

Paper LS-142

Sankey Diagram with Incomplete Data – From a Medical Research Perspective

Yichen Zhong, Merck & Co., Inc., Upper Gwynedd, PA USA

ABSTRACT

Sankey diagram is widely used in energy industry but relatively rare in medical research. Sankey diagram is an innovative tool in visualizing patient flow in longitudinal data. A SAS® Macro for generating Sankey Diagram with Bar Charts was published in 2015. However, it did not provide an intuitive solution for missing data or terminated data, which is common in medical research. This paper presents a modification of this macro that allows subjects with incomplete data to appear on the Sankey diagram. In addition, examples of using Sankey diagram in medical research are also provided.

KEYWORDS

Sankey Diagram, Medical Research, Missing Data, Terminated Data

INTRODUCTION

Sankey diagram is a type of flow chart showing factors' states and transitions with the width of the arrows representing the size of the flow. It was first developed in 1898 to track the energy flow in a steam engine.[1] Sankey diagram is now widely used as an effective graphic illustration of energy, material, or money flow. In recent years, the interest of using Sankey diagrams in medical research is growing. Sankey diagram has important advantages in visualizing cohort distribution, treatment pattern, longitudinal change of clinical characteristics, and disease progression.[2]–[5] It can also be used to explore underlying association among factors.[4] Currently available tools for creating Sankey diagrams with SAS® include SAS® Visual Analytics 7.1, a web-based product, and the “Getting Sankey with Bar Charts” macro, which was published at PharmaSUG in 2015.[6] SAS® Visual Analytics has built-in functions for easy graphic customizations, including title, color, label, layout, and other nodes/path options, however separate subscription for SAS® Visual Analytics is required. As an alternative, SAS® users can produce Sankey diagrams using the “Getting Sankey with Bar Charts” macro. Similar to SAS® Visual Analytics, the minimum data requirements for the “Getting Sankey with Bar Charts” macro are: event (e.g., state, characteristics), sequence order (e.g., time point), and transaction identifier (e.g., patient ID). However for both tools, subjects who do not have data at every sequence order are excluded in the final chart. Incomplete data is common and sometimes informative in medical research. An option to include subjects with incomplete data can make the macro more useful for medical researchers. This paper presents a modification of the “Getting Sankey with Bar Charts” macro to allow subjects who had missing data or terminated data to appear in the Sankey diagram.

DESCRIPTION OF SAS MACRO MODIFICATIONS

The previously published “Getting Sankey with Bar Charts” macro allows only subjects with data at every time point to appear in the Sankey diagram. However in medical research, patients may not have data at each of the time points for different reasons. The following modifications allow user to include patients with incomplete data in the Sankey diagram. The “Sankey with Bar Charts Macro” contains two submacros: %RawToSankey and %Sankey. The sections below describe major modifications made to each of the submacros.

1. Modification of the %RawToSankey submacro

This submacro converts one input dataset into a dataset with Sankey nodes (i.e., bars, patient distribution at each time point) and another dataset with Sankey links (i.e., bands, patient flow from one state to another). In the Sankey diagram, x values are sequence orders (e.g., time points) while y values are events (e.g., states). The original submacro includes codes that exclude subjects with number of non-missing records less than the maximum number of possible time points for each subject (i.e., &xmax). In this modification, these codes are commented out so these subjects can be included in the two summary datasets.

```
proc sql;
  create table _nodes30 as
  select *
  from _nodes20
  order by &subject
/*MODIFICATION: comment out codes selecting patients with data at every time points*/
  *group by &subject;
  *having count(*) eq &xmax;
;
quit;
```

2. Modification of the %Sankey submacro

- i. Macro parameter YFMT is added to allow user to specify legend format. Previous parameter PERCENTS is renamed to DATALABEL to allow user to specify whether data label will be displayed (i.e., yes vs. no). Macro parameter LABELTYPE is added to allow user to specify data label type (i.e., percent vs. frequency).

```
%sankey_modify
  (sankeylib= /*default: work*/
  ,colorlist= /*default: cxa6cee3 cx1f78b4 cxb2df8a cx33a02c cxfb9a99 cxe31a1c
                cxfdbf6f cxff7f00 cxcab2d6 cx6a3d9a cxffff99 cxb15928*/
  ,barwidth= /*default: 0.25*/
/*MODIFICATION 1: added macro parameter YFMT to allow user to specify legend format*/
  ,yfmt= /*e.g., risk_score_fmt.*/
  ,xfmt= /*e.g., time_fmt.*/
  ,legendtitle= /*e.g., %quote(Risk Score)*/
  ,interpol= /*valid value: cosine or linear*/
/*MODIFICATION 2: renamed macro parameter from PERCENTS to DATALABEL*/
  ,datalabel= /*valid value: yes or no*/
/*MODIFICATION 3: added macro parameter LABELTYPE to allow user to specify data label type*/
  ,labeltype= /*valid value: percent or frequency*/
  );
```

- ii. In the previously published macro, the bars are drawn using HIGHLOW statements on the percentage scale. In this modification, the bars are drawn on the frequency scale. The _ctfhl dataset is a frequency table (i.e., event by sequence order) generated from the nodes summary dataset. Cumulative frequencies by time (e.g., x) are then calculated in the below data step as the high/low values for the bars. Cumulative percentages by time are calculated for data labels.

```
data _highlow;
    set _ctfhl;
    by x;
    node = _N;
    /*MODIFICATION 1: change the bars and bands scale to frequencies*/
    retain cumcount;
    if first.x then cumcount = 0;
    low = cumcount;
    high = cumcount + frequency;
    cumcount = high;
    /*MODIFICATION 2: calculate percentages for data label*/
    retain cumpct;
    if first.x then cumpct = 0;
    low_pct = cumpct;
    high_pct = cumpct + rowpercent;
    cumpct = high_pct;

    keep x y node low high low_pct high_pct;
run;
```

- iii. The following modifications include options for data label type (frequency vs. percent) and legend format (no format vs. user specified format). The modifications are made to the data step that creates the _highlow_statements dataset. Legend label can be defined using user specified y-axis format. Data label type can be frequency or percent based on user's input.

```
/*MODIFICATION 1: modify the following codes to allow user to specify legend format*/
legendlabel_temp = "&&yvarord&jro";
legendlabel = put(legendlabel_temp,&yfmt.);
/*MODIFICATION 2: modify scatter statement to allow data label options*/
%if &labeltype.=frequency %then %do;
    mean = mean(low,high);
    freq = high - low;
    if freq >= 1 then do;
        mean&jc = mean;
        textb&jc = compress(put(freq,6.));
        scatter = "scatter x=xb&jc y=mean&jc / x2axis markerchar=textb&jc;";
    end;
%end;

%if &labeltype.=percent %then %do;
    mean = mean(low,high);
    percent = high_pct - low_pct;
    if percent >= 1 then do;
        mean&jc = mean;
        textb&jc = compress(put(percent,3.)) || '%';
        scatter = "scatter x=xb&jc y=mean&jc / x2axis markerchar=textb&jc;";
    end;
%end;
```

- iv. The following modifications revise the left edge of each band to allow missing data. The cumulative frequencies calculated in the `_highlow` dataset are merged with the links summary dataset. For each bar, the first band of the first event category starts at 0. The first band of the other event categories starts at the cumulative frequency of the last event category.

```
/*MODIFICATION 1: merge cumulative frequencies with links summary dataset*/
proc sql;
  create table _nodes_join_left as
  select a.*, b.high as cumthickness from
  links a inner join _highlow (where=(high~=low)) b
  on a.x1=b.x and a.y1=b.y
  order by x1, y1, x2, y2;
quit;
/*MODIFICATION 2: allow missing data in bands statement*/
data _links2 (drop=lastcumthickness yblow0);
  set _nodes_join_left;
  if cumthickness=. then cumthickness=0;
  by x1 y1 x2 y2;
  link = _N_;
  xt1 = x1;
  retain lastcumthickness yblow0 ybhigh1;
  if first.x1 then lastcumthickness = 0;
  if first.x1 or first.y1 then do;
    yblow0=lastcumthickness;
    ybhigh1=lastcumthickness;
  end;
  lastcumthickness = cumthickness;
  ybhigh1=ybhigh1+thickness;
  yblow1=ybhigh1-thickness;
  proc sort;
  by x2 y2 x1 y1;
run;
```

- v. Similarly, the right edge of each band is modified to allow missing data.

```
/*MODIFICATION 1: merge cumulative frequencies with links summary dataset */
proc sql;
  create table _nodes_join_right as
  select a.*, b.high as cumthickness from
  links a inner join _highlow (where=(high~=low)) b
  on a.x2=b.x and a.y2=b.y
  order by x2, y2, x1, y1;
quit;
/*MODIFICATION 2: allow missing data in bands statement*/
data _links3_temp (drop=lastcumthickness yblow0);
  set _nodes_join_right;
  if cumthickness=. then cumthickness=0;
  by x2 y2 x1 y1;
  xt2=x2;
  retain lastcumthickness yblow0 ybhigh2;
  if first.x2 then lastcumthickness = 0;
  if first.x2 or first.y2 then do;
    yblow0=lastcumthickness;
    ybhigh2=lastcumthickness;
  end;
  lastcumthickness = cumthickness;
  ybhigh2=ybhigh2+thickness;
  yblow2=ybhigh2-thickness;
run;

data _links3;
  merge _links2 _links3_temp(keep=x2 y2 x1 y1 xt2 ybhigh2 yblow2);
  by x2 y2 x1 y1;
run;
```

- vi. Since the scale of the original macro is percentage, the high/low values are multiplied by 100. In the modified data step that creates the `_band_statements` dataset, the scale is changed from percentage to frequency.

```
/*MODIFICATION: change the HIGHLOW bar chart scale from percentage to frequency.
*yblow&jc = yblow*100;
*ybhigh&jc = ybhigh*100;

yblow&jc = yblow;
ybhigh&jc = ybhigh;
```

- vii. In the SGPLOT procedure, change the y-axis label from "Percent" to "Frequency".

```
/*MODIFICATION: In the SGPLOT procedure, change the y-axis label to "Frequency"*/
yaxis offsetmin=0.02 offsetmax=0.02 label="Frequency" grid;
```

In the next section, examples of generating Sankey Diagram using medical data are provided.

EXAMPLES OF GENERATING SANKEY DIAGRAM USING MEDICAL DATA

Using Sankey diagram to visualize treatment pattern

In this example, treatment history up to the fourth line for 20 patients was collected. Patients might not have data for each line even with the assumption of no missing data. For example, incompleteness could occur when patients were still receiving early lines of therapy at the time of the analysis or patients discontinued early lines but did not receive further treatment. Data of three patients are shown below:

	id	line_number	therapy
1	1	1	1
2	1	2	3
3	2	1	1
4	2	2	2
5	2	3	4
6	2	4	3
7	3	1	2
8	3	2	1
9	3	3	4

→ Patient 1 received only two lines of therapy

→ Patient 3 received a total of three lines of therapy

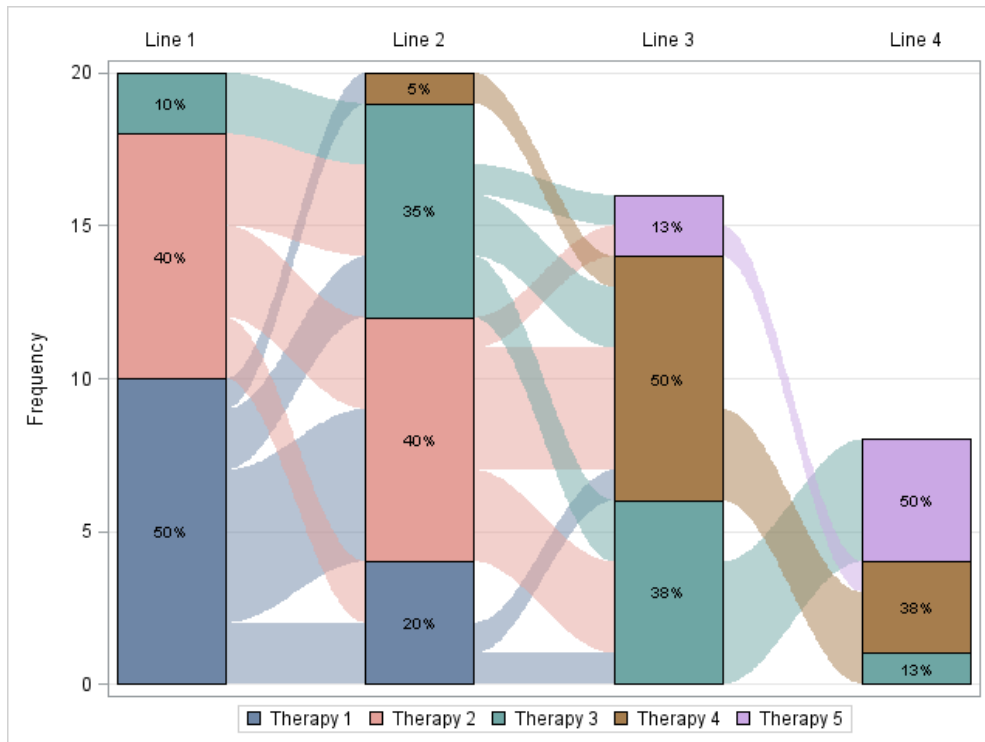
Option 1: Produce the Sankey diagram using original data

Macro calls for Example 1 (Option 1):

```
%rawtosankey_modify
  (data=example1_option1
   ,subject=id
   ,yvar=therapy
   ,xvar=line_number
   ,outlib=work
   ,yvarord=%str(1, 2, 3, 4, 5)
   ,xvarord=%str(1, 2, 3, 4)
  );

%sankey_modify
  (sankeylib=work
   ,colorlist=LIGB LIYPK LIBG MOO VPAV GWH BWH LIOLBR
   ,barwidth=0.45
   ,yfmt=therapy_fmt.
   ,xfmt=line_fmt.
   ,legendtitle=
   ,interpol=cosine
   ,datalabel=yes
   ,labeltype=percent
  );
```

Example 1 (Option 1). Sankey diagram for treatment pattern generated using original data



In this diagram, percentages reflect **the distribution of patients who received the specified line of therapy**. The diagram is clean and focused on patients receiving the actual treatment. This type of diagram is informative in market share research. Gaps between the left edges of the bands represent patients who did not receive further treatment.

Option 2: Modify the data and create separate categories for missing treatment lines

	id	line_number	therapy		id	line_number	therapy_missing
1	1	1	1		1	1	1
2	1	2	3		2	1	2
3	2	1	1		3	1	3
4	2	2	2		4	1	4
5	2	3	4		5	2	1
6	2	4	3		6	2	2
7	3	1	2		7	2	3
8	3	2	1		8	2	4
9	3	3	4		9	3	1
					10	3	2
					11	3	3
					12	3	4

Discontinued treatment and no subsequent line

Still on the third line

Macro calls for Example 1 (Option 2):

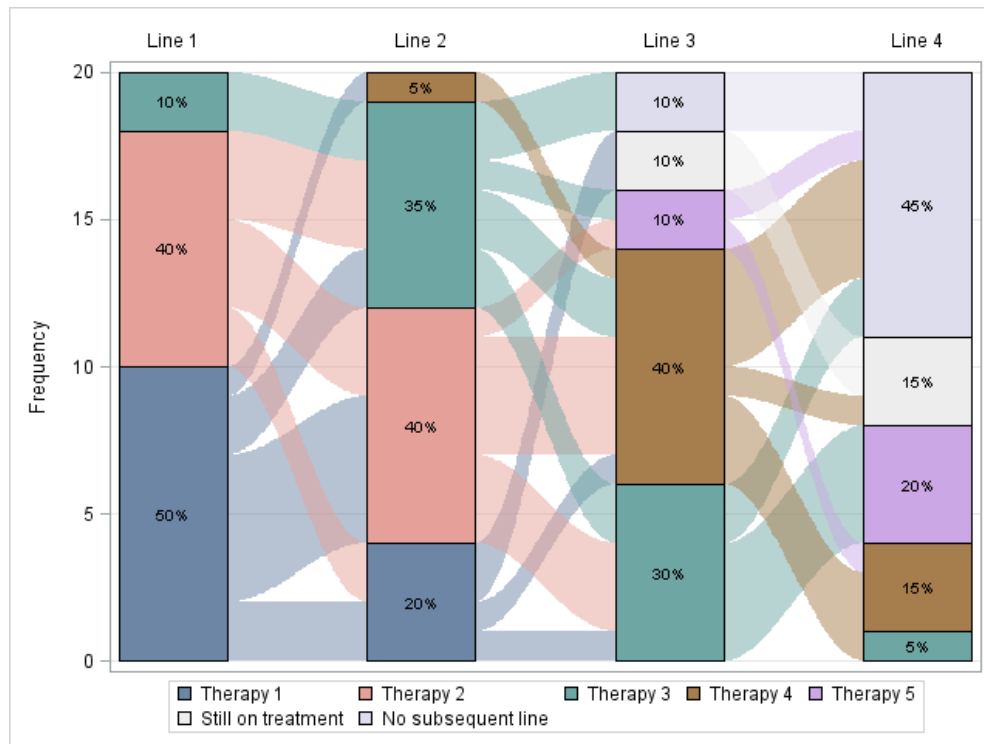
```

%rawtosankey_modify
  (data=example1_option2
  ,subject=id
  ,yvar=therapy
  ,xvar=line_number
  ,outlib=work
  ,yvarord=%str(1, 2, 3, 4, 5, 98, 99)
  ,xvarord=%str(1, 2, 3, 4)
  );

%sankey_modify
  (sankeylib=work
  ,colorlist=LIGB LIYPK LIBG MOO VPAV GWH BWH LIOLBR
  ,barwidth=0.45
  ,yfnt=therapy_fmt.
  ,xfnt=line_fmt.
  ,legendtitle=
  ,interpol=cosine
  ,datalabel=yes
  ,labeltype=percent
  );

```

Example 1 (Option 2). Sankey diagram for treatment pattern generated by adding categories for missing treatment lines



In this diagram, percentages reflect **the distribution of all patients who entered the study**. It provides information on patients who did not receive four lines of therapy. The percentages do not reflect the distribution of patients receiving the actual treatment. This type of diagram can be useful for studying cohort distribution. This diagram can also be generated using the previously published macro.

Using Sankey diagram to visualize longitudinal change of patient characteristics

In this example, 100 patients were followed monthly and the risk score was measured during each visit. In the real-world, patient often enter the study at different time. At the time of the analysis, patient may have different numbers of visits. Furthermore, patients may miss one or more visits during the follow-up. Data of three patients are shown below:

	id	visit	risk_score
1	1	1	0
2	1	2	1
3	1	3	1
4	1	4	0
5	1	5	2
6	2	1	0
7	2	2	2
8	2	3	1
9	2	4	2
10	2	5	2
11	3	1	0
12	3	2	0
13	3	4	2

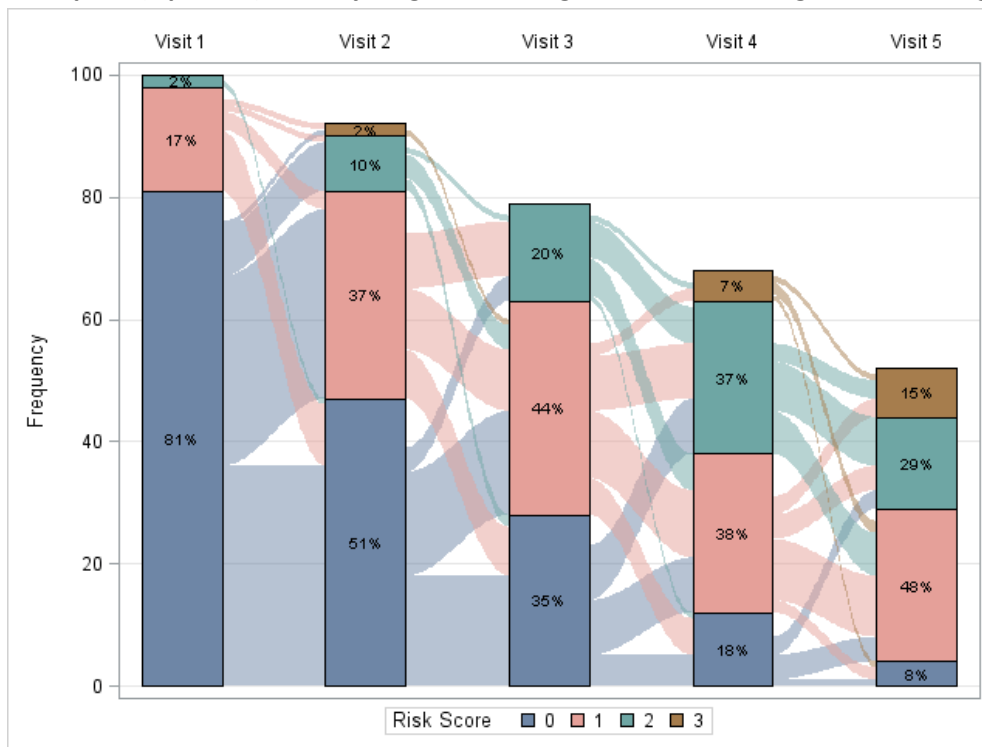
Patient 3 missed the third visit. The analysis was conducted before he/she had the fifth visit

Option 1: Produce the Sankey diagram using original data

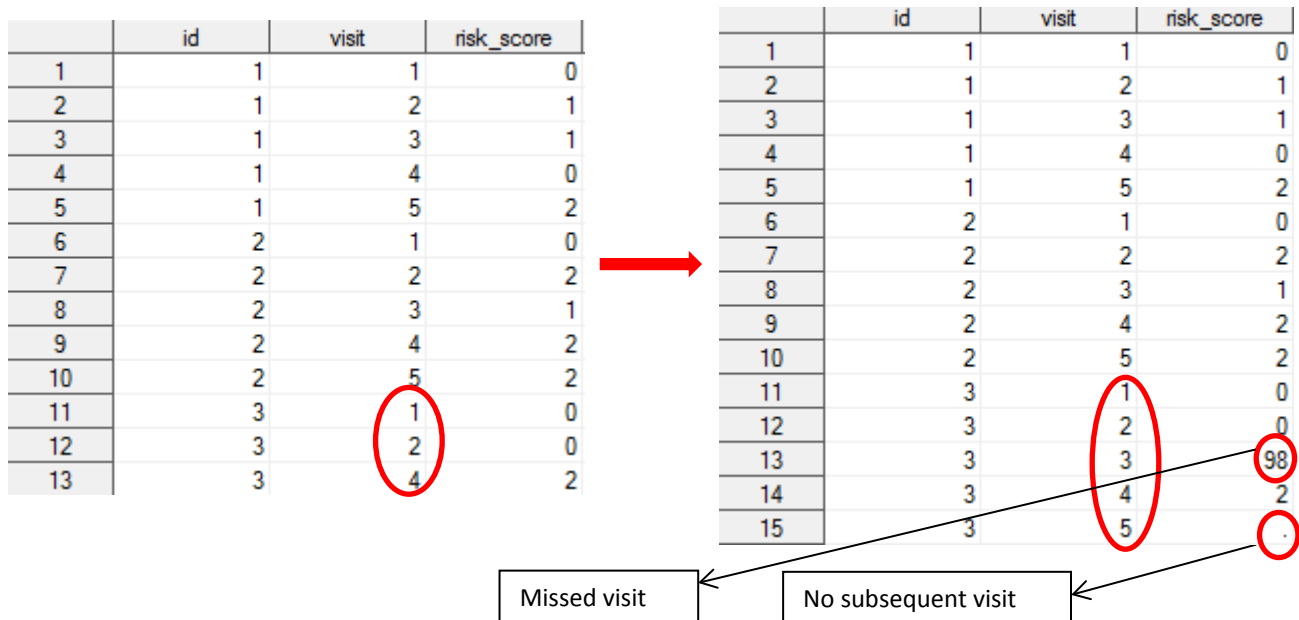
Macro calls for Example 2 (Option 1):

```
%rawtosankey_modify
  (data=example2_option1
  ,subject=id
  ,yvar=risk_score
  ,xvar=visit
  ,outlib=work
  ,yvarord=%str(0, 1, 2, 3)
  ,xvarord=%str(1, 2, 3, 4, 5)
  );

%sankey_modify
  (sankeylib=work
  ,colorlist=LIGB LIYPK LIBG MOO VPAV GWH BWH LIOLBR
  ,barwidth=0.45
  ,yfmt=risk_score_fmt.
  ,xfmt=visit_fmt.
  ,legendtitle=%str(Risk Score)
  ,interpol=cosine
  ,datalabel=yes
  ,labeltype=percent
  );
```

Example 2 (Option 1). Sankey diagram for longitudinal risk score generated using original data


In this diagram, percentages reflect **the distribution of patients who had the risk score measured during each visit**. Gaps between the left edges of the bands represent patients who missed the next visit or did not have subsequent visit. Gaps between the right edges of the bands represent patients who missed the last visit. This diagram shows the trends of patient flow and all data points collected. But it does not differentiate the reasons for missing data.

Option 2: Modify the data and include missed visit as a category of risk score


Macro calls for Example 2 (Option 2):

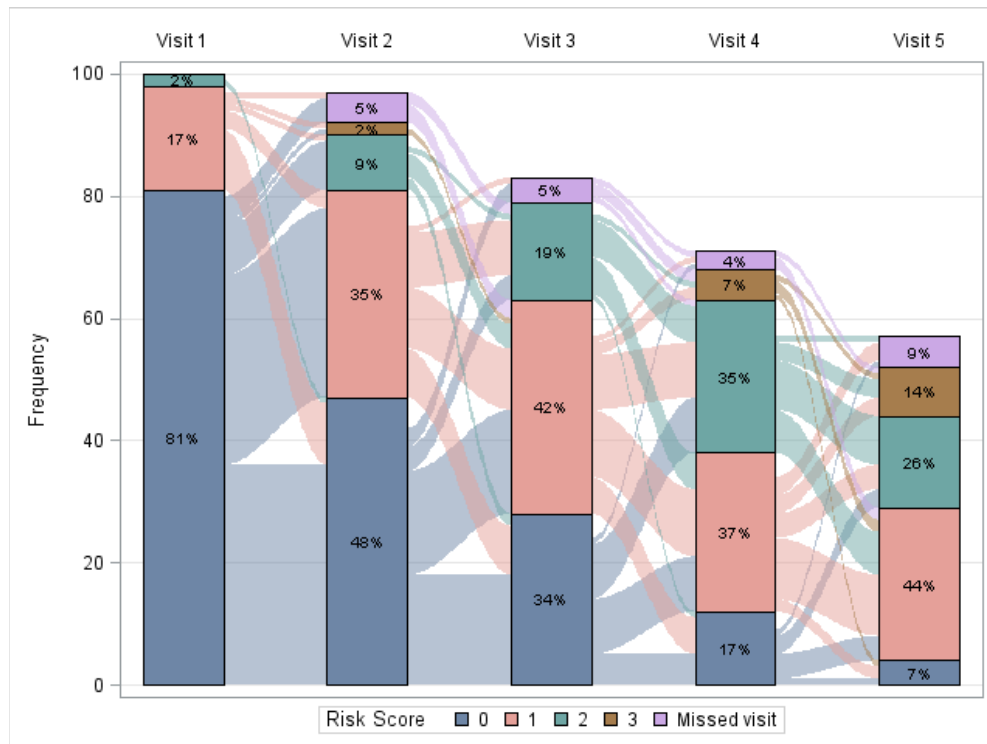
```

%rawtosankey_modify
  (data=example2_option2
  ,subject=id
  ,yvar=risk_score
  ,xvar=visit
  ,outlib=work
  ,yvarord=%str(0, 1, 2, 3, 98)
  ,xvarord=%str(1, 2, 3, 4, 5)
  );

%sankey_modify
  (sankeylib=work
  ,colorlist=LIGB LIYPK LIBG MOO VPAV GWH BWH LIOLBR
  ,barwidth=0.45
  ,yfmt=risk_score_fmt.
  ,xfmt=visit_fmt.
  ,legendtitle=%str(Risk Score)
  ,interpol=cosine
  ,datalabel=yes
  ,labeltype=percent
  );

```

Example 2 (Option 2). Sankey diagram for longitudinal risk score generated by adding a category for missed visit



In this diagram, percentages reflect **the distribution of patients who remained in the study, regardless of missed visits**. Gaps between the left edges of the bands represent patients who left the study. The diagram differentiates missed visit and end of follow-up. It also tracks the path of missed visits, which could be useful for missing data investigation.

Option 3: Modify the data and include missed visit and no subsequent visit as categories of risk score

	id	visit	risk_score		id	visit	risk_score
1	1	1	0	1	1	1	0
2	1	2	1	2	1	2	1
3	1	3	1	3	1	3	1
4	1	4	0	4	1	4	0
5	1	5	2	5	1	5	2
6	2	1	0	6	2	1	0
7	2	2	2	7	2	2	2
8	2	3	1	8	2	3	1
9	2	4	2	9	2	4	2
10	2	5	2	10	2	5	2
11	3	1	0	11	3	1	0
12	3	2	0	12	3	2	0
13	3	3	98	13	3	3	98
14	3	4	2	14	3	4	2
15	3	5	.	15	3	5	99

Macro calls for Example 2 Option 3:

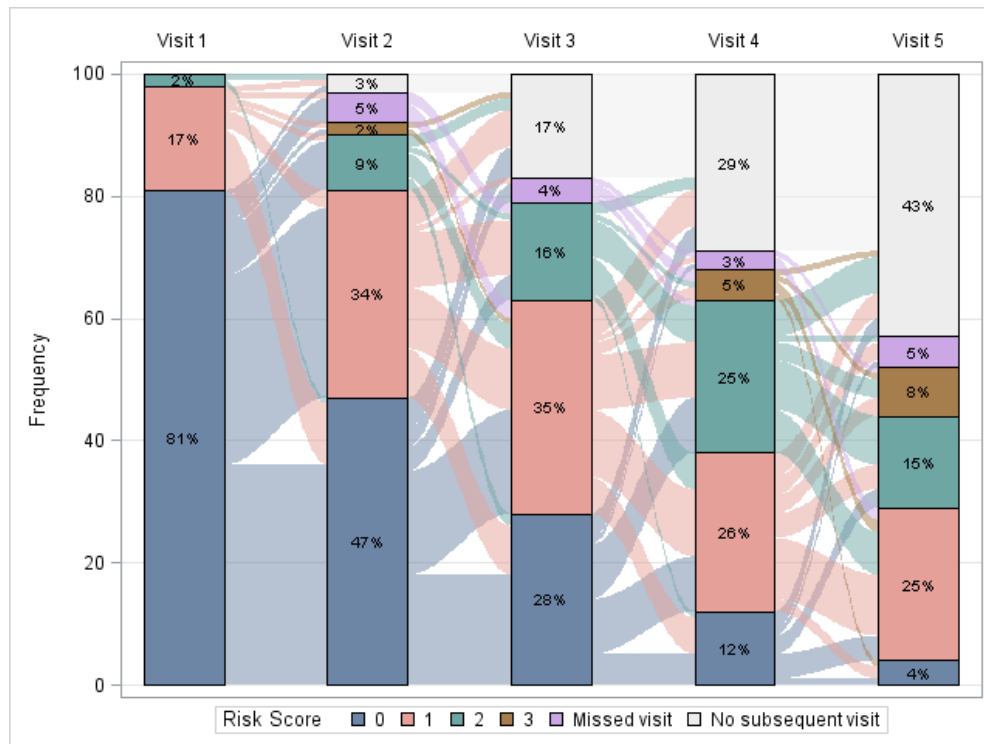
```

%rawtosankey_modify
  (data=example2_option3
  ,subject=id
  ,yvar=risk_score_missing
  ,xvar=visit
  ,outlib=work
  ,yvarord=%str(0, 1, 2, 3, 98, 99)
  ,xvarord=%str(1, 2, 3, 4, 5)
  );

%sankey_modify
  (sankeylib=work
  ,colorlist=LIGB LIYPK LIBG MOO VPAV GWH BWH LIOLBR
  ,barwidth=0.45
  ,yfmt=risk_score_fmt.
  ,xfmt=visit_fmt.
  ,legendtitle=%str(Risk Score)
  ,interpol=cosine
  ,datalabel=yes
  ,labeltype=percent
  );

```

Example 2 (Option 3). Sankey diagram for longitudinal risk score generated by adding categories for missed visit and end of follow-up



In this diagram, percentages reflect **the distribution of all patients who entered the study**. It can be useful for studying cohort distribution. This diagram can also be generated using the previously published macro.

CONCLUSION

Incomplete data are common in medical research. Although this modified macro allows patients with incomplete data to appear on the Sankey diagram, it is important for the user to understand the data and make sure that the missing and/or termination are intended. In addition, user should be aware that including incomplete data in a Sankey diagram may create different interpretations for the percentages on the bar chart. As limited data checks are available in this macro, user must check the input data carefully before creating the diagram.

REFERENCES

- [1] H. Sankey, "Introductory note on the thermal efficiency of steam-engines," *Minutes Proc. Inst. Civ. ...*, 1898.
- [2] M. G. Shin, M. S. McLean, and M. J. Hu, "Visualizing Temporal Patterns by Clustering Patients," *gotz.web.unc.edu*.
- [3] E. Hinz, D. Borland, H. Shah, V. West, and W. Hammond, "Temporal Visualization of Diabetes Mellitus via Hemoglobin A1c Levels," *researchgate.net*.
- [4] C.-W. Huang, R. Lu, U. Iqbal, S.-H. Lin, P. A. A. Nguyen, H.-C. Yang, C.-F. Wang, J. Li, K.-L. Ma, Y.-C. J. Li, and W.-S. Jian, "A richly interactive exploratory data analysis and visualization tool using electronic medical records.," *BMC Med. Inform. Decis. Mak.*, vol. 15, no. 1, p. 92, Jan. 2015.
- [5] M. Jones, R. Hockey, G. D. Mishra, and A. Dobson, "Visualising and modelling changes in categorical variables in longitudinal studies.," *BMC Med. Res. Methodol.*, vol. 14, p. 32, Jan. 2014.
- [6] S. Rosanbalm, "Getting Sankey with Bar Charts," in *PharmaSUG*, 2015, p. Paper DV07.

ACKNOWLEDGEMENTS

I would like to thank my manager, Margaret Coughlin, for reviewing this paper and providing valuable comments and suggestions. My colleagues from the Center for Observational and Real-world Evidence Center at Merck introduced me to Sankey diagram and provided inspirations for this project.

CONTACT INFORMATION

For more information, contact the author at:

Name: Yichen Zhong

Affiliation: Statistical Programing, Biostatistics and Research Decision Sciences, Merck & Co., Inc.

Enterprise: Merck & Co., Inc.

Address: 351 North Sumneytown Pike, Mail Stop UG-1CDS204A, North Wales, PA 19454

Email: yichen.zhong@merck.com

Work Phone: 267-305-1282

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.