECT IDENTIFIER LEVEL ANALYSIS DATASET

This dataset contains one row per subject for the final adjudication. Keep it simple by using few basic variables: **TRT1PN, TRT1P, SUBJID, AVISIT, AVISITN, ADY, AVALC, and BOR.**

```
proc sql;
    create table ADRS_subject as
    select a.trt1pn, a.trt1p, subjid, a.avisit, a.avisitn, a.ady, a.avalc, b.avalc as bor
    from
    (select distinct trt1pn, trt1p, usubjid, avisit, avisitn, ady, avalc from RS group by trt1p,
    usubjid, avisit ) a
    left join
    (select distinct trt1p, usubjid, avisit, avisitn, ady, avalc from
    BOR(where=(paramcd='BESTOR')) ) b
    on a.trt1p = b.trt1p and
    a.usubjid = b.usubjid and
    a.avisit = b.avisit and
    a.avisitn = b.avisitn
    order by a.trt1pn, a.trt1p, a.subjid, a.ady ; quit;
```

PARAMETER LEVEL ANALYSIS DATASET

This dataset contains one row per subject for each parameter every day on study. Keep it simple by using few basic variables: TRT1PN, TRT1P, SUBJID, PARAM, PARAMCD, AVISIT, AVISITN, ADY, and AVAL.

```
proc sql;

  create table ADRS_parm as
  select adsl.actarm as TRT1P,
         adsl.actarmvf as TRT1PN,
         adsl.subjid as subjid,
         RS.visit as avisit,
         RS.visitnum as avisitn,
         RS.rsorres as PARAM,
         RS.rstest as PARAMCD,

         ((RS.rsstresn - RS.BASE)/RS.BASE)*100 as AVAL
         intck('days',adsl.trtendt,best.rsdtdt) as ADY
  from
  (select usubjid, actarm, trtendt from analysis.adsl(where=(safll eq 'Y')) ) adsl
  left join
  (select distinct usubjid, rsseq, rstest, rstestcd, rsorres, visit,
   mean(rsstresn) as BASE,
   mdy( input(substr(rsdtdc,6,2),4.),
        input(substr(rsdtdc,9,2),4.),
        input(substr(rsdtdc,1,4),4.) ) as rsdt format=ISO8601da.,
   hms( input(substr(rsdtdc,12,2),4.),
        input(substr(rsdtdc,15,2),4.),00 ) as rstm format=ISO8601dt.,
   visit
  from CDISC.RS group by usubjid, rstest ) rs
  on adsl.usubjid = rs.usubjid
  left join
  (select usubjid, rsdtdc, rsstresn, visit, rstest,
   mdy( input(substr(rsdtdc,6,2),4.),
        input(substr(rsdtdc,9,2),4.),
        input(substr(rsdtdc,1,4),4.) ) as rsdt format=date9.,
   hms( input(substr(rsdtdc,12,2),4.),
        input(substr(rsdtdc,15,2),4.),00 ) as rstm
  from CDISC.RS(where=(RSTESTCD in ('BEST' ))) group by usubjid, rstest ) best
  on rs.usubjid = best.usubjid and
   rs.rstest = best.rstest
  order by adsl.actarm, adsl.usubjid, rs.rstest, rs.visit
  ; quit;
```

STEP 3: RENDER STANDARD DATA TABULATION MODEL

CDISC_RS (THE TUMOR RESPONSE DOMAIN MODEL)

There are a number of well know methods that determine if a subject is responding favorably to a cancer treatment. Here we'll use RECIST method to generate the standard data tabulation model Response Domain.

```
%macro resp(dmn=, resp= );
data &resp;
do until (last.trdtdc);
  length linkid $200 orresn 8;
  retain linkid orresn;
  set &dmn;
  by usubjid trlinkid trdtdc;
  if linkid ^= '' and trlinkid = linkid then do;

    if trstresn = 0 then RSORRES = '1'; /* CR Complete Response*/
    else if trstresn < orresn then RSORRES = '2' ; /* PR Partial Response*/
    else if trstresn = orresn then RSORRES = '3'; /* SD Stable disease*/
    else RSORRES = '4' ; /* PD Disease Progression*/

  end;

  linkid = trlinkid;
  orresn = trstresn; end;run;
```

```

%mend resp;

proc sql;
  create table rs1 as
  select a.*,
         case
           when b.tuorres = 'TARGET' then 'TRGRES'
           when b.tuorres = 'NONE-TARGET' then 'NTRGRES'
           when b.tuorres = 'NEW' then 'NEWLPROG'
           else ''
         end as rstestcd,
         case
           when b.tuorres = 'TARGET' then 'Target Response'
           when b.tuorres = 'NONE-TARGET' then 'Non-target Response'
           when b.tuorres = 'NEW' then 'New Lesion Progression'
           else ''
         end as rstest,
         case
           when a.trblfl eq 'Y' then a.trstresn
           else .
         end as baseline
  from cdisc.tr (where = (trtestcd = 'LDIAM')) as a left join
  (select distinct tulinkid, tuorres from cdisc.tu) as b
  on a.trlinkid = b.tulinkid
  order by a.usubjid, trlinkid, a.trdctc
  ;

create table rs2 as
  select a.*,
         case
           when b.tuorres = 'TARGET' then 'NTRGRES'
           else ''
         end as rstestcd,
         case
           when b.tuorres = 'TARGET' then 'Non-Target Response'
           else ''
         end as rstest,
         case
           when a.trblfl eq 'Y' then a.trstresn
           else .
         end as baseline
  from cdisc.tr (where = (trtestcd = 'SUMVDIAM' and trntl eq 'present')) as a left join
  (select distinct tulinkid, tuorres from cdisc.tu) as b
  on a.trlinkid = b.tulinkid and b.tulinkid ne 'NEW01'
  order by a.usubjid, trlinkid, a.trdctc
  ;

quit;

%resp(dmn=rs1, resp=TRGRES);
%resp(dmn=rs2, resp=NTRGRES);

data rs0;

  retain bl;
  set TRGRES NTRGRES;
  by usubjid trlinkid trdctc;
  DOMAIN = 'RS';
  RSGRPID = TRMETHOD;
  RSLINKID = TRLINKID;
  RSSEQ = .;
  RSCAT = TRCAT;
  RSORRESU = 'MM';
  RSSTRESN = TRSTRESN;
  RSSTRESN2 = ostresn;
  RSSTRESU = 'MM';
  RSREASND = '';
  RSMETHOD = 'RECIIST';
  RSSTAT = TRSTAT;
  RSBASE = baseline;
  RSDTC = TRDTC;
  RSDY = TRDY;

```

```

if first.trlinkid then bl = baseline;

if first.trdtc then output;

if RSORRES = '4' and first.trdtc then do;
    RSTESTCD = 'FIRSTPD';
    RSTEST = 'First Documented Progressive Disease';
    output;
end;

if RSORRES = '4' and TRPREVIR ne '1' then do;
    RSTESTCD = 'NRADPROG';
    RSTEST = 'Non-Radiological Progression';
    output;
end;

RSSTRESC = put (RSORRES,$rsorres.);

run;

proc freq data=RS0 order=freq noprint;
    table usubjid*rslinkid*rsorres / out=overall_tumor ;
    table usubjid*rsorres / out=overall ;
run;
proc sort data=overall; by usubjid; run;

data RS_overall;
merge RS0 overall;
by usubjid;

RSTESTCD = 'OVLRESP';
RSTEST = 'Overall Response';

run;
data RS_overall;
set RS_overall;
by usubjid;
if first.usubjid;
run;

proc sort data=overall_tumor; by usubjid rslinkid rsorres; run;
proc sort data=RS0; by usubjid rslinkid rsorres; run;

data RS_overall_tumor;
merge RS0 overall_tumor;
by usubjid rslinkid rsorres;

RSTESTCD = 'BEST';
RSTEST = 'Best Response (Adjudicated)';
RSEVAL = 'Final Reading';

run;
data RS_overall_tumor;
set RS_overall_tumor;
by usubjid rslinkid rsorres;
if first.rslinkid;
run;

data RS0 (Keep = STUDYID DOMAIN USUBJID RSGRPID RSLINKID RSTESTCD RSTEST RSCAT
RSORRES RSORRESU RSSTRESN RSSTRESC RSSTRESU RSREASND RSMETHOD
RSSTAT RSNAME RSEVAL RSBFL RSDTC RSDY );
set RS0
RS_overall
RS_overall_tumor;

by usubjid rslinkid;

RSORRES = put (rsorres,$best.);

* Have all the lesions disappeared? Yes No;
if (rsstresn - rsstresn2) <= 0 then RSNTL1 = 'Yes'; else RSNTL1 = 'No';

```

```

*Is there a 30% shrinkage from baseline? Yes No;
  if (rsstresn - baseline) <= .30*baseline then RSNTL2 = 'Yes'; else RSNTL2 = 'No';

*Is there a 20% increase from any previous minimum assessment? Yes No;
if (rsstresn - rstresn2) >= .20*rstresn2 then RSNTL3 = 'Yes'; else RSNTL3 = 'No';

run;

proc sort data=RS0; by usubjid rslinkid rsdtc; run;
data RS(label='RECIST');

  Retain STUDYID DOMAIN USUBJID RSSEQ RSGRPID RSLINKID RSTESTCD RSTEST RSCAT
        RSORRES RSORRESU RSSTRESN RSSTRESC RSSTRESU RSREASND RSMETHOD
        RSSTAT RSNAME RSEVAL RSBLFL RSDTC RSDY;

  set RS0;
  by usubjid rslinkid rsdtc;
  if (first.usubjid) then RSSEQ=0;
  RSSEQ + 1;
  label RSSEQ = 'Sequence Number';
run;

proc compare base=cdisc.RS compare=work.vRS; run;
proc datasets delete all;; run;

```

STEP 2: CONTROL YOUR VOCABULARIES

The key to transforming from investigator or radiology notes to that which can be used to show endpoints, is the use of controlled vocabularies. Both **SDTM** and **ADaM** data renderings are subject to standard vocabularies. Datasets are named according to an insight captured, like the model (RECIST). While variables are named by the kind of information they contain. In SAS categorical variables are named using FORMATS, and binomial variables are named using FLAGS. Similarly, in CDISC SDTM, numeric values are named by joining a two character domain modifier with the ending “STRESN”(e.g “RSSTRESN” where RS is the response domain). In CDISC ADaM, numeric values are named “AVAL” which is short for “Analysis Value”. All character valued counterparts are named using a “C” in place of the “N” for example “STRESC” for SDTM and “AVALC” for ADaM. For more information on these conventions refer the ADaM Implementation Guide, and the SDTM Implementation guide found at the CDISC website.

In addition to what the outside designation of the datasets and variables are called, controlled vocabularies also helps normalize source data, by telling us what write in the dataset or variable. Consider a LAB data domain “CDISC_LB” There are standard units’ contained in “LBSTRESU”. These often require numeric conversions involving molecular weights for all substances collected. The data collected from the investigator is saved in “LBORRESU” as it happens it is not a standard unit. This is addressed by putting in “LBORRESU*Conversion Factor” in LBSTRESU.

To capture character descriptions on ranges and limits for each substance we have controlled vocabularies DAIDS, and CTCAE. This is a dictionary that contains grades are established on modifiers like “LBSTRESU”. For example a lab test can have a measurement that is graded as “SEVERE” or “LIFE THREATENING”. For this instance we set a variable called “LBTOXGR” to “SEVERE” or “LIFE THREATENING”.

The NCI Dictionary of Cancer terms provides a list of entries we can make on our datasets and variables that are specific to treating cancer.

A record in the TU domain contains the following information: a unique tumor ID value(TULINKID); anatomical location of the tumor(TULOC); method used to identify the tumor(TUMETHOD); role of the individual identifying the tumor(TUEVAL); and timing information(TUDTC,VISIT,VISITNUM).

```

proc format library=work;
  value $test
    'OVLRESP'= 'Overall Response'
    'BEST'    = 'Best Response'
    'FIRSTPD' = 'First Documented Progressive Disease'
    'TRGRES'  = 'Target Response'
    'NTRGRES' = 'Non-target Response'
    'NEWLPROG'= 'New Lesion Progression'
    'NRADPROG'= 'Non-Radiological Progression'
  ;

```

```

value $best
'1' = CR
'2' = PR
'3' = SD
'4' = PD
'5' = NE
'6' = NA
;

value $rsorres
'1' = 'Complete Response'
'2' = 'Partial Response'
'3' = 'Stable Disease'
'4' = 'Progressive Disease'
'5' = 'Not Evaluable'
'6' = 'Not Assessed'
;

```

run;

The TU domain represents data that uniquely identifies tumors. T

```

ATTRIB STUDYID length=$200 label='Study Identifier'
DOMAIN length=$2 label='Domain Abbreviation'
SUBJID length=$200 label='Unique Subject Identifier'
TUGRPID length=8 label='Group ID'
TUSPID length=8 label='Sponsor ID'
TULINKID length=$200 label='Unique Tumor Identifier'
TUTESTCD length=$40 label='Tumor Identification Short Name'
TUTEST length=$200 label='Tumor Identification Test Name'
TUCAT length=$200 label='Category for Tumor Identification'
TUORRES length=$200 label='Tumor Identification Result'
TUSTRESC length=$200 label='Tumor Identification Result Std Format'
TULOC length=$200 label='Location of the Tumor'
TULAT length=$200 label='Laterality'
TUDIR length=$200 label='Directionality'
TUDSTREB length=$200 label='Distribution'
TUMETHOD length=$200 label='Method of Identification'
TUEVAL length=$200 label='Evaluator Specified'
VISITNUM length=8 label='Visit Number'
VISIT length=$200 label='Visit Name'
TUNTL length=$7 label='Lesions Present'
TUPREVIR length=$200 label='Previously Irradiated'
;

```

The TR domain represents quantitative measurements and/or qualitative assessments of the tumors identified in the TU domain. These 190 measurements are usually taken at baseline and then at each subsequent assessment to support response evaluations

```

ATTRIB TRGRPID length=8 label='Group ID'
TRLINKID length=$200 label='Unique Tumor Identifier'
TRTESTCD length=$40 label='Tumor Assessment Short Name'
TRTEST length=$200 label='Tumor Assessment Test Name'
TRCAT length=$200 label='Category for Tumor Assessment'
TRORRES length=$200 label='Result or Finding in Original Units'
TRORRESU length=$200 label='Original Units'
TRSTRESC length=$200 label='Character Result/Finding in Std Format'
TRSTRESN length=8 label='Numeric Result/Finding in Standard Units'
TRSTRESU length=$200 label='Standard Units'
TRSTAT length=$200 label='Tumor Assessment Status'
TRREASND length=$200 label='Reason Tumor Measurement Not Performed'
TRMETHOD length=$200 label='Method used to identify the Tumor'
TRBLFL length=$200 label='Baseline Flag'
TRDTC length=$200 label='Date/Time of Tumor Measurement'
TRNTL length=$40 label='Non-Target Lesions'
TRPREVIR length=$200 label='Previously Irradiated'
;

```

The RS domain represents the response evaluation determined from the data in TR. Data from other sources (in other SDTM domains) might also be used in an assessment of response for example, RECIST

ATTRIB

```

RSGRPID    length=$200 label='Group ID'
RSLINKID   length=$200 label='Unique Tumor Identifier'
RSTESTCD   length=$8    label='Short Name of Test/Response'
RSTEST     length=$200 label='Response Assessment/Test'
RSCAT      length=$200 label='Category for the Response'
RSORRES    length=$200 label='Result or Finding in Original Units'
RSORRESU   length=$200 label='Original Units'
RSSTRESC   length=$200 label='Character Result/Finding in Std Format'
RSSTRESN   length=8     label='Number Result/Finding in Standard Units'
RSSTRESN2  length=8     label='Previous minimum measurement'
RSSTRESU   length=$200 label='Standard Units'
RSMETHOD length=$200 label='Method Used to Assess Response'
RSSTAT     length=$200 label='Tumor Assessment Status'
RSREASND   length=$200 label='Reason Tumor Measurement Not Performed'
RSNAM      length=$200 label='Vendor Name'
RSEVAL     length=$200 label='Radiologist of Reader'
RSBASE     length=$200 label='Baseline'
RSDRVFL    length=$200 label='Derived Flag'
RSDTC      length=$200 label='Date/Time of Tumor Measurement'
RSDY       length=8     label='Response Days from First Dose'
RSSEQ      length=$200 label='Group ID'
RSNTL1     length=$200 label='Have all lesions disappeared'
RSNTL2     length=$200 label='Is there a 30% shrinkage from baseline'
RSNTL3     length=$200 label='Is there a 20% increase from any previous min assessment'
;

```

STEP 1: MINE YOUR DATA

Oncology Data is captured from immunoassay test, radiology test or by an investigator.

Collect and keep only the data that is important to either categorical or binomial decision making. Make certain the data is concise, accurate, and can show effectiveness. Recall that the primary objective for collecting data is to measure effectiveness and summarize. Avoid collecting, keeping and processing data that has little to no affect on your bottom line. In FDA terms; collect “substantial evidence” that the data is from a “well-controlled study of comparative effectiveness” (**21 CFR 314.126**).

Here are common examples for collecting and establishing your data for cancer studies. We use the dictionaries whenever possible, this helps the task of setting the data up to make the study process smoother. Before set up, have an idea of the question that needs an answer. Is it a list of things that can occur? Is it one or the other? Is the answer quantifiable and best expressed by a metric? Should the set up be inadequate just try again.

1. Categorical (A,B,C,D).
2. Binomial (1=Yes, 0=NO).
3. Nominal(>37,>35 , >35,+25)

```

/* If Tumors were too small to read on CT Scan, MRI, or Ultrasound they were
identified from Immunoassay test blood work using bio-marker Cancer Antigen
load per U/ML of blood. CA 19.9 is secreted by tumor on the pancreas, CA
27.29 is secreted by tumor on the breast, CA 125 is found on the surface of
ovarian cancer cells, PSA is found on the prostate gland.  ***/

```

```

proc format library=work;
  value $TUtest
    'A' = 'CARBOHYDRATE ANTIGEN 19.9'
    'B' = 'CARBOHYDRATE ANTIGEN 27.29'
    'C' = 'CARBOHYDRATE ANTIGEN 125'
    'D' = 'PSA'
  ;
run;

```

```

* Dose the pancreas carbohydrate antigen 19.9 in the blood exceed 37 U/ML? Yes No;
if TATEST='A' & TUSTRESN > 37 then do;
    TRCAT = 1; TATESTCD = 'TUMIDENT'; TUORRES = 'TARGET'; end; else do;
    TRCAT = 0; TATESTCD = 'BENIGNAB'; TUORRES = 'BENIGN'; end;

* Dose the Brest carbohydrate antigen 27.29 in the blood exceed 35 U/ML? Yes No;
if TATEST='B' & TUSTRESN > 35 then do;
    TRCAT = 1; TATESTCD = 'TUMIDENT'; TUORRES = 'TARGET'; end; else do;
    TRCAT = 0; TATESTCD = 'BENIGNAB'; TUORRES = 'BENIGN'; end;

* Dose the ovarian carbohydrate antigen 125 in the blood exceed 35 U/ML? Yes No;
if TATEST='C' & TUSTRESN > 35 then do;
    TRCAT = 1; TATESTCD = 'TUMIDENT'; TUORRES = 'TARGET'; end; else do;
    TRCAT = 0; TATESTCD = 'BENIGNAB'; TUORRES = 'BENIGN'; end;

* Has the prostate carbohydrate antigen in the blood exceeded 25% of baseline of U/ML?
Yes No;
if TATEST='D' & TUSTRESN > .25*TRBLFL then do;
    TRCAT = 1; TATESTCD = 'TUMIDENT'; TUORRES = 'TARGET'; end; else do;
    TRCAT = 0; TATESTCD = 'BENIGNAB'; TUORRES = 'BENIGN'; end;

/* Categories for Solid Tumors ***/

proc format library=work;
    value $TUtest
        'LDIAM' = 'LONGEST DIAMETER'
        'SUMVDIAM' = 'SUM OF VIABLE DIAMETER'
        'PCHGBL' = 'Percent Change From Baseline'
    ;
    value $TUeval
        'OVIN' = 'OVERALL ASSESSMENT INVESTIGATOR'
        'OVRA' = 'OVERALL ASSESSMENT RADIOLOGY'
        'TTEOVIN' = 'TIME TO OBJECTIVE RESPONSE INVESTIGATOR'
    ;
run;

```

As input data from the investigator or lab is examined we noticed there are many expressions which describe a distinct tumor location. Perhaps there's more than one investigator or Lab. Producing synonyms is a way data is normalized to the terms as they appear on the dictionary. Before this data can be summarized one unique name is taken from the dictionary. For example: The tumor location "ABDOMEN" is mined from four aliases.

```

proc format library=work;
    value $TULoc
        'ABDOMEN (LIVER)' = 'ABDOMEN'
        'ABDOMEN LIVER SEGMENT 7' = 'ABDOMEN'
        'AMT. ABDOMINAL WALL' = 'ABDOMEN'
        'ANTERIOR ABDOMINAL WALL' = 'ABDOMEN'
    ;
run;

```

CONCLUSION

Cancer is a disease that affects a growing number of people. If we are to help those suffering from it we must first understand it. We are that much closer to beating it by taking the time to know. This paper seeks to extend an invitation to readers that use this information. Facilitate understanding of how study data flows from gathering to endpoint. Enhance and expand on it knowing that if you mine your data correctly, then controlling your vocabulary will be easy, subsequently allowing you to render your data for study tabulation and analysis. By modeling your cancer treatment inference with your data you can quickly decide if you have a solid conclusion about a cancer. Be it yes or no, a choice from a list, or a quantifiable metric. With these insights we can try again and again until we have a cure for cancer. Thanks so much for your time and interest.

REFERENCES

- (1) Guidance for Industry Clinical Studies Section of Labeling for Human Prescription Drug and Biological Products – Content and Format
- (2) ADaM Implementation Guide Prepared by the CDISC ADaM Team <http://www.cdisc.org/standards/index.html>.
- (3) Oncology Disease-specific Therapeutic Area Supplement to the Study Data Tabulation Model Implementation Guide Prepared by the CDISC Submission Data Standards Oncology Sub-Team <http://www.cdisc.org/standards/index.html>
- (4) Bruce D. Cheson, Beate Pfister, et al. Revised Response Criteria for Malignant Lymphoma *Journal of Clinical Oncology*. Vol 25 Number 5 Feb 10 2007
- (5) DR Macdonald, TL Cascino, et al. Response criteria for phase II studies of supratentorial malignant glioma *Journal of Clinical Oncology*, Vol 8, 1277-1280

REPRESENTATION AND WARRANTIES, LIMITATION OF LIABILITY, AND DISCLAIMERS

CDISC Patent Disclaimers

The writer wishes conveys no guarantees. I cannot guarantee that a magic bullet treatment for cancer will result from understanding these concepts, however it's is a start.

CDISC Patent Disclaimers

It is possible that implementation of and compliance with this standard may require use of subject matter covered by patent rights. By publication of this standard, no position is taken with respect to the existence or validity of any claim or of any patent rights in connection therewith

ACKNOWLEDGEMENTS

This First I'd like to thank my lord and savior Jesus Christ, without him there is no inspiration for better.

Special thanks to Alfred Balch PhD, for allowing me to create, develop, test and his ideas and helping me gain a better understanding of statistics.

Special thanks to Andrew P. Beelen MD, for his un-exhaustible attention to detail, even helping with molecular weights DAIDS, CTCAE , and all things associated with oncology.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Karen Walker:
Walker Consulting LLC:
6481 West Gary Drive :
Chandler, Arizona , 85226 :
Work Phone:
Fax:
Kkwalker77@yahoo.com:
Web:

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.