**PhUSE 2007**

**Paper ST06**

# Data Mining in the Clinical Research Environment

Dave Smith, SAS, Marlow, UK

## ABSTRACT

Data mining has had wide adoption in recent years in many industries, largely because of the ability of mining techniques to rapidly yield answers to business questions in a short time and the availability of large quantities of data to exploit. This paper will discuss the topic of data and text mining in general, before focusing on applications in the clinical research field.

Of particular interest is the application of mining techniques to signal detection for adverse events. The value of these techniques is discussed, along with the context in which data and text mining appear in the overall architecture of a SAS solution for pharmacovigilance.

## WHAT IS DATA MINING?

Data mining is defined by SAS as the process of selecting, exploring, and modelling large amounts of data to uncover previously unknown patterns for business advantage. To expand on this in detail, it is important to realise that data mining is a continuous process where models are built, refined and managed over a period of time. The techniques used are largely iterative and empirical in nature, which implies a continuous process.

Several different techniques are employed to gain value from the data, including graphical exploration and many different modelling and modification techniques; data mining is not the same as data exploration.

Data volumes are generally very large, as data mining techniques are generally applied to circumstances where the problem is not well understood and traditional parametric statistics have either failed or not been applied because of the complexity of the situation. Data mining is also often applied where the problem statement cannot be easily stated, and where a hypothesis needs to be generated. For example the question could be “what significant associations exist between items in a typical shopping basket?” This might then lead to a question such as “do people that buy nappies also buy beer at the same time most of the time?” (This is apparently true!).

Data mining should always be done for business advantage, so being able to measure the outcome in business terms and then use that measure to compare models from the data mining process adds value and understanding.

### THE DATA MINING PROCESS – SEMMA

In order to improve the usability of the SAS Enterprise Miner™ tool and provide a framework to assist users in getting the most out of the tool, SAS has developed the SEMMA process:

- **Sample** the data by creating one or more data tables. The samples should be large enough to contain the significant information, yet small enough to process. You may need to apply stratified sampling techniques to obtain valid analysis of rare events, or not sample the data at all if there is insufficient volume to do so. Many data mining techniques (such as tree models or neural networks) employ learning algorithms and therefore require that the data is divided into two or ideally three parts to allow the algorithms to develop iteratively.
- **Explore** the data by searching for anticipated relationships, unanticipated trends, and anomalies in order to gain understanding and ideas. This is a very important stage in determining the success of the modelling stage; for example a graph of the data might indicate that it should be transformed, or that outliers should be removed. It is also likely to show variables that add no value and can be safely removed.

- **Modify** the data by creating, selecting, and transforming the variables to focus the model selection process. There are a number of techniques that apply to the removal/replacement of outliers that apply here.
- **Model** the data by using the analytical tools to search for a combination of the data that reliably predicts a desired outcome. These tools include clustering, self-organizing maps / Kohonen, variable selection, trees, linear and logistic regression, and neural networking.
- **Assess** the data by evaluating the usefulness and reliability of the findings from the data mining process. This is usually a matter of comparing the models in business terms (profit, lift) to determine which is best.

You may or may not include all of these steps in your analysis, and it may be necessary to repeat one or more of the steps several times before you are satisfied with the results. After you have completed the assess phase of the SEMMA process, you apply the scoring formula from one or more champion models to new data that may or may not contain the target. Scoring new data that is not available at the time of model training is the end result of most data mining problems.

In Enterprise Miner the SEMMA data mining process is driven by a process flow diagram, which you can modify and save. The GUI is designed in such a way that the business analyst who has little statistical expertise can navigate through the data mining methodology, while the quantitative expert can go "behind the scenes" to fine-tune and tweak the analytical process.

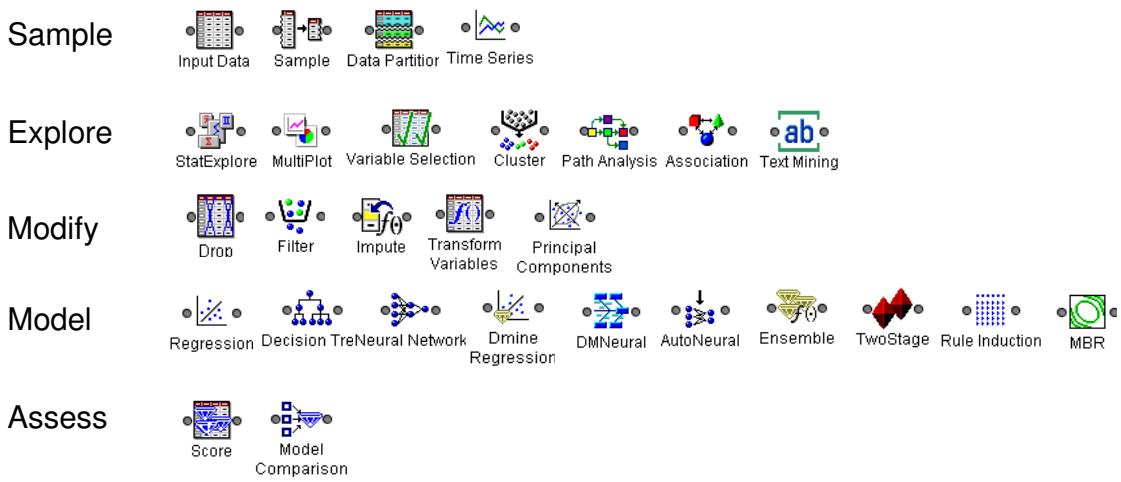


**Figure 1. The SEMMA framework**

SEMMA is not meant to be a complete data mining methodology but serves as a logical organization of Enterprise Miner tools for carrying out the core tasks of data mining. SEMMA is not a data mining methodology and should not be conveyed as such. SAS has developed its own methodology, the *SAS Data Mining Projects Methodology*, to address the comprehensive process of building models to address problems/opportunities including both precursor data mining activities (e.g. problem formulation and data mining case set preparation) as well as post-SEMMA tasks (e.g. model deployment and management).

**WHICH TECHNIQUES ARE RELEVANT?**

As can be seen from Figure 1, there are a large number of tools available within Enterprise Miner. Those that get used most are as follows

**StatExplore** – Generate descriptive statistics to understand correlations between variables etc.

**MultiPlot** – quick generation of plots of all the variables to understand distributions. The MultiPlot and StatExplore nodes together will drive many of the Modify steps, especially data transformation.

**Data Partition** – divide data into partitions for the training of learning algorithms such as neural networks. This is used on nearly all diagrams.

**Variable Selection** – remove unwanted variables or those that add nothing to the model

**Cluster** – group data into data driven clusters to generate hypotheses

**Text Mining** – group textual information into clusters to generate hypotheses

**Transform Variables** – modify data to deal with lack of normality, missing values etc.

**Regression** – perform logistic regression, usually well understood and easy to explain

**Decision Tree** – perform decision tree modelling; usually well understood and performs acceptably with non-normal data
**Neural Network** – perform Neural Network modelling; not easy to explain but very powerful
**Model Comparison** – compare models in business terms to select the best modelling technique and the best implementation of that technique

#### WHERE HAS THIS BEEN USED SUCCESSFULLY?

Within the life sciences sector one of the most successful uses of data mining was a US healthcare provider who generated predictive models for hospital admissions due to heart disease and asthma; the model was used to reduce hospitalisations by 80% by providing early preventive interventions

## WHAT IS TEXT MINING?

Text mining allows you to classify documents into predefined or data driven categories and find explicit relationships or associations between those documents.
Text mining is a multi-step process: accessing the unstructured text, parsing the text and turning it into actionable data and analyzing the newly created data. Within SAS® Text Miner the flow is typically:
**Text parsing** – automatically extract terms and phrases from parts of speech, as well as "stemming" to reduce words to their root forms (e.g. run, ran, running would all map to run).
**Automatic Text cleaning** – automated spell checking in the specified language
**Dimension Reduction** – using techniques such as Singular Value Decomposition to automatically relate similar terms and documents and avoid having to generate industry-specific ontologies (categories of words or phrases)
**Text Clustering** – Group documents into common themes and topics based in their content

The clusters generated are then used to either generate hypotheses or as additional inputs into another more traditional data mining model

#### WHERE HAS THIS BEEN USED SUCCESSFULLY?

An example of text mining in the life sciences is a company that uses text miner to categorise journal abstracts, making great efficiencies in their scientists researches by cutting down on number of abstracts they sift through before finding one that is of interest.

## DATA AND TEXT MINING TOGETHER

One of the main benefits of using Enterprise miner is that the tools for data and text mining are available on the same workbench, allowing the clusters that come from the text mining node to be easily combined with quantitative variables to produce a combined model that as been shown to provide greater insight than either technique alone

## APPLICATIONS OF DATA AND TEXT MINING IN PHARMACEUTICALS

Although in comparison with some industry sectors such as telecommunications and retail the data volumes in pharmaceutical industry are relatively small, there is still plenty of rich data from clinical trial history to exploit using data mining techniques. Just as recent applications in telecommunications and retail have focused upon understanding the dynamics of their business (answering questions such as which customers would respond best to special offers), so the opportunity exists for pharmaceutical companies to use data and text mining techniques to understand more about their own business.

One application might be to model the behaviours and attributes of investigators from previous trials and use this to predict which attributes suggest a suitable investigator for a particular trial domain, and then use this to drive recruitment policy.

Another related area might be to model the propensity of patients to withdraw from trials, adjusting for factors such as therapeutic area and drug class. This could be used to reduce recruitment to the minimum number that still maintains a very low risk of having to re-open recruitment towards the end of a trial.
Perhaps the most promising application of data mining within R&D is in pharmacovigilance, which will be discussed in the next section.

Outside R&D there are many potential applications of data mining, from the modelling of healthcare providers to understand prescribing behaviours and increase sales to the modelling of pharmaceutical manufacturing processes to predict batch failures and reduce costs.

## PHARMACOVIGILANCE

Recent high profile safety incidents have focused the minds of pharmaceutical companies, regulators and other agencies on pharmacovigilance, which is defined by the World Health Organisation as "the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problems".

Aside from the obvious public health issues, the potential costs to companies of drug safety issues are huge, particularly if there is a withdrawal from the market. For example, the withdrawal of Baycol was estimated to have cost Bayer in the region of $1Bn through refunds, lost operating earnings and out of court settlements. Some estimates have given Merck's potential costs from Vioxx lawsuits alone at over ten times that, and even when the evidence is based on a meta-analysis (and I leave others to debate the value of this technique) as with concerns over Avandia the impact on earnings can be enormous, with the drop in sales of Avandia wiping an immediate 9% off the GSK share price.

### SIGNAL DETECTION TOOLS IN PHARMACOVIGILANCE

It as become relatively common practice to screen pharmacovigilance databases for early signals of safety issues using screening techniques such as Proportional Reporting Ratios or Lincoln Technologies' MGPS (Multi-item Gamma Poisson Shrinker). There are many such techniques, and each has their own characteristics, but they all essentially do the same job, which is to sift through the many possible associations between drug and adverse effect and determine whether or not a "signal" is worth further investigation. It then takes human interaction to determine whether these signals are real, or whether they are due to one of a number of other factors. For example it could be that the association shows up between the compound and the indication it was prescribed for, or that a common concomitant medication with known adverse effects was prescribed alongside the compound. A physician would also be able to tell whether or not the association was clinically significant or already well known (such as NSAIDS and gastrointestinal effects).

### DATA MINING TOOLS IN PHARMACOVIGILANCE

Data mining techniques can add another perspective to the science of pharmacovigilance, allowing progressive investigation of the database to generate hypotheses outside the traditional methods, particularly using text mining, association analysis and clustering. Text mining techniques can also be used on other data sources, such as internet discussion forums where the luxury of accurate classification is not available. In this way data and text mining techniques can allow a move from reactive to proactive analysis.

### DIFFICULTIES IN SIGNAL DETECTION

Many conditions (e.g. Cancer, heart disease) are age related – risk increases with age. Certain common classes of preventive compounds may be prescribed following health screening, typically in patients' forties. The chance of heart disease and the chance of receiving certain classes of compounds are therefore likely to be related, and determining the difference between compounds that have causal links and those that have a beneficial effect requires high numbers of observations to allow age stratification, especially where the effects are small.

Pharmacovigilance often detects safety issues either early in a product lifecycle or once a critical mass of data has been accrued. The early detection is usually achieved through screening techniques (PRR, MGPS etc) and confirmed by thorough medical review of the cases to determine mechanisms and confirm causality. Later detection occurs when the compound has been in the market for a considerable period and where the volume of data permits the statistical separation of small differences.

Where text mining can add value is as a proxy for detailed medical review to boost the detection rates of standard techniques without reliance upon the larger volumes of data that come with a longer exposure to the wider population. In this way the detail from the spontaneous adverse event reports that might have been missed in simple coding can add to the body of evidence allowing an earlier indication that a safety issue is emerging. This is an artefact of the structure of coding systems and coding dictionaries, which might apportion related events to different disease hierarchies. An automated screening system based upon signal detection algorithms is likely to miss such a relationship where a physician would not, and it is this association that text mining should detect.

The timeline therefore could look like this:

**Pre-launch** – Safety issues detected by detailed controlled analysis of trial data
**Immediately post launch** – Safety issues detected by traditional monitoring methods (physician review of early spontaneous reporting post-launch)
**2-3 years post launch** – text mining and data mining together detects early signals that would be missed by signal detection screening algorithms. Too early to detect smaller effects using meta-analysis
**5+ years post launch** – meta-analysis or similar techniques indicate that a safety issue exists with a compound and it is withdrawn

Clearly if effects are detected earlier they could prevent public harm and cost to the pharmaceutical company.

#### PRACTICAL STEPS TO IMPLEMENT A PHARMACOVIGILANCE SOLUTION

One of the common issues with the delivery of any analytical application is the preparation of data, and it from experience most data mining related projects are around 80% data preparation and 20% data mining. The additional burden on data preparation in the pharmacovigilance arena is that the data must be coded to a common system, both for the adverse event classification and the medications used. Data warehouses built for clinical decision making are also subject to GCP and therefore need to be validated; this might suggest that a typical pharmacovigilance data mining project would be nearer to 5% data mining and 95% data preparation. This rather gloomy assessment is mitigated by the fact that many pharmaceutical companies have already constructed validated data stores to feed the data screening techniques such as PRR and MGPS, so much of this work has been done already. However it is likely that other data feeds will be necessary to produce the full value of data mining, and therefore the creation of analytical data stores needs to take into account which data are validated and which unvalidated; conclusions need to be tempered accordingly. It is therefore recommended that any selected solution have full management of all the relevant metadata.
Once data has been assembled for this purpose it is relatively straightforward to generate some of the screening measures as part of the data mart, although others will require specialist tools. These measures can be built into distributed reports, analyses and user alerts.
It is also strongly recommended that considerable effort is spent in improving and managing the data quality of the input data, both in ensuring that the data are valid and coded correctly, but also that there are no duplicates or poorly reconciled data between different sources.
A possible layout is shown in Figure 2.

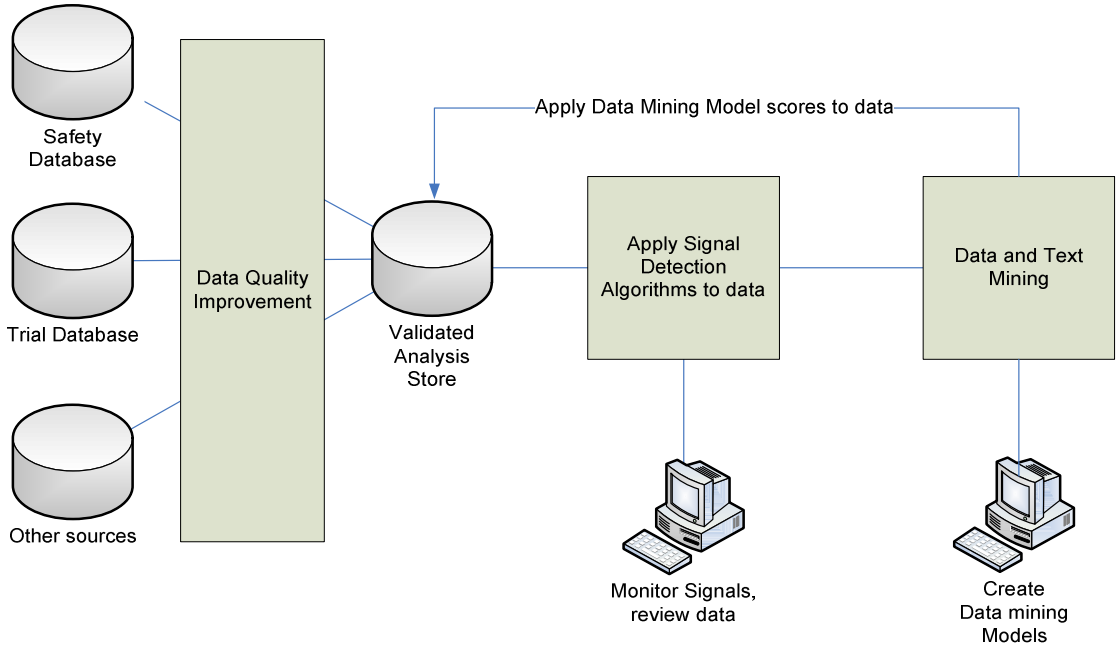


**Figure 2 – A Pharmacovigilance solution design**

## CONCLUSION

Data mining and text mining are powerful techniques that can add understanding to many aspects of the clinical research environment. These are perhaps strongest in the pharmacovigilance area, where the value of even small improvements in detection and management of adverse events has the potential to prevent disastrous consequences.

## REFERENCES

From Detection to Prediction - SAS® for Pharmacovigilance and Proactive Risk Management. A SAS White Paper.

Manfred Hauben and Andrew Bate - Data Mining in Drug Safety. Side effects of drugs essay, Side Effects of Drugs Annual, Volume 29, 2007, pages xxxiii – xlvi.

## CONTACT INFORMATION

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

Dave Smith
SAS Institute
Wittington House
Henley Road
Medmenham
Marlow
Bucks
SL7 2EB

Work Phone: 01628 404379
Fax: 01628 490550

Email: david.smith@suk.sas.com
Web: http://www.sas.com