Paper DG12-2011

Area under a Curve: Calculation and Visualization

Patricia M. Herbers, M.S., Cincinnati Children's Hospital Medical Center, Cincinnati, OH
Deborah A. Elder, M.D., Cincinnati Children's Hospital Medical Center, Cincinnati, OH
Jessica G. Woo, Ph.D., Cincinnati Children's Hospital Medical Center, Cincinnati, OH

Abstract

At times it is necessary to summarize longitudinal data with a single value for analysis purposes. When each subject has repeated, changing measurements over time, one way to summarize the data is to graph the measurement across time and calculate the area under the curve (AUC) relative to a baseline value. Such a summary necessarily involves a loss of information. For example, subjects with similar AUC values may have very different trajectories, peak amplitudes, or peak timing, which may be important to visualize. To more fully describe the data, a panel of graphs of each subject's original data with the summary AUC value inset may be included with the analysis.

SAS programming is presented to calculate the AUC via Riemann sums for data organized in a stacked data set, that is, multiple observations per subject. The use of indicator variables and scatter statement options are then used to produce individual graphs with insets using the SGPANEL procedure, which lacks an INSET option.

Introduction

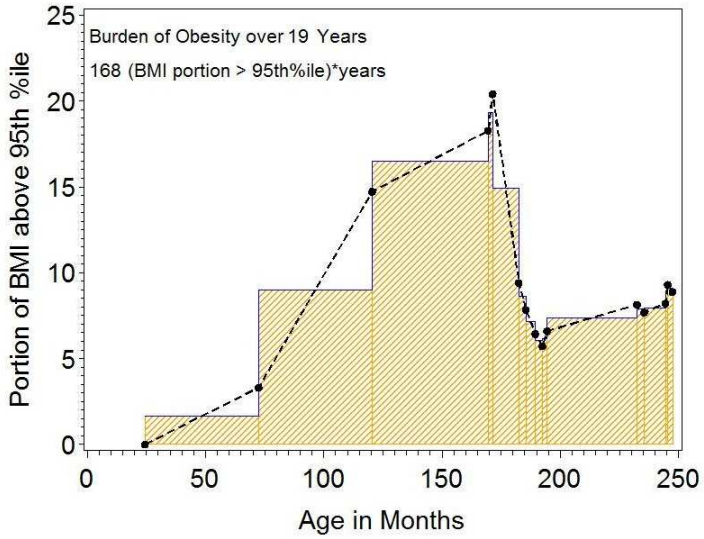
Longitudinal data abound in biomedical research. Examples presented here include such disparate data as body mass index (BMI) recorded throughout childhood, the percent of an infant's feedings that are made up of breastmilk recorded weekly over the first six months of life, and the amount of circulating insulin in an individual's blood stream recorded multiple times during the two hours following the administration of dextrose.

The area under the curve (AUC) relative to a baseline value can be a useful tool to summarize such data. SAS DATA steps along with the MEANS or SQL procedure can be used to calculate the AUC for longitudinal data arranged in a stacked data set.

However, such a summary necessarily involves a loss of information. Subjects with similar AUC values may have very different trajectories, peak amplitudes, or peak timing, which may be important to visualize. The SGPANEL procedure provides a simple method to quickly produce longitudinal plots for multiple individuals. PROC SGPANEL does not have an inset option, so in order to inset each individual's AUC value into the appropriate plot, indicator variables must be created and plotted.

Calculating the Area under a Curve

Riemann sums were used to estimate the area under a curve. The width of each rectangle is the length of time between measurements. The height of each rectangle is the mean of two consecutive measurements. The area of the rectangle is the width multiplied by the height. This is the same as calculating using the trapezoidal rule. The areas of all the rectangles are summed together to estimate the area under the curve.



Mathematically, the summation can be represented by the formula

$$AUC = \sum_{i=2}^{n} \frac{1}{2}(m_{i-1} + m_i) * (t_i - t_{i-1})$$

where m_i corresponds to the i th measurement and t_i corresponds to the i th time point. This is accomplished using the LAG function in SAS. A stacked data set containing multiple observations per subject, one observation for each time point, is sorted by subject ID and time. The following code is used to calculate the area of each rectangle:

```
data forArea;
     set LongData;
     by subjectID;
      prevTime = lag(time);
     prevMsr = lag(measure);
     timeDiff = time - prevTime;       *width of rectangle;
     areaRect = timeDiff*(measure + prevMsr)/2;     *multiply width by mean of;
     if first.subjectID                             * consecutive measurements;
            then do;                   *There is no rectangle for a subject's first;
                  prevTime = .;        *data point alone. Be sure not to use the;
                  prevMsr = .;         *last data point from the previous subject.;
                  timeDiff = .;
                  areaRect = .;
            end;
run;
```

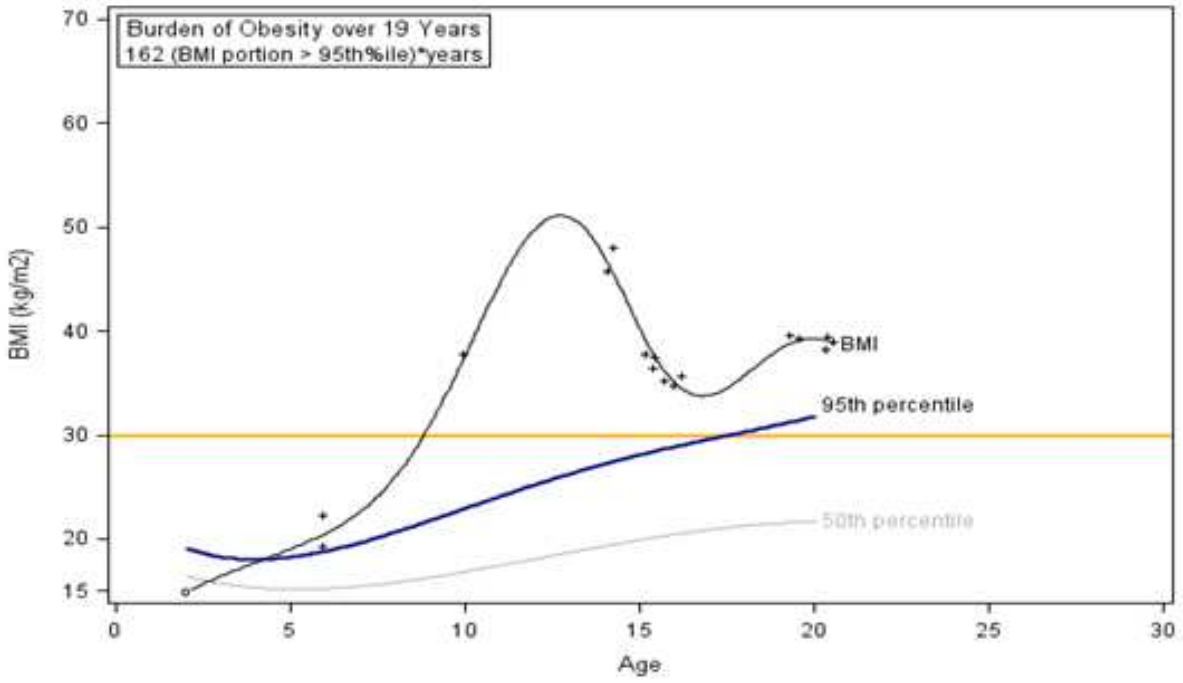
Each subject's rectangular areas must be added together to estimate the AUC. One way to sum the areas together is to use the MEANS procedure with the SUM option. The OUTPUT statement outputs the sum of the areas into a SAS data set which is then merged with the original dataset. The code required is reproduced below:

```
proc means data=forArea sum;
     class subjectID;
     var areaRect;
     output out=summation sum=AUC;
run;

data forArea2;
     merge summation(where=(subjectID ne .) drop _type_ _freq_);
           forArea;
     by subjectID;
run;
```

The new variable, *AUC*, estimates the area under the curve for each individual over the time for which data are available. It is appended to each of the individual's observations.

Example: Burden of Obesity

In children, BMI is indexed to age- and sex-specific growth curves, with ≥95th percentile of the curve indicating clinical obesity. The burden of obesity is an attempt to summarize both the extent and duration of obesity in childhood, by quantifying how far above the 95th percentile an individual's BMI is for how long.

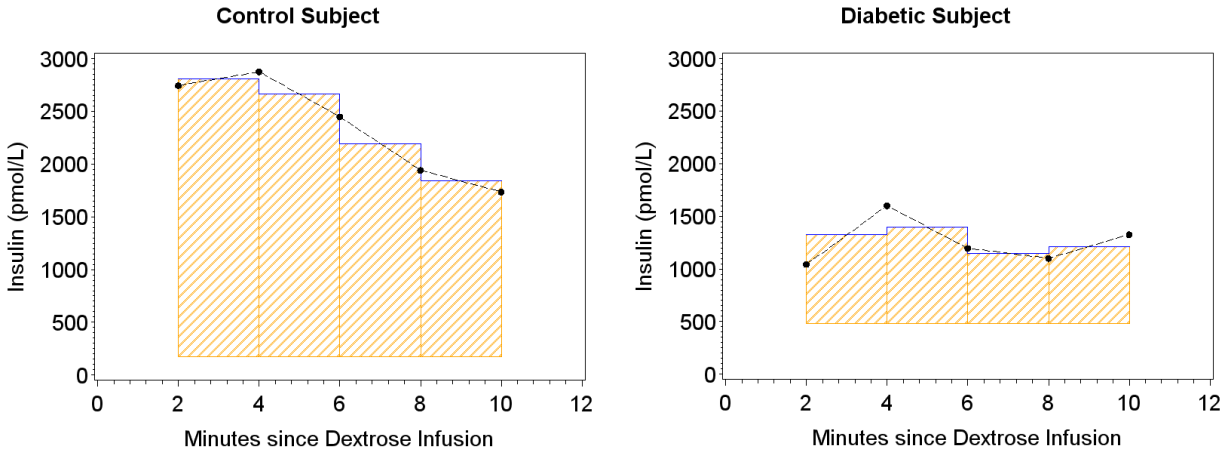


See appendix for code to create graph.

The area between the individual's BMI curve and the 95th percentile for BMI is to be estimated. In order to do that, the difference between the individual's BMI and the 95th percentile is calculated for each time point at which the BMI measurement is above the 95th percentile. If the BMI is lower than the 95th percentile, the difference is set to zero. Calculations proceed as in the code above.

Example: Acute Insulin Response to Dextrose (AIR_{dex})

In non-diabetic individuals, the insulin level in the blood spikes after the administration of dextrose and slowly returns to normal over time. In diabetic individuals, the insulin spike is minimal. The graphs below show the change in insulin levels from baseline from 2 to 10 minutes following an infusion of dextrose. The non-diabetic individual, or control subject, had a much higher insulin spike than the type 2 diabetes (T2DM) individual.



See appendix for code to create graphs.

The area between an individual's insulin response curve and his baseline insulin line (AIR_{dex}) is a measure of insulin secretion, which is suppressed in diabetics. AIR_{dex} can be calculated in the same manner as the burden of obesity. The individual's baseline insulin level is subtracted from the measurements taken after the administration of dextrose. The following code estimates the AIR_{dex} using the AUC methodology. The time of each measurement is stored in the variable *time*, and the insulin measurement is stored in the variable *ins*. *Dex_calc2* contains the longitudinal data.

```
data forDex; set dex_calc2;
     by id;
     prevtime=lag(time);
     prevIns=lag(ins);
     timediff=time-prevtime;
     areaRect=timediff*(ins+prevIns)/2;
     if first.visit
            then do;
                  prevtime = .;
                  prevIns = .;
                  timediff = .;
                  areaRect = 0;
            end;
run;

proc means noprint data=fordex sum;
     class id;
     var areaRect;
     output out=summation2 sum=AIR_dex;
run;

data computeVars3;
     merge computeVars2
     summation2(where=(id ne .) drop = _type_ _freq_);
     by id;
run;
```

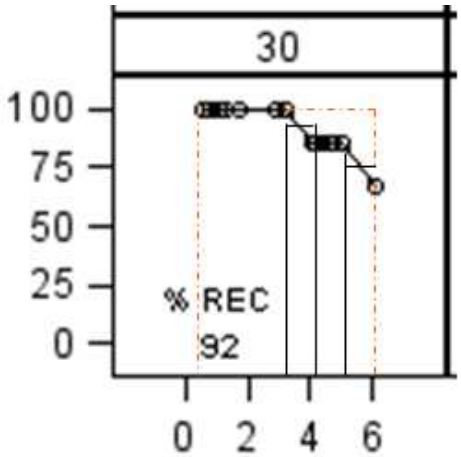
Visualizing the Area under a Curve

While estimating the AUC results in a convenient point estimate to use in analysis, information is lost in regards to the trajectory and shape of the curve. Two individuals with similar AUC estimates may have very different looking curves. A panel of individual plots can be used to look at these differences. The SGPANEL procedure greatly simplifies the creation of a large group of plots that share the same axes. However, it is useful to include the AUC estimate as an inset with each of the plots. PROC SGPANEL does not contain an inset option in SAS 9.2. A new variable containing the title of the inset, along with variables containing the location of the desired inset must be added to the data set. The method is detailed in the following example.

A Breastfeeding Example

The American Academy of Pediatrics (AAP) recommends exclusive breastfeeding for the first six months of an infant's life. Mother-infant pairs were followed for six months and data were collected on what they fed to their infants. In order to see how close they came to meeting the AAP recommendation, the percent of daily feedings that were breastmilk was calculated and the results plotted.

In this example, the proportion of the AAP recommendation met is estimated. A family meeting the AAP recommendation will have a graph that is a horizontal line at 100% for the entire 6 months. The area to calculate would be represented by the orange box in the graph at right. Most families did not meet the recommendation. The ratio of the area under the breastfeeding curve (the black curve) to the area of the orange dotted box estimates the percent of the recommendation that is met.



The initial data set, called food, contains longitudinal data, one observation for each time point for each individual. *AgeDate* is the age of the infant in days, *AgeMonth* is the age of the infant in months, and *pctBMilk* is the percent of feedings that are breastmilk (range 0 to 100) at a given time. The areas of the rectangles are calculated as before:

```
data SixMonths;
     set food(where=(AgeDate <= 182) keep=familyID ageDate ageMonth pctBMilk)
     by familyID;
     prevAge = lag(ageDate);
     prevPct = lag(pctBMilk);
     AgeDiff = ageDate - prevAge;
     areaRect = ageDiff*(pctBMilk + prevPct)/2;
     if first.familyID then do;
                  prevAge = .;
                  prevPct = .;
                  ageDiff = .;
                  areaRect = .;
     end;
run;
```

The SQL procedure gives an alternative method for summing the areas of the rectangles. It will also be used to find the minimum and maximum ages for which data are available, which are needed to calculate the width of the orange rectangle. A new data set called auc is created. It contains the variables from the data set SixMonths and three new variables. *BFInten* is the sum of the rectangular areas for a given subject, *minAge* is the minimum age in days at which data were collected, and *maxAge* is the maximum age in days of data collection. The GROUP BY statement tells SAS to calculate each of those three variables for each family individually. The ORDER BY statement sorts the data set by familyID and age, which will be necessary for the SGPANEL procedure.

```
proc sql;
      create table auc as
            select familyID, ageDate, ageMonth, pctBMilk, prevAge, prevPct, AgeDiff,
            areaRect, finalAge, sum(areaRect) as BFInten, min(ageDate) as minAge,
            max(AgeDate) as maxAge
     from SixMonths
     group by familyID
     order by familyID, ageDate;
quit;
```

The *BFInten* variable must be divided by the area of the orange, dotted rectangle, indicating 100% compliance, to estimate the percent of breastfeeding recommendation met. That rectangle has a height of 100 and a width of *maxAge – minAge.*

```
data auc2; set auc;
     PctRecBM = (BFInten/(100*(maxAge-minAge)))*100;
run;
```

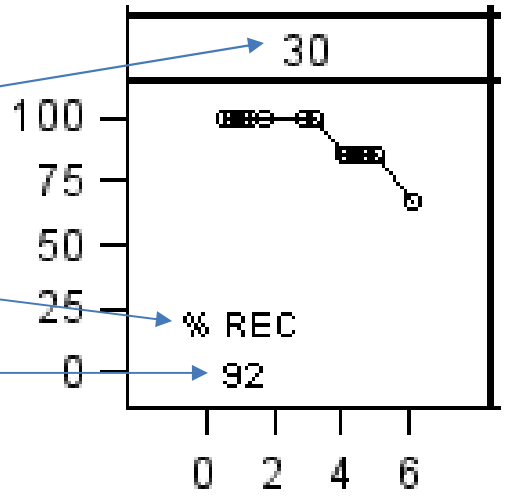
Finally, the data set used to create the panel graph can be created. Four additional variables are necessary to add the inset. *xLoc* contains the x-coordinate for the placement of the inset on each individual graph. *yLoc_pct* contains fhe y-coordinate for the placement of the percent recommendation variable and *yLoc_title* contains the y-coordinate for the placement of the title. The values of the coordinates will vary with from application to application. Finally, the variable *text* contains the text of the inset title. The inset should only be placed on the graph once, so these variables are only assigned values for the first observation for each family and are set to missing for succeeding observations.

```
data auc3; set auc2; by familyID;
     if first.familyID then do;
                  xLoc = 1;
                  yLoc_pct = 0;
                  yLoc_title = 20;
                  text = "% REC";
     end;
     else do;
                  xLoc = .;
                  yLoc_pct = .;
                  yLoc_title = .;
                  text = " ";
     end;
run;
```

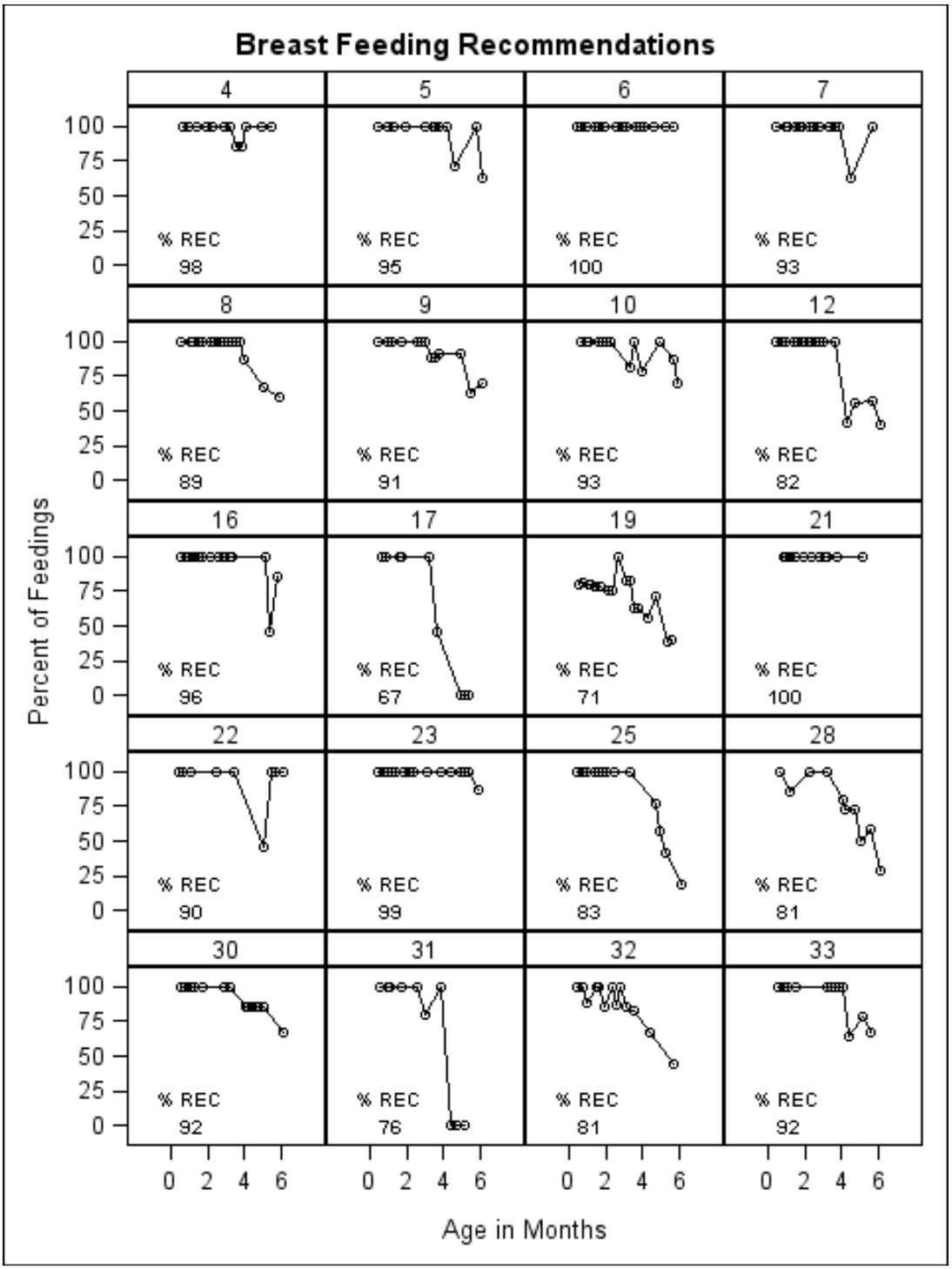
The following code produces the set of graphs and stores them as a set of 5 cell by 4 cell panels, PercRec.png, PercRec1.png, PercRec2.png, etc., using the SGPANEL procedure.

- The NOAUTOLEGEND in the PROC SGPANEL statement suppresses the printing of a default legend.
- The FORMAT statement limits the inset value, *PctRecBM*, to integer values of no more than three digits. This allows for any of the possible values of *PctRecBM*, which runs from 0 to 100, inclusive, and avoids long decimals.
- The PANELBY statement causes individual graphs to be made for each family, with five rows and four columns (20 graphs) per page.
- The NOVARNAME option suppresses the default "FamilyID =" before each family ID in the heading for the individual graphs.
- The first SCATTER statement adds the inset title, which was defined to be the phrase "% REC" and is contained in the variable "text." The markerchar option uses the value of *text* as the scatter plot marker. The coordinates for the placement of the title were stored in the variables *xLoc* and *yLoc_title*.
- The second SCATTER statement places the inset value, the percent of compliance with the AAP recommendation, below the title. Each family's percent of the breast feeding recommendation met is stored in *PctRecBM* which is used as the scatter plot marker. The coordinates for the placement of the inset value were stored in the variables *xLoc* and *yLoc_pct*.
- Finally, the SERIES statement graphs the individual data points for each individual, joined by line segments.
- The ROWAXIS and COLAXIS statements define the tick values and titles for the axes. Note that the ROWAXIS statement refers to the values for the y-axis and the COLAXIS statement refers to the values for the x-axis.

```
ods graphics on / imagename="PercRec";
title "Breast Feeding Recommendations";
proc sgpanel data=auc3 noautolegend;
     format PctRecBM 3.;
     panelby familyID
            / rows=5 columns=4 novarname;
     /*Place inset title*/
     scatter y=yloc_title x=xLoc
            / markerchar = text
              markercharattrs = (weight=normal);
     /*Place inset value*/
     scatter y=yloc_pct x=xLoc
            / markerchar = PctRecBM
              markercharattrs = (weight=bold);
     /*Graph individual data as a series plot*/
     series y=pctBMilk x=ageMonth
            /markers markerattrs=(symbol=Circle);
     colaxis values=(0 to 6 by 2)
             label = "Age in Months";
     rowaxis values=(0 to 100 by 25) label = "Percent of Feedings";
run;
ods graphics off;
```



The first panel created by this code is reproduced on the following page. An individual cell is recreated on a larger scale above.



Note that families 17 and 19 have similar values for percent compliance with the AAP guideline, at 67% and 71%, respectively. Without seeing the breastfeeding graphs of each family, one might conclude that they were similar in terms of breastfeeding. However, notice that their breastfeeding trajectories look very different. The infant in family 17 was breastfed exclusively for about three months, after which the percent of breastmilk feedings dropped off precipitously to 0% by about four months. In contrast, the infant in family 19 received feedings that were not breastmilk from shortly after birth but was still receiving breastmilk as about 33% of feedings at six months.

Both pieces of information, the percent compliance with AAP breastfeeding recommendations and the trajectory of each family's breastfeeding data, shed light on what breastfeeding looks like for these families during the first six months of life.

Conclusion

The SAS DATA step can be utilized to estimate the area under a curve (AUC) which is a useful summary of longitudinal data. However, as with other summary values, information is necessarily lost in the process. Adding a graphical representation of the longitudinal data can preserve the information and clarify differences in trajectory, peak timing, and peak amplitude between subjects with similar AUC values. The SGPANEL procedure provides a simple way to visualize the data, and the summary value can be added to each cell with minimal difficulty.

References

Carpenter, Art. 1999. *Annotate: Simply the Basics.* Cary, NC. SAS Institute Inc.

Acknowledgements

I would like to thank Meredith Tabangin, Jesse Pratt, and Matthew Fenchel for their insights and help.

Contact Information

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

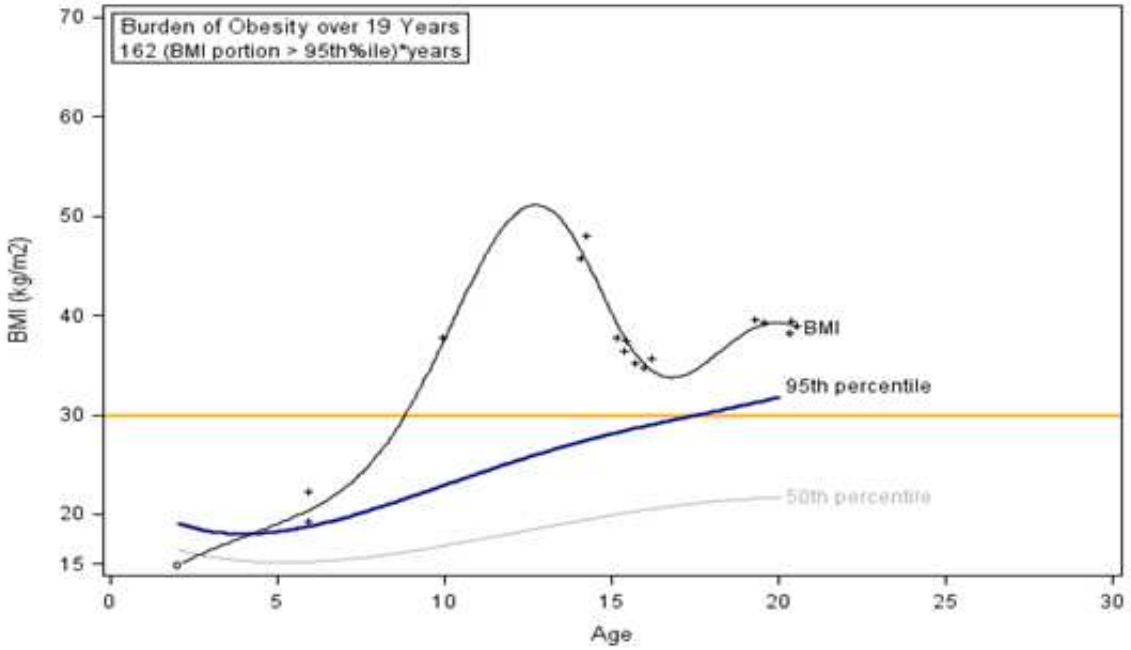
Patricia M Herbers
Division of Biostatistics and Epidemiology
Cincinnati Children's Hospital Medical Center
MLC 5041, 3333 Burnet Avenue
Cincinnati, OH 45229

E-mail: Patricia.Herbers@cchmc.org

Appendix

Burden of Obesity Graph

The following macro creates a series of plots, one for each of 15 obese adolescents undergoing gastric bypass surgery, depicting their BMI trajectories over time. The annual burden of obesity is inset in the plot.



The data set called "index" is used to create a set of macro variables. It includes the following variables:

subjectID = subject identifier
early = age in months at earliest BMI measurement
late = age in months at most recent BMI measurement
sex = gender of subject
BurObYrs = burden of obesity in years
NumYears = number of years between initial and most recent BMI measurements

The data set called "MyLib.WgtHgtBMI" contains longitudinal BMI and BMI percentile data for the subjects in the study. BMI percentiles are based on CDC growth charts.

ageCat = age in months at measurement
sex = gender of subject
BMI = BMI measurement
BMIpct = BMI percentile

The data set called "MyLib.BMIperc" contains BMI percentiles from CDC growth charts. It is used to add the 50th and 95th percentile curves to the graph.

ageCat = age in months at measurement
sex = gender for percentiles
p50 = 50th percentile BMI values
p95 = 95th percentile BMI values

The TRANSREG procedure was used to estimate the shape of each subject's BMI curve using pbsplines. See SAS online documentation for further explanation.

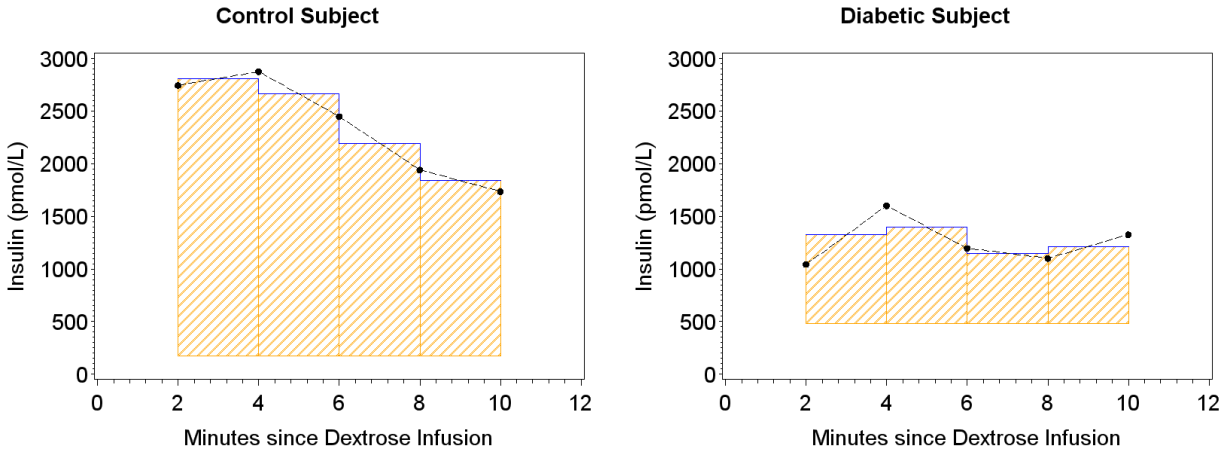
The SGPLOT procedure generates the graphs.

- The REFLINE statement adds a horizontal reference line.
- The SERIES statements add the three curves, one for the individual's trajectory and the others for the 50th and 95th percentiles. Notice that the SERIES statement for the BMI curve uses the x2axis, which is in months instead of years.
- The SCATTER statement adds the individual data points. This is useful to see how the data compare to the estimated curve.
- The INSET statement adds the inset in the upper left corner.
- The X2AXIS statement then hides the display of the second x-axis.

```
proc format;        /*Format for symbols on graph*/
      value mark 0 = 'o'
                 1 = '+';
run;

%macro PlotIt;
      %do i = 1 %to 15;
            data _null_; set index;
                  if _n_ = &i then do;
                  call symput('ID',trim(left(put(subjectID,8.))));
                  call symput('low',trim(left(put(early,8.))));
                  call symput('high',trim(left(put(late,8.))));
                  call symput('gender',trim(left(put(sex,8.))));
                  call symput('BOB',trim(left(put(BurObYrs,8.))));
                  call symput('Years',trim(left(put(NumYears,8.))));
            end;
            run;
/*Data for individual subject*/
            data sub&ID; merge MyLib.WgtHgtBMI(where=(subjectID=&ID) in=inWtHt)
                   MyLib.BMIperc(where=(sex=&gender) keep=sex ageCat p50 p95);
                  by sex ageCat;
                  format over95 mark.;
                  if (inWtHt =1 and BMIpct ne .) then do;
                        if BMIpct >= 95 then over95 = 1;
                              else if 0 < BMIpct < 95 then over95 = 0;
                  end;
                  else BMIpct = .;
/*Generate data predicted values for BMI for age using penalized B-splines*/
            proc transreg data=sub&ID solve ss2;
                  model identity(BMI) = pbspline(ageCat/nknots=20 degree=3);
                  output out=regOutput predicted pdp=p;

            proc sort data=regOutPut; by ageCat; run;

            data regOut2; set regOutput; by ageCat;
                  if first.ageCat;
                  keep ageCat pBMI;

            data ForPlot; merge regOut2 (where=(&low <= ageCat <= &high)) sub&ID;
                  by ageCat;

            ods graphics / imagename="BMI_loess &ID";
            proc sgplot data=ForPlot description="ID&ID";
                  label ageCat = "Age";
                  label BMI = "BMI (kg/m2)";
                  label over95 = "BMI at or above 95th percentile";
                  refline 30 / axis=y lineattrs = (color=Orange thickness=2);
                  series x=ageCat y=p50/ transparency = 0.7
                                         curvelabel="50th percentile" x2axis;
                  series x=ageCat y=p95 / lineattrs = (color=DarkBlue thickness=2)
                                          curvelabel="95th percentile" x2axis;
                  scatter x=age y=BMI/ nomissinggroup markerchar=over95;
                  series x=ageCat y=pBMI/ curvelabel="BMI" x2axis;
                  inset "&BOB (BMI portion %nrstr(> 95th%ile))*years"
                        / position = TOPLEFT border
                          title = "Burden of Obesity over &Years Years";
                   x2axis display=(nolabel noticks novalues)
                          values=(0 to 360 by 60);
                  xaxis values=(0 to 30 by 5);
            run;
            ods graphics off;  *stop graphics output from transreg procedure;
      %end;
%mend PlotIt;

%PlotIt; run;
```

Acute Insulin Response to Dextrose (AIR_{dex}) Graph

The following code uses the GPLOT procedure and ANNOTATE data sets to create the graphs showing the change in insulin levels from baseline for the first 10 minutes following an infusion of dextrose. Refer to *Annotate: Simply the Basics* by Art Carpenter for a very readable introduction to annotate data sets.



The data set called "MyLIb.BaselineAnalysis" contains the necessary data, but it is not in the correct format. Each subject has one observation containing all of the insulin measurements. It must be transposed to a stacked data set with five observations per subject, one each at 2, 4, 6, 8, and 10 minutes after the infusion.

id = subject identifier
Bdex_ins = baseline insulin level before dextrose infusion
I2_p2 = insulin measurement two minutes after dextrose infusion
I2_p4 = insulin measurement four minutes after dextrose infusion
I2_p6 = insulin measurement six minutes after dextrose infusion
I2_p7 = insulin measurement eight minutes after dextrose infusion
I2_p10 = insulin measurement ten minutes after dextrose infusion

```
/*Create stacked data set*/
data forPlot; set MyLIb.BaselineAnalysis(where = (id in ('001c' '055d')));
       length time 3 ins Bdex_ins 4;
       time = 2; ins = round(I2_p2); base=round(Bdex_ins); output;
       time = 4; ins = round(I2_p4); base=round(Bdex_ins); output;
       time = 6; ins = round(I2_p6); base=round(Bdex_ins); output;
       time = 8; ins = round(I2_p8); base=round(Bdex_ins); output;
       time = 10; ins = round(I2_p10); base=round(Bdex_ins); output;
       keep time ins id base;
run;

/*Need previous and current data to draw rectangles*/
data forPlot2; set forPlot;
       by ID;
       prevTime = lag(time);
       prevIns = lag(ins);
       timeDiff = time - prevTime;
       if first.ID then do;
              prevTime = .;
              prevIns = .;
              timeDiff = .;
       end;
       midY = (ins + prevIns)/2;
       midX = Time - TimeDiff/2;
run;

proc sort data=forPlot2; by ID time; run;
```

```
/*Annotate data set to make bars indicating the rectangles used in area estimation*/
data step; set forPlot2(where=(prevTime ne .));
       by ID;
       length color function $8;
       retain color 'blue' xsys ysys '2';
              y = midY;
              x = prevTime;
              if first.ID then function = 'move';
                     else function = 'draw';
       output;
               x = Time;
              function = 'draw';
       output;
       keep ID function x y color xsys ysys;
run;

/*Annotate data set to fill bars orange*/
data bar; set forPlot2(where=(prevTime ne .));
       by ID;
       length color function style $8;
       retain color 'orange' xsys ysys '2' style 'R3' line 0;
              function = 'move';
              x = prevTime;
              y = base;
       output;
              function = 'bar';
              x = time;
              y = midY;
       output;
       keep ID function x y color xsys ysys style line;
run;

data step2; set step;
       n = _n_ + 10000;           /*to keep proper order in combined set*/
run;

data bar2; set bar;
       n = _n_;
run;

data StepBar; set bar2 step2; run;

proc sort data=StepBar; by id n; run;

filename aucout "H:\Presentations\MWSUG2011\stepbar001c.gif";
goptions reset=all;
goptions gsfname=aucout device=gif gsfmode=replace;
symbol1 v=dot i=join c=black l=3;
axis1 order=(0 to 3000 by 500) label=(angle=90 rotate=0 "Insulin (pmol/L)");
axis2 order=(0 to 12 by 2) label=("Minutes since Dextrose Infusion");
title "Control Subject";
proc gplot data=forPlot2;
       where ID='001c';
       plot ins*time=1 / anno=stepBar(where=(ID='001c')) vaxis=axis1 haxis=axis2;
run;quit;

filename aucout "H:\Presentations\MWSUG2011\stepbar055d.gif";
goptions reset=all;
goptions gsfname=aucout device=gif gsfmode=replace;
symbol1 v=dot i=join c=black l=3;
axis1 order=(0 to 3000 by 500) label=(angle=90 rotate=0 "Insulin (pmol/L)");
axis2 order=(0 to 12 by 2) label=("Minutes since Dextrose Infusion");
title "Diabetic Subject";
proc gplot data=forPlot2;
       where ID='055d';
       plot ins*time=1 / anno=stepBar(where=(ID='055d')) vaxis=axis1 haxis=axis2;
run;quit;
```