A graphical approach to examine the completeness of epidemiological data: Patient Profile Plots (PPPs) and Centre Profile Plots (CPPs)

Methodios Typou (Independent Consultant), Joseph Kim (Amgen)

RATIONALE

Investigators performing epidemiologic studies (also referred to as observational or existing data studies) often wonder about the completeness of data available to them for statistical analysis. Unlike data collected for interventional studies, the completeness of data for observational studies can vary widely depending on the quality of the data capture process.

Here are the main characteristics of the epidemiological data that we had to work with:

- Large patient population (over 10,000)
- Data collected over an extended period of time (over 2 years)
- Data collection in natural settings and at unscheduled time-points
- Data with repeated measures over time
- Longitudinal data e.g. medication prescriptions (see data structures 2 and 3 below)
- Data captured in large number of variables (over 50)
- As a result of all of the above, large volume of data

Structurally our data falls into three main types:

1.

patientid	date	measurement_value
X000001	13-Jan-05	300
X000001	16-Feb-05	500
X000001	28-Jun-05	400

2.

patientid	prescription_start	prescription_end	dose
X000001	03-Jan-05	15-Jan-05	5000
X000001	16-Jan-05	04-Feb-05	6000
X000001	28-Feb-05	.	4500

3.

patientid	change_of_prescription_date	dose
X000001	05-Jan-05	400
X000001	12-Jun-05	350
X000001	02-Feb-06	500

In reviewing the PPPs and CPPs in the sections on the left, consider the simple scenario of investigating the relationship between use of medication B (**Med B**) and Laboratory measure A (**Lab A**), and the association of this relationship with events/outcomes.

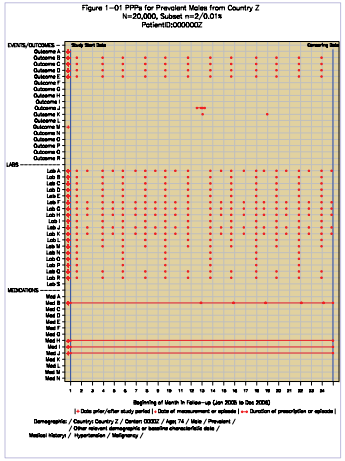
CONCLUSIONS

The Patient Profile Plots (PPPs) and Center Profile Plots (CPPs) are novel graphical tools that allow the investigator to evaluate the completeness of epidemiological data in a simple, easy to use graphical approach.

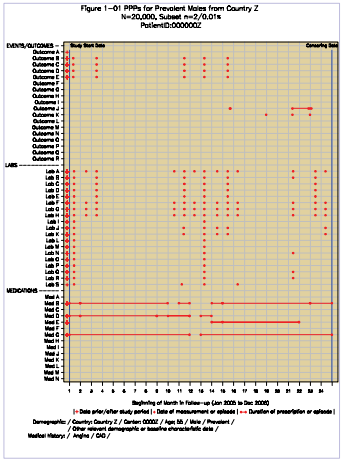
ACKNOWLEDGEMENTS

The work presented in this poster was developed as part of the ARO project at Amgen (Uxbridge, UK). The research initiative, Analyzing data, Recognizing excellence, Optimizing outcomes (ARO), is a multi-year project that seeks to improve the quality of medical care for patients with chronic kidney disease (CKD) and associated illnesses through data analyses and research. This poster presentation was sponsored by Amgen Limited.

PATIENT PROFILE PLOTS (PPPs)

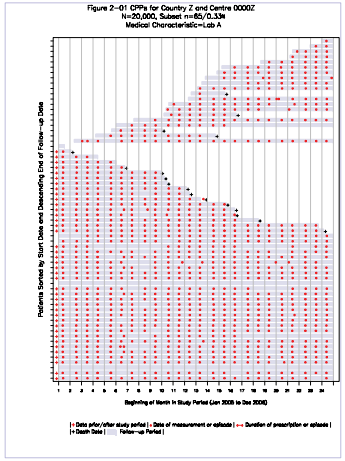


Example of a well populated data availability profile: Both Med B and Lab A data available almost throughout the follow-up period.

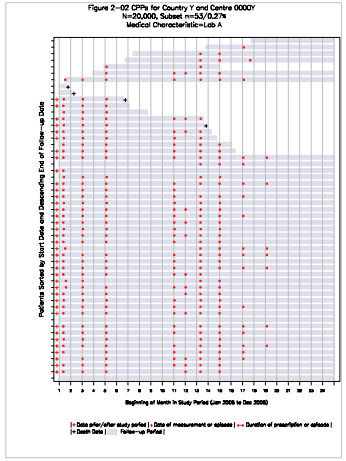


Example of a poorly populated data availability profile: Data completeness of Med B is very good, but Lab A data is not available during long periods.

CENTRE PROFILE PLOTS (CPPs)



Example of a well populated data availability profile: Excellent availability of Lab A data in the particular centre!



Limited completeness of Lab A data at this centre. A common seasonal pattern is observed in all patients. Operating problems in the centre? …or data extraction issues?

HOW TO DO IT

LABS

patientid	date	measurement_value
X000001	13-Jan-05	300
X000001	16-Feb-05	500
X000001	28-Jun-05	400
X000002	28-Feb-05	400
X000002	05-May-05	400

MEDS

patientid	prescription_start	prescription_end	dose
X000001	03-Jan-05	15-Jan-05	5000
X000001	16-Jan-05	04-Feb-05	6000
X000001	28-Sep-05		4500
X000002	26-Mar-05	04-Jul-05	6000
X000002	10-Jun-05	14-Jul-05	5000

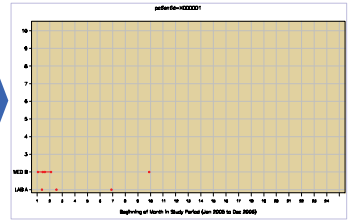
Data Transformation

ALL

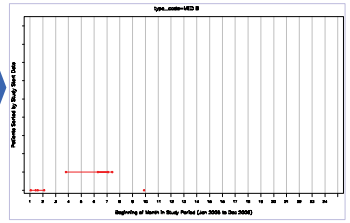
patientid	type	type_code	date
X000001	LAB A	1	13-Jan-05
X000001	LAB A	1	.
X000001	LAB A	1	16-Feb-05
X000001	LAB A	1	.
X000001	LAB A	1	28-Jun-05
X000001	LAB A	1	.
X000001	MED B	2	03-Jan-05
X000001	MED B	2	15-Jan-05
X000001	MED B	2	.
X000001	MED B	2	20-Jan-05
X000001	MED B	2	04-Feb-05
X000001	MED B	2	.
X000001	MED B	2	28-Sep-05
X000001	MED B	2	.
X000002	LAB A	1	28-Feb-05
X000002	LAB A	1	.
X000002	LAB A	1	05-May-05
X000002	LAB A	1	.
X000002	MED B	2	26-Mar-05
X000002	MED B	2	04-Jul-05
X000002	MED B	2	.
X000002	MED B	2	10-Jun-05
X000002	MED B	2	14-Jul-05
X000002	MED B	2	.

Data Plotting

```
proc format;
  value type 1='LAB A'
             2='MED B';
run;

options … ;
goptions …;
axis1 …; axis2 …;
symbol v=dot height=0.4 width=2 c=red interpol=join;

proc gplot data=all (keep=patientid type_code date);
  plot type_code*date / cframe=paoy frame skipmiss
        haxis=axis1 lhref=1 autohref cautohref=ligr
        vaxis=axis2 lvref=1 autovref cautovref=ligr ;
  by patientid;
  format type_code type.;
run;quit;
```



```
data all2;
  set all;
  retain patientid_code 0;
  by patientid;
  if first.patientid then patientid_code+1;
run;
proc sort data=all2;
  by type_code;
run;

axis2 …;

proc gplot data=all2 (keep=patientid_code type_code date);
  plot patientid_code*date / frame skipmiss
        haxis=axis1 lhref=1 autohref cautohref=ligr
        vaxis=axis2;
  by type_code;
  format type_code type.;
run;quit;
```



HOW IT WORKS

The plot statement plots each data point in the plot area. The INTERPOL=JOIN option of the symbol statement, connects data points with straight lines. If the data contain missing values, the observations are omitted, but the plot line is not broken. It is the use of the SKIPPMISS options that breaks the plot line, thereby creating the representation of the time period.

ADDITIONAL FEATURES

We presented how you can plot your data points using a simple and straightforward technique. The resulting output can be enhanced further to feature customised annotations, through the use of the annotate facility. One such example is the duration of each patient's follow-up period in the CPPs, which is represented by the light blue stripes.