PhUSE 2015

Paper CS05

Hands-on data mining, digging up clinicaltrials.gov data with SAS 9.

Mahdi About, Novartis, Basel, Switzerland

ABSTRACT
The amount of data in our industry seems to be growing exponentially, both in quantity and diversity. The question is – how can we harness this? To illustrate how we can make use of this data, we will present a real-world example of data-mining techniques using clinicaltrials.gov data SAS® 9 procedures, and Enterprise miner®.

KEYWORDS
ClinicalTrials.gov, XML to SAS, Multiple Correspondence Analysis MCA, PROC CORRESP, PROC CLUSTER, hierarchical classification, data mining.

DISCLAIMER
The opinions expressed in this paper (and in the related presentation and slides) are solely those of the presenter and not necessarily those of Novartis. Only public domain data has been used for analysis, Novartis does not guarantee the accuracy or reliability of the information provided herein.

INTRODUCTION - CLINICALTRIALS.GOV
In 1997, the Food and Drug Administration Modernization Act (FDAMA) was introduced to prepare the agency for working in the 21st century. The FDAMA produced various initiatives and led to the foundation of www.clinicaltrials.gov (CT.gov): NIH and the Food and Drug Administration worked together to develop the site, which was made available to the public in February 2000, and which provides patients with a way to search clinical trials information:

Figure 1 - Screenshot of clinicaltrials.gov

The patient can fill one or more fields, and the search engine will return the clinical trials matching the criterion (Figure 1). For this paper, we will search for all clinical trials in oncology from January 2000 to June 2015 executed by private sponsors. Results can be downloaded in a non-proprietary data format: XML (13269 XML files precisely as per Figure 2):

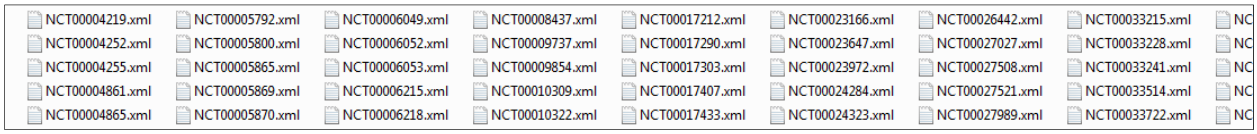

Figure 2 - Folder screenshot

XML AND SAS

SUMMARY

XML stands for Extensible Markup Language, and is designed to describe and store data in a software and hardware independent format. XML files are human and machine readable, self-described, hierarchical and internationalized. An example of XML is given in Figure 3 :

```
▼<clinical_study>
  ▼<!--
      This xml conforms to an XML Schema at:
         https://clinicaltrials.gov/ct2/html/images/info/public.xsd
      and an XML DTD at:
         https://clinicaltrials.gov/ct2/html/images/info/public.dtd
   -->
  ▼<required_header>
    ▼<download_date>
       ClinicalTrials.gov processed this data on August 30, 2015
     </download_date>
     <link_text>Link to the current ClinicalTrials.gov record.</link_text>
     <url>https://clinicaltrials.gov/show/NCT01816594</url>
   </required_header>
  ▼<id_info>
     <org_study_id>CBKM120F2203</org_study_id>
     <nct_id>NCT01816594</nct_id>
   </id_info>
  ▼<brief_title>
     NeoPHOEBE: Neoadjuvant Trastuzumab + BKM120 in Combination With Weekly Paclitaxel in HER2-positive
     Primary Breast Cancer
   </brief_title>
   <acronym>NeoPHOEBE</acronym>
  ▼<official_title>
     NeoPHOEBE: Pi3k Inhibition in Her2 OverExpressing Breast cancEr: A Phase II, Randomized, Parallel
     Cohort, Two Stage, Double-blind, Placebo-controlled Study of Neoadjuvant Trastuzumab Versus Trastuzumab
```

Figure 3 - Example of XML file from CT.gov

SAS/BASE® does not support XML parsing natively; instead, XML mapper® is available for free, which allows mapping of XML files from hierarchical data to horizontal dataset by using a MAP file along with the libname XML engine. I do not aim to create another tutorial to handle XML with XML Mapper/SAS®; the complete method has been covered already in many valuable papers[1].

INPUT DATA

The above query results in 13269 xml files representing the 13269 clinical trials in oncology from the specified period. After selection of interest variables, XML variables are mapped to SAS dataset columns, and one dataset is created for each clinical trial. Finally these 13269 datasets are appended to form one single source dataset. The whole process is covered by the macro import_xml.sas in appendix. It imports data from XML file to SAS datasets using the MAP file; XMLs are process sequentially from file 1 to file 13269).

ENCODING CONSIDERATIONS

XML files are internationalized files using the Unicode character set which covers every language. XML uses the UTF-8 character set, a web standard used by many other programming languages. Often, when a sponsor representative enters clinical trial information into FDA system, they use special characters which are not part of the SAS default WLATIN1 character set , this can lead to the below error message (Figure 4: 'some code points did not transcode') during XML data treatment by SAS:

```
NOTE: Copying SXLELIB.Trial_summary to WORK.TRIAL_SUMMARY (memtype=DATA).
ERROR: Some code points did not transcode.
       occurred at or near line 62, column 124
ERROR: XML parsing error. Please verify that the XML content is well-formed.
ERROR: File WORK.Trial_summary.DATA has not been saved because copy could not be completed.
NOTE: Statements not processed because of errors noted above.
```

Figure 4 - Prospective error message during XML import

An easy workaround is to start your SAS session using "Unicode support" additional language. You can also change the encoding of the session using system option at SAS startup.

OUTPUT DATASET

An example of the output dataset is given in Figure 5.

	nct_id	brief_title	official_title	agency	agency_class	source	authority
1	NCT00075192	CP-675,206 With Neoadjuvant Hormone Therapy in Patients With High Risk Prostate Cancer	A Phase 1, Open Label, Non-Randomized, Dose Escalation Study to Evaluate the Safety of CP-675,206 in Combination With Neoadjuvant Androgen Ablation and a Phase 2, Open Label, Randomized Study to Evaluate the Efficacy of CP-675,206 in Combination With Neoadjuvant Androgen Ablation and Androgen Ablation Alone in Patients With High Risk Prostate Cancer	AstraZeneca	Industry	AstraZeneca	United States: Food and Drug Administration
2	NCT00075218	A Study To Assess The Safety And Efficacy Of SU11248 In Patients With Gastrointestinal Stromal Tumor(GIST)	A Phase III, Randomized, Double-Blind, Placebo-Controlled Study Of SU011248 In The Treatment Of Patients With Imatinib Mesylate (Gleevec Tm, Glivec)-Resistant Or Intolerant Malignant Gastrointestinal Stromal Tumor	Pfizer	Industry	Pfizer	United States: Food and Drug Administration
3	NCT00075270	Paclitaxel With / Without GW572016 (Lapatinib) As First Line Therapy For	A Randomized, Multicenter, Double-Blind, Placebo-Controlled, 2-Arm, Phase III Study of Oral GW572016 in	GlaxoSmithKline	Industry	GlaxoSmithKline	United States: Food and Drug

Figure 5 - View of the result dataset

DATA EXPLORATION

Before starting complex analysis, one should of course, explore the data. There are many tools for this purpose. One could use PROC UNIVARIATE or more interactive software like JMP® or Enterprise miner®. The latter provides some nice features for exploring data in a more interactive manner, for instance the Multiplot node which produce automatically relevant graphs.

After initial investigation with SAS® the input dataset was seen to comprise **13269** clinical trials (which were recorded between January 2000 and June 2015) focusing on **689** distinct diseases. There are **3748** phase I, **5732** phase II and **1986** phase III trials. The rest are a mixture of phase 0 and combined phases. **11225** trials are on the action of chemical compound, **848** on biological products, **349** on procedures, **231** on radiation, **144** on devices, **44** on genetic and **417** on others (11 missing).

The **20** biggest private sponsors performed **30%** of the clinical trials. The company which performed the most trials is Novartis with **489** clinical trials during the last 15 years.

CORRESPONDENCE ANALYSIS: PROC CORRESP

SUMMARY

The CORRESP procedure performs simple and multiple correspondence analyses[2]. This procedure helps you to generate low-dimensional graphical representation of your data. This kind of analysis is very handy for extracting information (trends, similarities, association etc.) from highly dimensional datasets, i.e. observation with high number of variables. Here, we would like to identify similarities between the specialties of the big players in the pharmaceutical industry over the last 15 years (drug, biological products, genetic etc.).

GENERAL SYNTAX

```
PROC CORRESP < options > ;
TABLES < row-variables, > column-variables ;
VAR variables ;
BY variables ;
ID variable ;
SUPPLEMENTARY variables ;
WEIGHT variable ;
```

INPUT DATA

PROC CORRESP can read several type of input datasets: raw, binary, Burt and contingency tables.

Burt table: If in the TABLES statement, you do not specify a row variable and specify MCA option, then PROC CORRESP generates a BURT table, which is a cross tabulation of each variable with itself and every other variable. A Burt table is symmetric (see Figure 6). If the binary matrix is Z, then the Burt table is Z'Z.

Burt Table

	Blond	Brown	White	Short	Tall	Female	Male	Old	Young
Blond	3	0	0	2	1	1	2	1	2
Brown	0	6	0	2	4	2	4	3	3
White	0	0	2	1	1	1	1	2	0
Short	2	2	1	5	0	2	3	4	1
Tall	1	4	1	0	6	2	4	2	4
Female	1	2	1	2	2	4	0	2	2

Figure 6 - Burt table

Binary table: If you specify BINARY option, a binary table is generated which consists of one row for each input data set observation and one column for each category of each TABLES statement variable (see Figure 7). You can generate a binary table using PROC TRANSREG with the DESIGN option.

Binary Table

	Blond	Brown	White	Short	Tall	Female	Male	Old	Young
1	0	0	1	1	0	0	1	1	0
2	0	1	0	0	1	1	0	0	1
3	0	1	0	1	0	0	1	1	0
4	0	0	1	0	1	1	0	1	0
5	0	1	0	1	0	1	0	1	0
6	1	0	0	0	1	0	1	0	1
7	0	1	0	0	1	0	1	0	1
8	1	0	0	1	0	0	1	1	0

Figure 7 - Binary table

Contingency table: The TABLES statement and the VAR/ID statements are mutually exclusive but lead to same analysis. The statement choice depends on the input dataset shape. We used the TABLES statement because it avoids a lot of data preparation by reading the raw dataset up front and is easier to use, although it is less efficient. You can generate a contingency table using the TABLES statement with a row variable and one or more column variables (see Figure 8):

Contingency Table

	Blond	Brown	White	Short	Tall	Female	Male	Old	Young	Sum
Colman	1	0	0	1	0	1	0	0	1	4
Delafave	0	1	0	0	1	0	1	1	0	4
Ernst	0	0	1	0	1	1	0	1	0	4
Jones	0	0	1	1	0	0	1	1	0	4
Kasavitz	0	1	0	1	0	0	1	1	0	4
Kasinski	1	0	0	1	0	0	1	1	0	4
Myers	0	1	0	0	1	0	1	0	1	4
Singer	0	1	0	0	1	0	1	0	1	4

Figure 8 - Contingency table

If you prefer more effective processing because of the size of your dataset, you can still use PROC TRANSREG or PROC FREQ to directly generate a binary or Burt table respectively.

OUTPUT DATASET

PROC CORRESP provides many results and diagnostics datasets, for instance frequencies, and row and column contributions to axes which are mandatory to give a meaning to the axes. The most important dataset for our purpose is the coordinate dataset, which allows graphical representation of our data by using dim1 and dim2 variables (Figure 9).

	TYPE	_NAME_	Quality	Mass	Inertia	Dim1	Dim2
17	OBS	GlaxoSmithKline	0.697790702	0.071230159	0.055050544	-1.42633222	-0.87517476
18	OBS	Hoffmann-La Roche	0.001243503	0.070833333	0.018248743	0.030892458	0.026630658
19	OBS	Janssen Research & Development, LLC	0.034326106	0.014285714	0.02893147	0.567712705	0.196678263
20	OBS	Johnson & Johnson Pharmaceutical Research & Development, L.L.C.	0.006677706	0.011309524	0.041628037	0.33825779	-0.11495001
21	OBS	MedImmune LLC	0.122827792	0.010119048	0.057267711	-1.88078555	-0.26875281
22	OBS	Merck KGaA	0.011272432	0.010912698	0.03085982	-0.37928287	-0.14721266
23	OBS	Merck Sharp & Dohme Corp.	0.048535837	0.042063492	0.050109939	-0.33690236	-0.43213265
24	OBS	Millennium Pharmaceuticals, Inc.	0.036557363	0.024801587	0.01687479	0.335030955	0.130050075
25	OBS	National Cancer Institute (NCI)	0.934172165	0.028769841	0.059022435	-1.30241283	2.873222528
26	OBS	Novartis	0.256301813	0.09702381	0.048816535	0.818294341	-0.00418043
27	OBS	Pfizer	0.020995509	0.055555556	0.020791732	0.201965899	0.003407621
28	OBS	Pharmacyclics	0.016165315	0.008730159	0.019337568	0.232591977	0.363088023
29	OBS	Sanofi	0.023309616	0.056547619	0.019490429	0.204239157	-0.00228387

Figure 9 - Coordinates dataset

GRAPHICAL REPRESENTATION

The graphical representation is the beauty of correspondence analyses. We can group the big players in the pharmaceutical industry by the treatment in which they tend to specialize (for Oncology), i.e. drug, biological product, device, genetic etc. The amount of work to achieve an in-depth analysis would be tremendous, but using this method, with little or no background in the details of the analysis methods, one can easily generate a broad profile, with each company being positioned on a 2 (or more) dimension graphic (Figure 10) where the axis meaning is determined by the column contributions (here the intervention type) to axis.

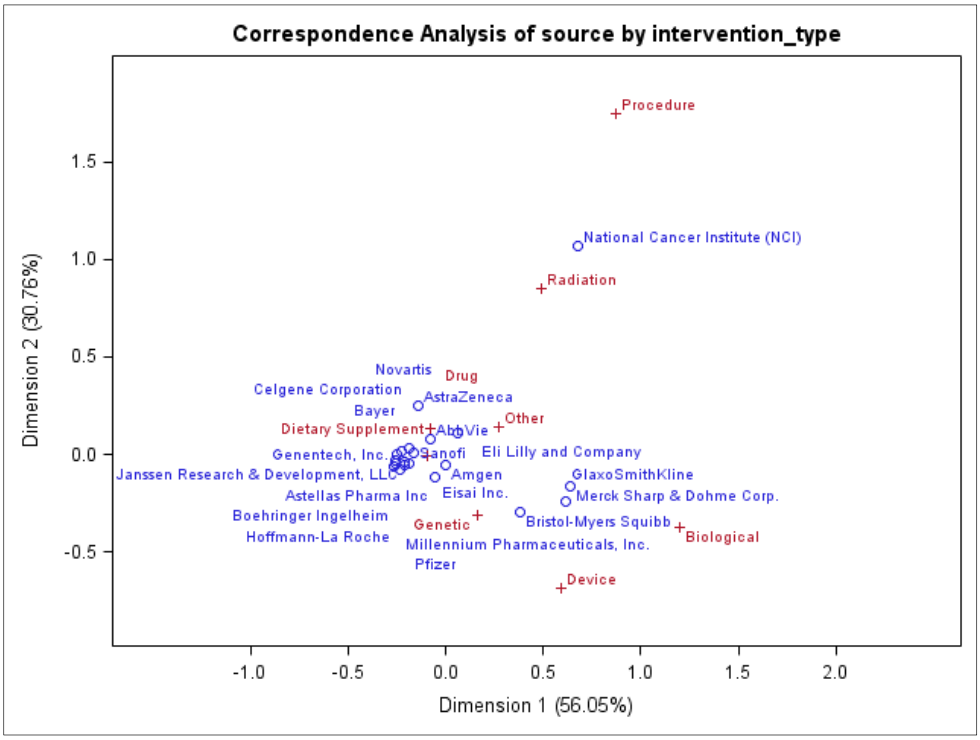

Figure 10 - MCA on company with > 70 trials since Jan 2000

HIERARCHICAL CLASSIFICATION: PROC CLUSTER

SUMMARY
We are now in possession of the coordinates of big pharma companies, and while we can appreciate the proximity of the points, we would like to use a quantitative method to group the companies into cluster based on their similarity (given by their relative position in an Euclidian space).
The CLUSTER procedure clusters the observations from a dataset by using a defined algorithm. The procedure computes Euclidian distances and processes the clustering using one of 11 methods. At the beginning of the iterative process, each observation is a 'cluster' itself and then the two closest clusters are merged to become a higher level cluster and so on. The process is repeated until we have one single cluster.
If you are analyzing huge datasets, you may have to use PROC FASTCLUS instead of PROC CLUSTER.

GENERAL SYNTAX

```
PROC CLUSTER METHOD = name <options> ;
BY variables ;
COPY variables ;
FREQ variable ;
ID variable ;
RMSSTD variable ;
VAR variables ;
```

COMMON METHODS[3]

To proceed with clustering, you have to specify a method: the most common are given in Figure 11, though SAS provides more methods (see this page for details).

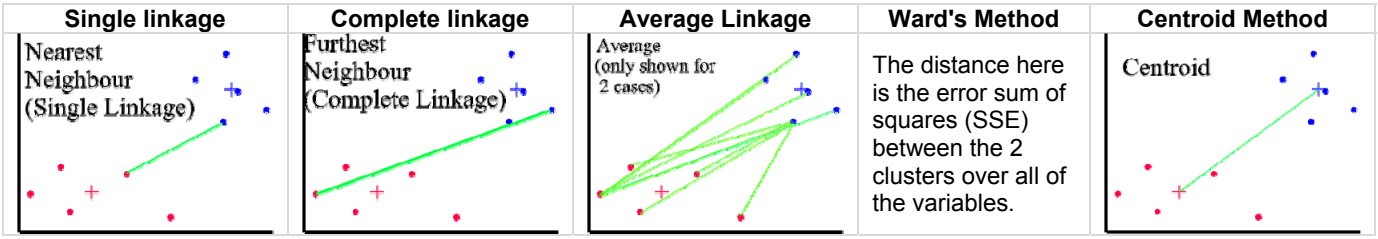

Figure 11 - Different clustering methods

INPUT DATA

Usually, the input data comprises quantitative variables, especially coordinates.

OUTPUT DATA

PROC CLUSTER produces a dataset which contains one observation for each observation from the input dataset, plus one observation for each node in the cluster tree. A parent variable identifies the parent node. The method used to cluster can be easily identified from the column labels (see Figure 12).

	Name of Observation or Cluster	Parent of Observor	Num of Clust	Freq of Clust	Distance Between Cluster Centroids	Root-Mean-Square Standard Deviation	Semi-Partial R-Squared	R-Squared
1	Daiichi Sankyo Inc.	CL10	11	1	0	0	0	1
2	Pharmacyclics	CL10	11	1	0	0	0	1
3	CL10	CL9	10	2	0	0	0	1
4	Seattle Genetics, Inc.	CL9	11	1	0	0	0	1
5	Abbott	CL8	11	1	0	0	0	1
6	Teva Pharmaceutical Industries	CL8	11	1	0	0	0	1
7	Gilead Sciences	CL7	11	1	0	0	0	1
8	Johnson & Johnson Pharmaceutical Research & Development, L.L.C.	CL7	11	1	0	0	0	1
9	CL8	CL6	8	2	0.0156605881	0.0156605881	0.0000245254	0.9999754746
10	CL7	CL6	7	2	0.0478342105	0.0478342105	0.0002288112	0.9997466634
11	CL6	CL5	6	4	0.069368966	0.0636591774	0.0009624107	0.9987842527
12	CL9	CL5	9	3	0	0	0	1
13	MedImmune LLC	CL4	11	1	0	0	0	1
14	Merck KGaA	CL4	11	1	0	0	0	1
15	CL5	CL3	5	7	0.1178711424	0.0998270955	0.0047635221	0.9940207306

Figure 12 - Cluster dataset

GRAPHICAL REPRESENTATION

From the output dataset, a tree can be generated either from PROC CLUSTER itself, or by using PROC TREE:

```
proc tree data=treeds nclusters=3 horizontal DISSIMILAR out=d1b;
id _name_ ;
run;
```

The procedure creates a tree diagram also known as Dendrogram (see Figure 13). The root of the tree is at the right. The x-axis distance is indicative of the distance between clusters. Below we can see the presence of 3 clusters, one comprising the National Cancer Institute, which specialized in surgery (procedures), a second cluster made up of Merck, GSK and BMS, which are known as biological products specialist, and one which is a giant cluster made of the chemical drug specialists. Of course, this doesn't mean that these companies are focusing only on chemical compounds, but this is highlighting the trend of these companies to develop more drugs than biological treatments over the last 15 years.

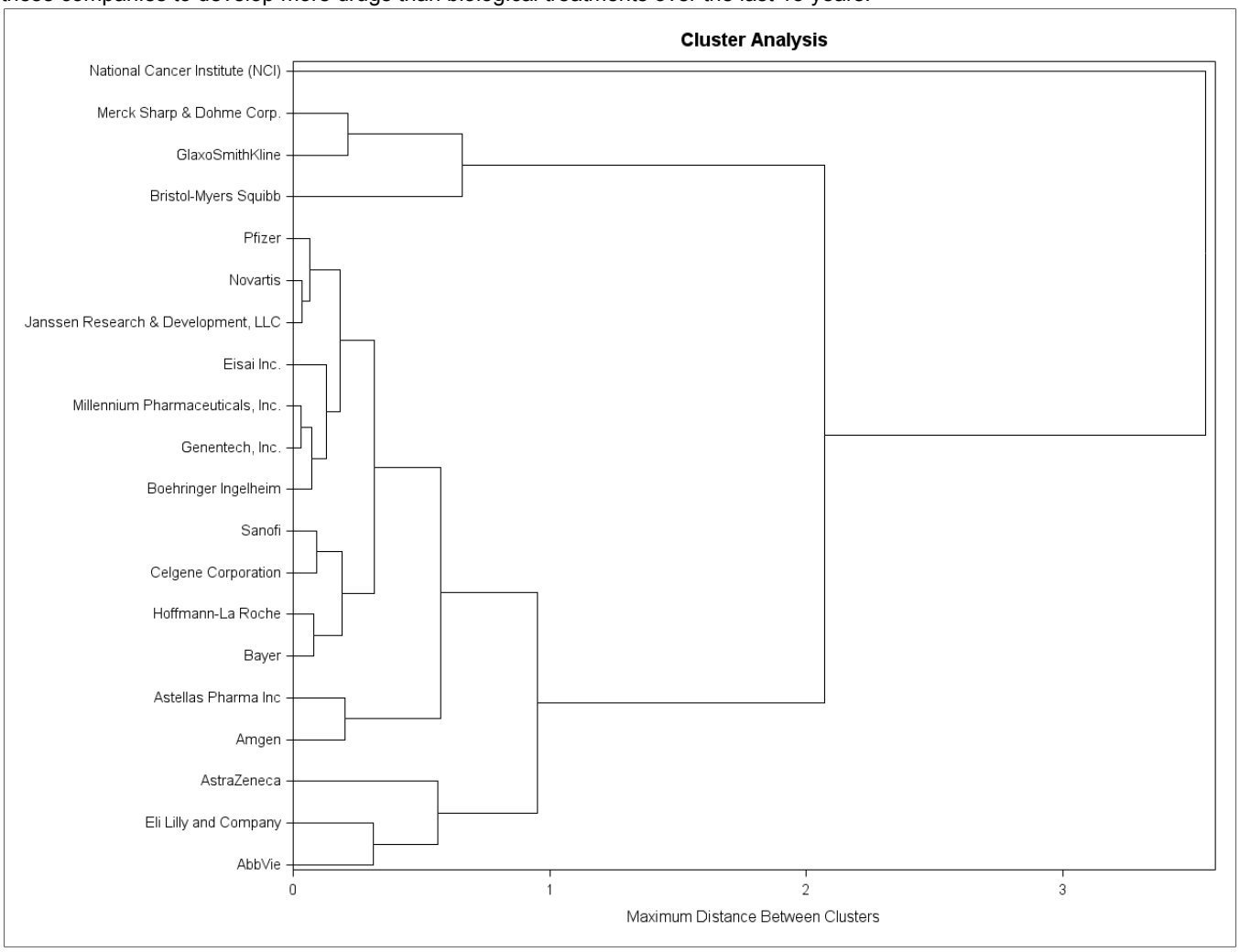

Figure 13 - Dendrogram

TEXT MINING

The above analyses can be further developed by text mining. This is a method which focuses on data summarization and the extraction of information from vast amounts of text. Text mining software solutions understand syntax and vocabulary and perform linguistic analyses. We used SAS® Text Miner, which is a component of SAS Enterprise Miner, meaning that it must be installed on the same machine[4].

ANALYSIS SEQUENCE

SAS ® Enterprise Miner™ is a data mining solution which uses a graphic user interface. The software runs on any platform that supports JAVA. The philosophy of the tool is drastically different from that of SAS base, but people used to SAS Enterprise Guide will not have too many problems.

The analysis process is based on diagrams (Figure 14), by which you design the analysis flow that will be performed: First we have to specify a data source; this can be a SAS dataset or raw text files.

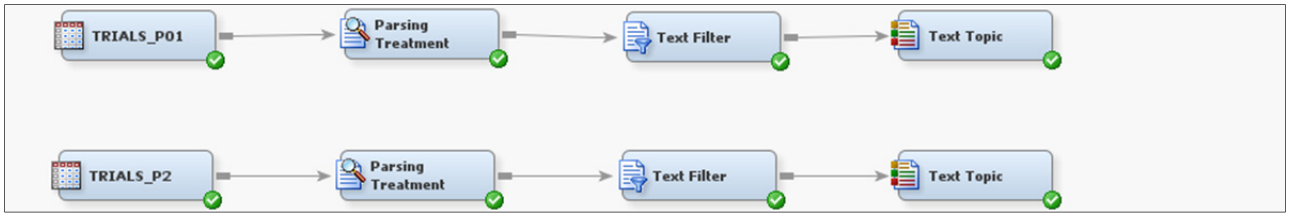

Figure 14 - Diagram Text mining in Enterprise Miner

Secondly, we parse the text using the parsing treatment node, during this step SAS Text Miner parses and analyzes the terms present in the data (Figure 15):

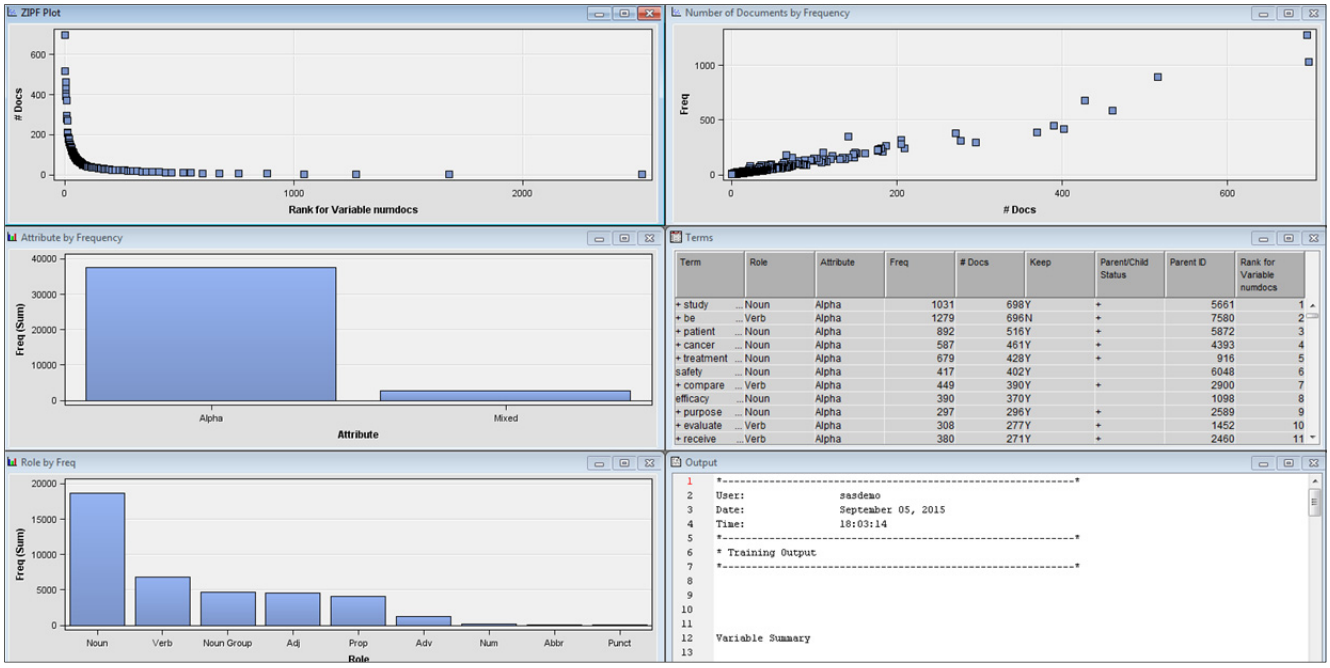

Figure 15 - Text parsing results

Terms are classified by their grammatical role, and regrouped under their stem form. The next step is the text filter, which allows you to drop or regroup terms. Finally, you can proceed with text profiling and topic exploration.

TEXT PROFILING

The text profile node associates descriptive terms with different levels of a target variable. We propose here to profile by the therapeutic areas on which companies focus, by analyzing the diseases mentioned in the clinical trials which they conducted.

Value	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8
	urinary/Adj	urinary bladder neo...	bladder/Prn	gastrointestinal stromal tu...	triple/Adj	negative/Adj	triple negative breast/Ng	stromal/Prn
AbbVie	leukemia/Nn	non-hodgkin/Prn	lymphoid/Prn	uterine/Adj	glioblastoma/Prn	stomach/Nn	lymphoma/Nn	cell/Prn
Amgen	neoplasm/Nn	metastasis/Prn	renal/Adj	colorectal neoplasm/Ng	colorectal/Adj	melanoma/Nn	sarcoma/Nn	stomach/Nn
Astellas Pharma Inc	prostatic/Nn	acute/Adj	leukemia/Nn	myeloid/Prn	pancreatic/Adj	pancreatic neo...	lung/Nn	ovarian neopl...
AstraZeneca	lung/Nn	breast/Nn	thyroid/Nn	ovarian neoplasm/Ng	ovarian/Adj	head/Nn	neck/Nn	colorectal neo...
Bayer	carcinoma/Nn	hepatocellular/Prn	prostatic/Nn	renal/Adj	renal cell/Ng	cell/Nn	non-hodgkin/Prn	colorectal neo...
Boehringer Ingelheim	carcinoma/Nn	non-small-cell/Prn	lung/Nn	head/Nn	neck/Nn	acute/Adj	myeloid/Prn	breast/Nn
Bristol-Myers Squibb	melanoma/Nn	leukemia/Nn	lung/Nn	myeloid/Prn	breast/Nn	cell/Prn	hepatocellular/Prn	plasma/Prn
Celgene Corporation	syndrome/Nn	cell/Prn	plasma/Prn	neoplasm/Nn	lymphoma/Nn	preleukemia/Prn	leukemia/Nn	lymphoid/Prn
Eisai Inc.	syndrome/Nn	sarcoma/Nn	breast/Nn	melanoma/Nn	thyroid/Nn	t-cell/Prn	hepatocellular/Prn	renal/Adj
Eli Lilly and Company	lung/Nn	breast/Nn	carcinoma/Nn	non-small-cell/Prn	metastasis/Prn	sarcoma/Nn	stomach/Nn	hepatocellular...
Genentech, Inc.	non-hodgkin/Prn	breast/Nn	colorectal neoplasm/...	colorectal/Adj	lung/Nn	lymphoid/Prn	lymphoma/Nn	cell/Nn
GlaxoSmithKline	breast/Nn	melanoma/Nn	non-hodgkin/Prn	lymphoma/Nn	cell/Nn	lymphoid/Prn	renal cell/Ng	leukemia/Nn
Hoffmann-La Roche	breast/Nn	melanoma/Nn	lung/Nn	lymphoid/Prn	colorectal neoplasm/Ng	colorectal/Adj	leukemia/Nn	glioblastoma/...
Janssen Research & D...	prostatic/Nn	lymphoma/Nn	plasma/Prn	cell/Prn	neoplasm/Nn	mantle-cell/Prn	metastasis/Prn	sarcoma/Nn
Merck Sharp & Dohme ...	melanoma/Nn	preleukemia/Prn	sarcoma/Nn	uterine/Adj	syndrome/Nn	breast/Nn	lung/Nn	leukemia-lym...
Millennium Pharmaceu...	neoplasm/Nn	lymphoma/Nn	cell/Prn	plasma/Prn	prostatic/Nn	follicular/Adj	mantle-cell/Prn	acute/Adj
Novartis	leukemia/Nn	myeloid/Prn	breast/Nn	tumors/Prn	neuroendocrine/Prn	glioblastoma/Prn	stromal/Prn	gastrointestin...

Figure 16 - Text profile terms (sponsor as target variable)

The program parses all the terms associated with a company name (Figure 16). For instance, we can see that Novartis is associated with the terms leukemia, myeloid, breast, tumors, neuroendocrine, glioblastoma, stromal and gastrointestinal. We believe this approach is less precise than a correspondence analysis in case you have very well structured data (disease code list properly defined). On the other hand, if you are working with unclean data, the text mining approach is much more powerful, because the software is handling synonyms, similar wording and even spelling mistakes, which gives to this solution a critical advantage when working with real world data.

TEXT TOPIC

The topic node allows you to extract the topic discussed from a collection of text. For instance, in our case, we were interested to see if the topic extracted by ENTERPRISE MINER from all the phase 1 contained in the database (Figure 17).

maximum,+tolerate,+mtd,+determine,+dose
+receive,+progression,+disease,+treatment,+anticipate
+drug,body,blood,+effect,+time
+solid,+advanced,+tumor,+tolerability,+malignancy
+paclitaxel,+carboplatin,+combination,+chemotherapy,+cisplatin
+escalation,+expansion,+cohort,+phase,+enrol
+objective,+primary,+profile,+secondary,pk
+relapse,+refractory,+multiple,myeloma,+lymphoma
+cell,lung,+cancer,+non-small,+growth
+therapy,standard,+exist,+progress,japanese
+subject,solid,+tumor,+clinical,+activity
+day,mg,+single,+oral,+effect
+dose-escalation,+multicenter,+phase,+open-label,iv
+day,+week,+infusion,+cycle,iv
+participant,+cancer,+metastatic,+advance,+breast

Figure 17 - Topic extraction results

The result seems to reflect clinical reality. The software presents 2 types of topics: single terms and multi terms, we will use here only multi terms. The topics covered by the phase I clinical trials are related “dose escalation”, “PK profile”, “maximum tolerated dose”. We have presented here the software’s simplest features. Nevertheless, the software provides a complete set of tweaks and tools to answer more complex analysis: such as detection of rare terms, weighting and user defined topics...

CONCLUSION

We have tried to present in this paper useful examples, stretching from data acquisition to final results, based on a real online repository. We believe these methods are relevant not only to pharma, but also to other sectors such as insurance, finance, mergers-and-acquisitions, research etc. The US government and FDA plan to make more and more databases available to everyone, opening a very exciting time for clinical R&D. Furthermore, private companies are also acting: biggest players of the sector are making clinical trial data available to researchers. Simultaneously, Eli Lilly, Pfizer and Novartis launched an initiative to improve the quality of the data provided through CTs.gov. They aim to ease the access of clinical trials to patients. Because

of all these coming challenges, we will need more innovative programmers and simpler tools to access this wealth of information.

ACKNOWLEDGMENTS

With thanks to Hong Yan, Andy Chopra, Danny Tuckwell, Nassim Sleiman and Insa Gathmann.
Particular thanks to Daniel Christen from SAS institute, for his support, and to Carmelo Lantosca for his time and explanations.

CONTACT INFORMATION

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

Mahdi ABOUT
Novartis Oncology
WSJ-103.3.10.17, Novartis Pharma AG
Postfach CH-4002 Basel, SWITZERLAND
mahdi.about@novartis.com
www.novartis.com

REFERENCES

1. http://www2.sas.com/proceedings/sugi29/119-29.pdf,
   https://support.sas.com/resources/papers/proceedings12/220-2012.pdf,
   http://www2.sas.com/proceedings/forum2008/099-2008.pdf.
2. http://en.wikipedia.org/wiki/Multiple_correspondence_analysis
3. http://hlab.stanford.edu/brian/forming_clusters.htm
4. http://support.sas.com/software/products/txtminer/
5. http://www.ncbi.nlm.nih.gov/books/NBK50886/

APPENDIX

IMPORT OF XML FILES

```
%macro input_xml() ;

*** Set options ;
option mprint mlogic bufsize=0 ;

*** CLEAN WORK lib ;
proc datasets lib=WORK kill memtype=data FORCE NOWARN NODETAILS NOLIST;
run;quit ;

*** Declare xml files, and folder to store imported data ;
filename indata pipe 'dir "C:\Users\...\PHUse\*.xml" ' ;

*** List and count XML files present into the folder, store number of files into fnum ;
data file_list ;
  retain cnt 0 ;
  length filen short_fname $50 ;
  infile indata truncover length=reclen;
  input ;
  if index(_infile_,'.xml')>0 then
       do ;
              filen = scan(_infile_,-1,' ') ;
```

```
            short_fname = tranwrd(filen,'.xml','') ;
            cnt + 1 ;
            output ;
      end;
  call symputx('fnum',cnt) ;
run;

*** sort the files by names ;
proc sort data = file_list ;
      by filen ;
run ;

*** CLEAR the pipe filename ;
filename indata clear ;

*** LOOP on the number of XML files into folder to transfer each into datasets ;
%do j=1 %to &fnum. ;

    *** Point on the n_th file and store its name into macrovariables ;
      data _null_ ;
            pointer = &j. ;
            set file_list point=pointer ;
            call symputx('fname',filen) ;
            call symputx('sfname',short_fname) ;
            stop ;
      run ;

      *** Declare XML file and MAP file associated ;
      filename  SXLELIB "C:\Users\...\PHUse\&fname." encoding="utf-8" ;
      filename  SXLEMAP "C:\Users\...\PHUse\auto_gen2.map" ;
      libname   SXLELIB xmlv2 xmlmap=SXLEMAP access=READONLY;

      *** Preprocess during the first iteration ;
      %if %sysevalf(&j.=1) %then %do ;

            *** Copy content from XML file to work lib and export name of datasets into
member_work ds ;
            proc datasets lib=SXLELIB NODETAILS NOLIST ;
                  contents directory data=_all_ out=work.member_work(keep=memname) noprint
;
                  copy in=SXLELIB out=work ;
            run;
            quit ;

            *** Get the name list of datasets and their total number ;
            proc sql noprint ;
                  select distinct strip(memname), count(distinct memname) into:dslist
separated by ' ',:numds TRIMMED from member_work ;
            quit ;

            *** Rename datasets to keep from one iteration to an other ;
            proc datasets lib=WORK NODETAILS NOLIST ;
            change
                  %do k=1 %to &numds. ;
                        %scan(&dslist.,&k.) = _%scan(&dslist.,&k.)
                  %end;
            ;
            run;
            quit ;

      %end ;
      %else %do ;
            *** Copy content from XML file to work lib ;
```

```
            proc datasets lib=SXLELIB NODETAILS NOLIST ;
                  copy in=SXLELIB out=work ;
            run;
            quit ;

            *** Apprend current datasets to Master dataset ;
            %do k=1 %to &numds. ;
                  proc append base=_%scan(&dslist.,&k.) data=%scan(&dslist.,&k.) ;
                  run ;
            %end;
            *** Delete datasets except master datasets ;
            proc datasets lib=WORK memtype=data NODETAILS NOLIST ;
            save
                  %do k=1 %to &numds. ;
                        _%scan(&dslist.,&k.)
                  %end;
                  file_list member_work
            ;
            run;
            quit ;

      %end ;
      %symdel fname sfname / NOWARN ;
%end;

** Store Final datasets into permanent lib ;
%do k=1 %to &numds. ;
      data perm.%scan(&dslist.,&k.) ; set work._%scan(&dslist.,&k.) ; run ;
%end;

*** Clear filenames ;
filename _all_ clear ; libname _all_ clear ;
%mend input_xml ;
%input_xml ;
```