SESUG Paper RIV-66-2017

Visualization of Patient Electronic Records to Support Exploratory Analysis and Variable Derivation of Categorical Data

Steven Thomas, Costel Chirila, Mary Beth Ritchey; RTI Health Solutions

ABSTRACT

BACKGROUND

Electronic medical records (EMR) have become a standard data source for epidemiological, outcomes, and health services research. However, there are challenges caused by the size and complexity of EMR data. Data are collected continuously across multiple systems and are stored in a variety of structures. Systems and structures can include free text, long or wide forms, and complex temporal information. These complexities make EMR data similar to an evolving ecosystem rather than a static source found in most studies. In a natural ecosystem, data sources are assessed to ensure that information is consistent with expectations. Institutions should approach EMR data in a similar manner to provide insight and to build confidence among team members with diverse backgrounds. New tools and processes need to be developed that support assessment of analytic decisions and are available to all members of the team.

METHODS

This paper proposes visual tools to use in exploratory analyses before variable derivation. These tools are designed to promote discussion and build consensus between team members using EMR data. They allow examination of individual patient records and trends across time so common operational considerations (e.g., defining variables via multiple features, selection of time windows) are addressed using both the data and therapeutic expertise. This paper will present SAS® graphic language templates for patient profiles, cumulative heat maps, and Sankey diagrams with example discussions and decisions that each visual is designed to support.

RESULTS/CONCLUSIONS

Studies seeking to maximize use of EMR data involve multiple stakeholders that need to understand nuances in the data. Visualizations can facilitate team discussions and improve the process of feature extraction, variable construction, and project planning. These visuals can be particularly useful for sequential analysis, treatment patterns, and defining episodes of care, but all studies using EMR can benefit from the use of visuals.

INTRODUCTION

The use of electronic medical records (EMRs) and other medical data repositories has increased in past years with data sources such as the Clinical Practice Research Datalink (CPRD), Truven Marketscan, data from the Centers for Medicare and Medicaid (CMS), and the National Patient-Centered Clinical Research Network (PCORnet) becoming common names in industry and academic research. Each repository of data is a complex reflection of temporal information held in a variety of structures, which presents major challenges for clinical and analytic teams that process the information into a condensed analytic form.

The path from source records to analysis-ready variables depends on each team member's experience and background. It can take hours to reconcile clinical knowledge with data limitations. For a physician, this may be the path of care observed in their own practice. An epidemiologist might focus on population trends in health care, whereas a statistician may use standards from previous studies. Each approach can conflict with current data as EMR systems adapt to changes in the standard of care, or as data providers change the way they use systems and revise collection tools.

To address these issues and gain a better understanding of our data, team members need access to tools that promote an open discussion of data definitions and careful assessment of expectations. The process would involve review of individual records, group summary statistics, and careful consideration of temporal elements. The templates presented here are designed to build confidence and consensus among team members with differing perspectives and experience.

BACKGROUND KNOWLEDGE CHRONIC KIDNEY DISEASE

In this article, we will present two examples along with interpretations and possible decisions, which are based on simulated data. The objective in the first example is to select patients with chronic kidney disease (CKD) from the simulated dataset. The objective in the second example is to define cohorts of drug use based on review of prescription data.

CKD is a condition where kidney function decreases due to hypertension and complications related to diabetes. A clinical diagnosis is confirmed with blood tests that estimate a patient's glomerular filtration rate (eGFR) and a urine test for proteinuria. Defining this condition from a database can be difficult as CKD is a progression of conditions that classify patients into five disease stages. In addition, each condition may be captured under different codes or measured directly through vitals and lab tests.

Treatment of CKD will vary based on the stage of the disease. When identified early, treatments focus on careful management of hypertension and diabetes, which are the underlying causes of CKD. In end-stage CKD, treatment options consist of transplant and dialysis. Our examples focus on two unspecified drugs, 'Drug A' and 'Drug B,' and event records related to CKD.

EXAMPLE DATA STRUCTURE

Examples we present use two simulated datasets representative of processed data from an EMR system. Each dataset is structured in a long format with multiple events per patient. It is not the purpose of this paper to review processing of electronic records. Instead we assume the presented fields can be constructed from an EMR source with information on diagnosis codes, labs, vitals, hospitalizations and records of prescription use. Table **1** and Table **2** provide a snapshot of a single patient with all necessary fields.

The general event dataset contains multiple records per patient (using a unique subject identification for each patient, "Usubjid"). Each record is a derived event related to CKD with fields for event description, timing and health status. Possible events include records on eGFR status, codes for CKD, hypertension, kidney function, diabetes type I or type II, hospitalization, referrals to a nephrologist, systolic blood pressure, and prescription for treatment. The time of the event is in days relative to the first recorded diagnosis of CKD, the index event. The patient's health status at the time of the event is considered 'Pre-onset' or 'Symptomatic' when events are recorded before the index event. Event classification after the index event is based on the number of confirmatory conditions such as hypertension, kidney function,

abnormal eGFR and prescription use. So a patient classified as 'Post 3/4' would have a diagnosis of CKD record, and three or four distinct confirmatory conditions. The prescription event data are in a similar structure but include a start and end duration of use.

Usubjid	Events	Patient Heath Status	Days From Index Event
TRTA1173	Hospitalization	Pre	-227
TRTA1173	Code (DM type II)	Symptomatic	-18
TRTA1173	Code (DM type II)	Symptomatic	-18
TRTA1173	Code (DM type II)	Symptomatic	-14
TRTA1173	Code 1 (CKD DX code)	DX	0
TRTA1173	Code (Evidence of hypertension)	Post 1	22
TRTA1173	Code (Kidney related)	Post 2	30
TRTA1173	Code (Evidence of hypertension)	Post 2	33
TRTA1173	Vital (elevated SBP)	Post 2	112
TRTA1173	LAB (abnormal eGFR)	Post 3 / 4	136
TRTA1173	LAB (abnormal eGFR)	Post 3/4	136
TRTA1173	Vital (elevated SBP)	Post 3/ 4	173

Table 1**. General Event Data Print Out of Single Patient**

Usubjid	Cohorts	DRG_Start	DRG_End
TRTA1173	Fixed-Dose Combination	0	90
TRTA1173	Drug A	88	133
TRTA1173	Drug A	134	179
TRTA1173	Drug A	182	227
TRTA1173	Drug B	314	344

Table 2**. Prescription Event Data Print Out of Single Patient**

PATIENT PROFILE PLOT

A good practice of any exploratory analysis is to review the data at the unit of information to be analyzed within the study. However electronic records are complex, numerous, and rarely in an accessible structure for all team members. A graphical representation addresses such issues by presenting temporal relationships between records.

Our approach uses a DATAPANEL layout to display multiple patients and retain common features across their respective spaces on the graphic. Event records are shown on the Y axis with a SCATTERPLOT overlaid with a BLOCKPLOT for health states. The values of each health state are centered at the top of the plot. An asterisk is substituted if the band is too small to display a label. Our profile plots support several key decisions and are primarily intended for use by clinical team members who have limited access or experience with previously collected EMR data. These profile plots were made using the event data in Table **1** and the following template:

```
proc template;
 define statgraph mygraph;
 dynamic yvar "required" grptitle "optional" pnum "required";
 begingraph;
 entrytitle grptitle;

 discreteattrmap name='attrmap';
       value 'pre'          / fillattrs=(color=cxffaf0a);
```

```
        value 'symptomatic' / fillattrs=(color=cxffeb46);
        value 'dx'          / fillattrs=(color=green);
        value 'post 1'      / fillattrs=(color=cxffc31e);
        value 'post 2'      / fillattrs=(color=cxfffe1e);
        value 'post 3/4'    / fillattrs=(color=cxffd732);
 enddiscreteattrmap;
 discreteattrvar attrvar=attrblock var=pat_heath_status
attrmap='attrmap';

 layout datapanel classvars=(usubjid)/ columns=1 rows=3
panelnumber=pnum
      border=false headerlabeldisplay=value headerborder=false
      rowaxisopts=(display=(tickvalues line) griddisplay=on
      linearopts=(tickvaluelist=(1 2 3 4 5 6 7 8 9 10 11)))
      columnaxisopts=(display=(tickvalues line label)
      linearopts=(tickvaluelist=
                 (-180 -120 -90 -60 -30 0 30 60 90 120 180)));

 layout prototype /  cycleattrs=true walldisplay=standard ;
 referenceline x=-90 / lineattrs=(color=gray pattern=dash);
 referenceline x=-30 / lineattrs=(color=gray pattern=dash);
 referenceline x=30  / lineattrs=(color=gray pattern=dash);
 referenceline x=90  / lineattrs=(color=gray pattern=dash);

 scatterplot y=yvar x=pat_time/name='main' group=yvar
   groupdisplay=cluster datatransparency=.1 markerattrs=
                  (symbol=circlefilled size=10pt color=blue);

 blockplot x=pat_time block=attrblock / display=(fill values)
      valuevalign=top datatransparency=0.5 valuehalign=center
      filltype=multicolor;

  endlayout;
  endlayout;
  endgraph;
  end;
run;
```

Note the following features in Figure 1:

- Multiple patient events and health states are contained in separate blocks.
- A common color scheme and labels is retained across patients.
- Disease progression is clear with a common horizontal axis.
- These patients became symptomatic 30 to 90 days before their initial diagnosis of CKD.
- Early symptomatic events involve diagnosis of diabetes, but the data do not contain records of elevated eGFR or SBP values.
- Codes for kidney function and hypertension occur within 30 days of the index event.
- Abnormal eGFR values are not observed until well past 90 days.

These findings may differ from expectations but could potentially be explained by lab, billing, or coding practices. CKD patients typically present with more numerous symptomatic events. These patients indicate possible issues with how those records are captured. The patient records also highlight the need

to carefully consider any temporal conditions and number of confirmatory codes when constructing a population of CKD patients from electronic records.

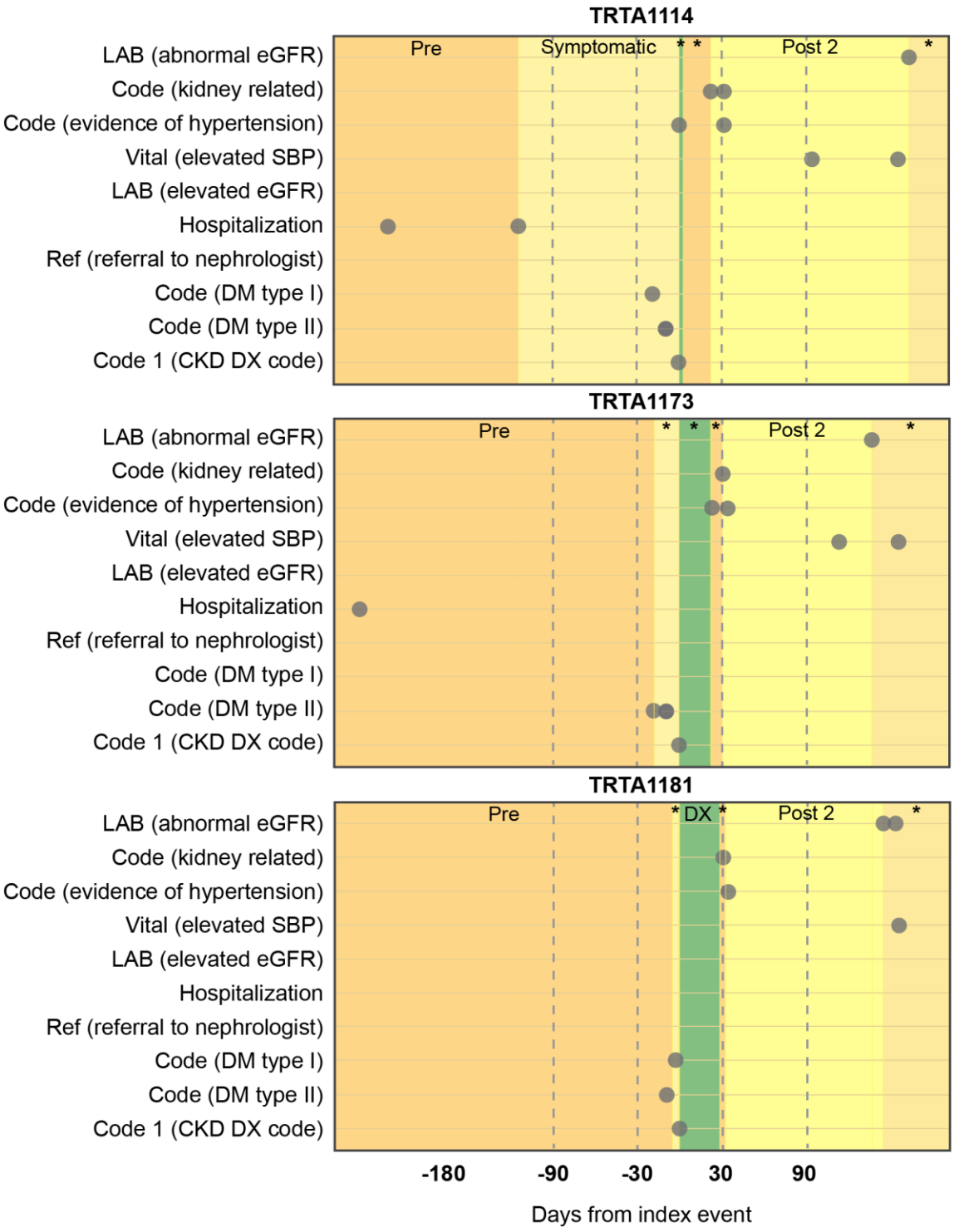

Figure 1. Patient Profile Plot of Chronic Kidney Disease Events

POPULATION PROFILE SUMMARY

Review of individual records can assist with team discussions and build familiarity with how the data are captured, but many key decisions are made at the population level. Our next figure tabulates cumulative sums and percentages at set intervals. Teams will find it useful in the selection of time windows and evaluation of algorithm development. Our examples use a single-day interval, but a larger interval can be constructed to reduce computing requirements. Again, records are benchmarked to an index event in both temporal directions, before and after the index event. All key events are displayed on the Y axis with temporal information on the X axis. A color gradient is added to present the cumulative percentage tabulated as we progress from the index event. Individual patient records are displayed with a SCATTERPLOT to indicate when events start or occur in low percentages. The example graphic and the corresponding cumulative values are constructed from the event data in Table **1** and the following template:

```
proc template;
 define statgraph profilesummary;
 dynamic yvar "required" grptitle "optional";
 begingraph ;
 entrytitle grptitle;

 rangeattrmap name="rmap";
 range 0 - 100 / rangecolormodel=(white orange blue );
 endrangeattrmap;
 rangeattrvar attrmap="rmap" var=cum_pct attrvar=pcolor;

 layout overlay / border=false
 yaxisopts=(display=(tickvalues)
 linearopts=(tickvaluelist=(2 3 4 5 6 7 8 9 10 11)))
 xaxisopts=(display=(tickvalues) griddisplay=off
 linearopts=(viewmin=-120 viewmax=120));

 heatmapparm x=pat_time y=yvar colorresponse=pcolor / display=(fill)
 discretey=true xvalues=leftpoints xendlabels=true name="heatmap";

 continuouslegend "heatmap";
 referenceline x=0 / lineattrs=(color=gray pattern=dash);
 scatterplot x=pat_time y=yvar / markercharacterposition=left
 markerattrs=(size=2pt symbol=circle color=black transparency=.1);

  endlayout;
 endgraph;
end;
run;
```

Note the following features in Figure 2 using the CKD event data:

- Percentages increase independently as events occur farther from the index record.
- Codes for diabetes are common with over 40% of the population having a coding event within 30 days before their CKD initial diagnosis, the index event.
- Evidence of hypertension and other kidney-related coding events are common 30 days postindex

event.

- Elevated eGFR and Systolic Blood Pressure rarely occur before or after the index event.
- Other confirmatory events, such as an abnormal eGFR Codes, rarely occur postindex.

These findings would strongly affect our opinion of the data quality and how we construct a population of CKD patients from these electronic records. Patients with CKD *should* experience much higher rates of abnormal or elevated eGFR values. This would lead the team to suspect quality issues with the lab data. While the coding events, kidney related and hypertension, seem to meet expectations.

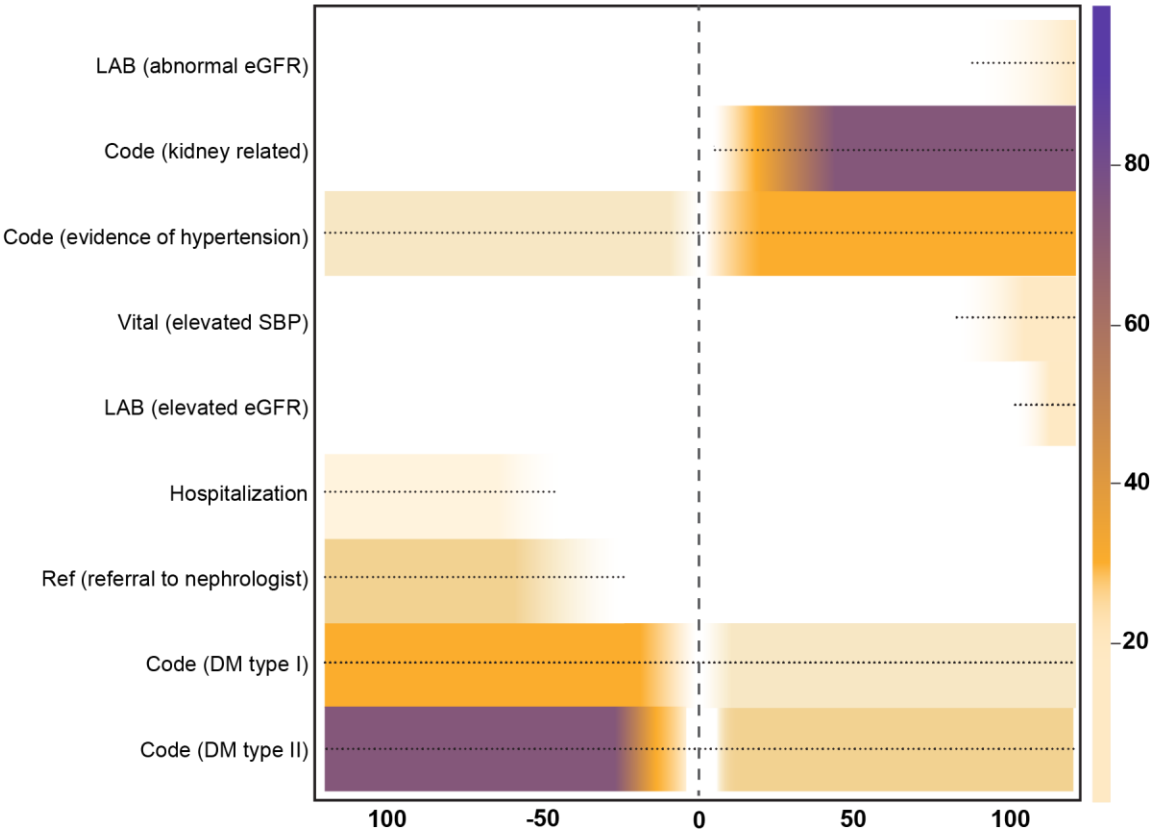

Figure 2. Cumulative Percentage of Patients with Conditions Relative to Initial Diagnosis of Chronic Kidney Disease

TREATMENT PROFILE PLOT

A data panel plot can be constructed for events with interval information as in Figure 1. In the example below, we present a graphic using the prescription event data from Table **2**. This graphic was designed to highlight possible combinations of therapy for two different drugs prescribed alone or in combination. Each duration of a prescription is presented with a high low plot by drug grouping. Overlapping intervals can be seen in darker areas and gaps in prescriptions by empty space.

```
proc template;
 define statgraph mygraph2;
 dynamic yvar "required" grptitle "optional" pnum "required";
 begingraph;
 entrytitle grptitle;
 layout datapanel classvars=(usubjid)/ columns=1 rows=5
panelnumber=pnum
```

```
        border=false headerlabeldisplay=value headerborder=false
        rowaxisopts=(display=(tickvalues line) griddisplay=on
        linearopts=(tickvaluelist=(1 2 3)))
        columnaxisopts=(display=(tickvalues line));

    layout prototype / cycleattrs=true walldisplay=standard ;
    highlowplot y=yvar high=drg_end low=drg_start/ group=yvar type=bar
    groupdisplay=cluster datatransparency=0.4 barwidth=1 name='main' ;

    endlayout;
    sidebar / align=bottom;
    discretelegend 'main'/across=3 border=false
           sortorder= ascendingformatted;

     endsidebar;
    endlayout;
   endgraph;
   end;
   run;
```

Note the following features in Figure 3:

- For these patients, the typical daily supply varies from 30 to 90 days.
- Switching between Drug A, Drug B or combinations appears to be common.
- The prescriptions can overlap or have gaps between supplies.

These findings would suggest teams need to carefully consider the effect of stockpiling medications, switching between medications and possible gaps in exposure. The template shares many of the features for a patient profile of key events. A HIGHLOWPLOT replaces the SCATTERPLOT and BLOCKPLOT statements. In addition, reference lines and other options are removed.

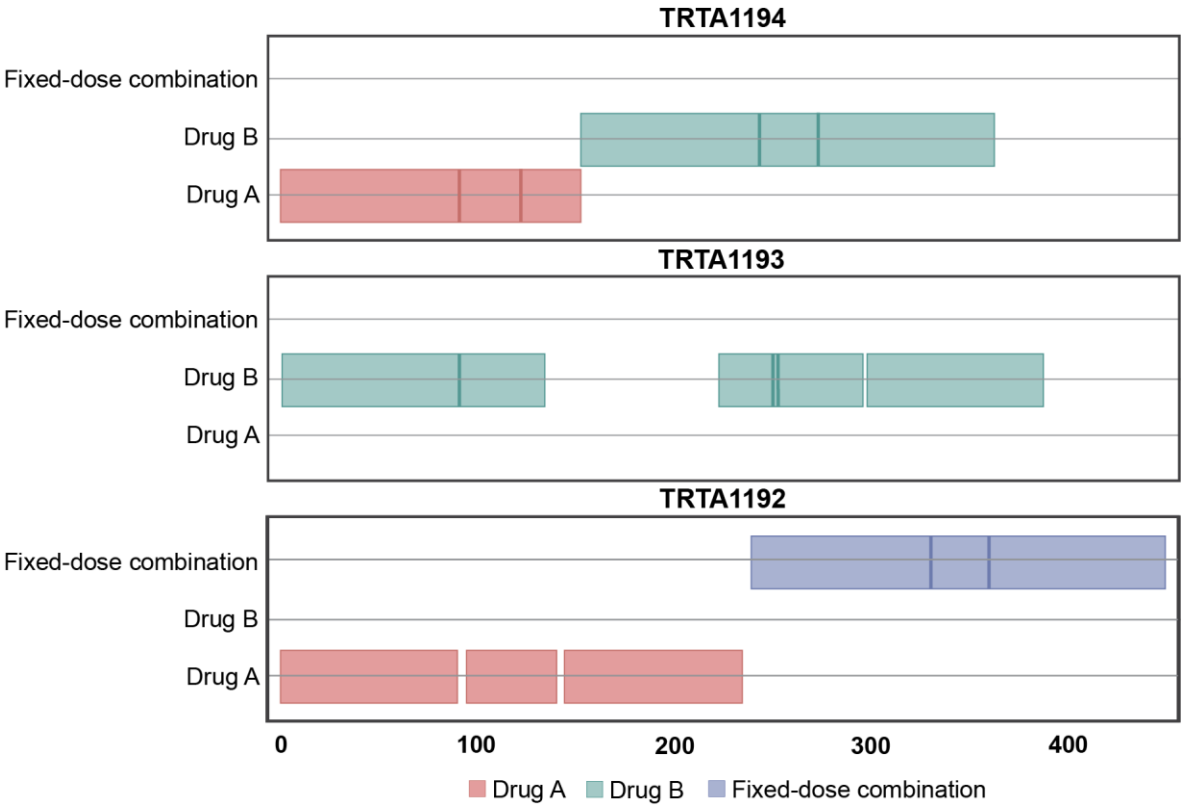

Figure 3. Patient Profile Plot of Prescription Records

DRUG USE SUMMARY – SANKEY DIAGRAM

A Sankey diagram is an ideal graphic for visualizing changes over time by group. In our example, we follow shifts in our study population from their initial prescription using 15 day windows. At each time window, we tabulated the percentage of patients in each treatment grouping, Drug A, Drug B, Fixed-Dose Combination or a Gap based on the prescription's day of supply. This status is then linked to the next time window and its grouping. To add to the visual appeal of the graphic, additional data points are constructed using a cosine function and simple do loop. Each set of data points for the link is shifted to start and end between the bars of each time window. This can take a bit of practice but is a relatively simple process once established.

```
proc template;
 define statgraph bar_line_2;
 dynamic grptitle "optional";
 begingraph ;
 entrytitle grptitle;

discreteattrmap name='attrmap';
  value 'Drug A'               / fillattrs=(color=cxd55e00);
  value 'Drug B'               / fillattrs=(color=cx009e73);
  value 'Fix Dose Combination' / fillattrs=(color=cxe69f00);
  value 'Gap'                  / fillattrs=(color=cx999999);
enddiscreteattrmap;
discreteattrvar attrvar=attrnode var=node_status attrmap='attrmap';

discreteattrmap name='attrmapb';
  value 'Drug A'               / fillattrs=(color=cxd55e00);
  value 'Drug B'               / fillattrs=(color=cx009e73);
  value 'Fix Dose Combination' / fillattrs=(color=cxe69f00);
  value 'Gap'                  / fillattrs=(color=cx999999);
enddiscreteattrmap;
discreteattrvar attrvar=attrband var=bgroup attrmap='attrmapb';

layout overlay / border=false xaxisopts=( label='days from initial rx'
      type=linear griddisplay=off display=(tickvalues label)
      linearopts=(viewmin=0 viewmax=60 tickvaluelist=(0 15 30 45 60)))
      x2axisopts=(type=linear griddisplay=off display=none
      linearopts=(viewmin=2.5 viewmax=57.5) )
      yaxisopts=(label='percentage population' griddisplay=off
      display=(label tickvalues) linearopts=(tickvaluesequence=(start=0
      end=100 increment=25)) );

bandplot x=window2 limitlower=ylow limitupper=yhigh /
      datatransparency=0.5 group=attrband xaxis=x;

barchart x=window y=percent /xaxis=x2 groupdisplay=stack group=attrnode
      barwidth=0.35  dataskin=pressed datatransparency=0 name='a';
scatterplot x=window y=node_y / markercharacter=textpct
      markercharacterposition=center markercharacterattrs=(size=9pt
      color=white) xaxis=x2 datatransparency=0 dataskin=crisp;

discretelegend 'a' /across=4 border=false sortorder=ascendingformatted
      valign=top;
 endlayout;
 endgraph;
 end;
```

```
run;
```

Note the following features in Figure **4**:

- The proportion of patients on Drug A, Drug B, or in a combination is approximately one-third at any given time.
- Switching between Drug A, Drug B, or combinations appears to be minimal, with about 5% of a given cohort changing therapy.
- After 45 days from the initial prescription, patients start to accumulate gaps in therapy.
- The links or lines that connect each time window follow an elegant cosine curve that is proportional to the percentage of patients transitioning to the next cohort.

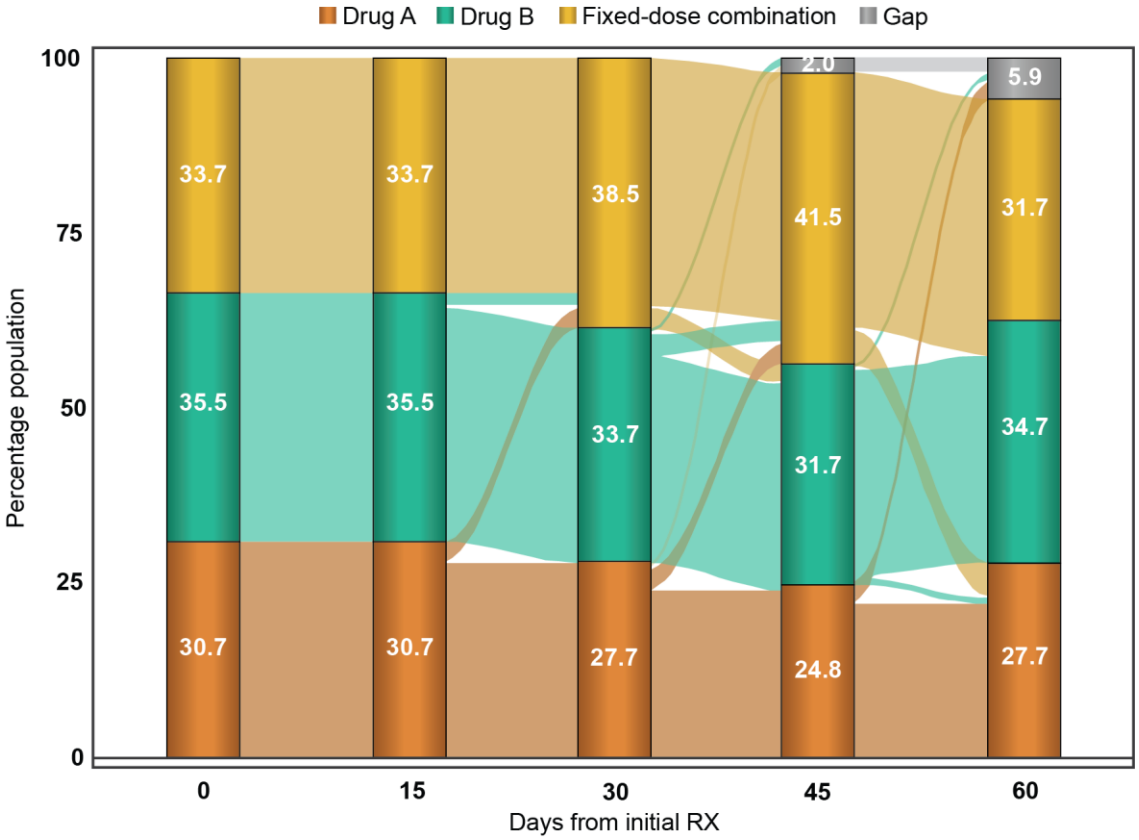

Figure 4. Patient Drug Flow Chart – Sankey Diagram

CONCLUSION

While large database studies have an enormous amount of information, it is only useful when all team members understand and trust their data. By developing and using visual tools, teams can improve communication and conduct operational tasks such as variable derivation and outcome validation in a timely manner. Our figures present records both at the individual patient level and aggregated across the patient population. So the nuance of individual patients can be reviewed but the effect of decisions can be evaluated across the entire population.

The graphic templates provided can be adapted to the user's needs but may require significant data manipulations or statistical calculations. The code presented is intended as a starting point for experienced analysts familiar with the SAS graphic template language.

CONTACT INFORMATION

Please direct any comments and questions to:

Steven Thomas, Statistician
RTI Health Solutions
Phone: +1 919.316.3133
E-mail: sthomas@RTI.org