

## Data Mining in a pharmaceutical environment

Franky De Cooman, Business & Decision, Brussels, Belgium

### ABSTRACT

Until now, most of the statistics are used in the R&D department. It is also observed that DM techniques are seldom used in a Pharmaceutical environment. This paper will show that these techniques can be easily and successfully used outside of their normal *habitat*

### INTRODUCTION

This paper will focus on the different departments where data mining can be applied and extensive examples will be given. On top of that, a case study will be given of a project done on Pharmaco-Vigilance data

### DIFFERENCE BETWEEN DM AND STATISTICS

Business question	Vs	Repository of Data
Formulate Hypothesis	Vs	Need to exploit data
Design data collection	Vs	Look in the data ( <i>dredge</i> )
Collect data	Vs	Formulate Hypothesis
Perform modeling & and check validity of model	Vs	Perform models
		Compare models

### DM & STATISTICS IN THE PHARMS AREA

Following are some examples of (possible) situations where Data Mining and Statistics can be used outside of the normal 'Clinical Trial' settings.

#### RESEARCH

##### **UNSTRUCTURED DISCOVERY OF NEW MOLECULES**

Imagine that all molecules being created at the department until now are grouped via clustering analysis into  $n$  groups via the characteristic chemical properties of the molecule. Once that these are grouped, one could find the elements most influencing the probability of belonging to group  $k$  and not to group  $l$ . If a new molecule is created one could then use this to see into which group out of the  $n$  the molecule might belong to. Like this, one can more easily find the clue of to which therapeutic group (i.e. what disease it might cure) the molecule belongs.

##### **STRUCTURED CREATION OF NEW MOLECULES**

If we can measure the chemical activity of a molecule on one way or another for a specific disease, we could try to find out from which part of the molecule it comes from. Then we could try to combine several molecules to construct a super-molecule, or trying to find a molecule with the good healing properties and getting rid of the adverse events provoking it.

#### DEVELOPMENT

The problem with analyzing the Adverse Events in the Pharmaco-Vigilance department is threefold:

The complaints reaching the department are most of the time spontaneous, not all problems get to the ears of the pharmaceutical companies.

We can only internally compare with other drugs from the same company, not with the drugs from the competing

companies.

We don't have the information of the patients taking the drug, and not having any Adverse Event.

Why not reuse the data from the clinical trials more often? For the purpose of the AE's, we could for example pool all the data, and analyze these, giving us more complete information!

#### **PHARMACO-VIGILANCE**

Unfortunately, drugs do not only cure diseases, one might also get Adverse Events. For the authorities, the mandatory PSUR exist (Periodic Safety Update Report), but these are only listings and tables, no inference is done. Using Data Mining techniques, we can find out which Adverse Events occur more frequently with specific drugs. If the MD prescribes a drug for a patient, he can then alert the patient for eventual side effects, or choose another drug if the side effects might be too dangerous for the specific patients.

We can even go a step further, once we know the Adverse Events related to a specific drug, we can try to find out if there are other factors influencing the probability of getting the Adverse Event. Think of sex, obesity, other diseases, and concomitant drugs. If we know that a certain drug is perfectly safe for 95% of the population, but is tremendously hazardous for 5%. Then it might be worthwhile trying to sufficiently describe these 5%, such that in the future, they don't get the drug, but the rest can still take advantage of the blockbuster.

#### **OUTCOMES RESEARCH**

Normally, for proving the efficiency of a drug, the rules for playing the game are described. Take for example a drug for hay fever, where one knows how to measure (relief) to compare the drugs. With Data Mining techniques, we could try to find alternative measures of relief, and promote the drug in another way: on top of curing the disease in a standard way, with our drug you get some extras compared to the competitor, e.g. a better Quality of Life.

#### **PHARMACO-ECONOMICS**

If we want to compare the cost of two drugs being equally efficient, until now, we only compare the exact cost of the drug on itself, isolated. But what if one of the two drugs gives more adverse events?? Trying to abate these also costs money (doctor visits, other drug use, hospital visit maybe), which makes the price ticket from the safer drug even more attractive! Data Mining can help comparing the total price ticket of a drug.

#### **MANUFACTURING**

If we can figure out in which time period/season a certain drug is sold (e.g. drugs for hay fever), we can adapt our production scheme, and make prediction on the number of boxes to produce. As with all manufacturing plants, storing products costs a lot of money!

One can also try to trace down production problems and see where they might come from.

Why not try to find the best suppliers for the chemical entities used by your plant, think of the new hype of Supplier Relationship Management (SRM).

#### **QUALITY CONTROL**

We can trace down the factors influencing the quality of the drugs being produced, identifying those factors introducing flaws in the production process.

#### **FINANCE**

One can try to find out which are the most costly phases in the clinical trials queue and try to predict like this what will be the costs of the clinical trials to come for a new molecular entity. Like that, management can simulate on their budgeting rounds.

#### **MARKETING**

Publicity becomes more and more important nowadays. In Europe, the rules are quite restrictive for the moment. But Europe will follow the attitude of the U.S.A., where direct marketing of drugs is not forbidden at all (and even comparative publicity on drugs is allowed). We can think of Customer Relationship Management of the patients, or look for the impact of television campaigns on the sales figures of a drug.

### **TYPICAL DM PHILOSOPHY**

#### **TYPICAL EXPERTISE INVOLVED IN DM PROJECT**

In a successful DM project, three competencies play a crucial role. These can be found in one and the same person, or a team can be built.

#### **BUSINESS EXPERT**

The person conducting the activity should have enough knowledge of the business he is working in. Like that, he can easily translate the business question in a modelling question

#### **DATA EXPERT**

As more than 60% of the effort of conducting a DM project is devoted to the set-up, validation and manipulation of the data, the person should be able to juggle with `RETAIN`, `FIRST..LAST.`, `PROC SQL`, ...

### **ANALYTICAL EXPERT**

As most of the problems can be downsized to classical statistical techniques, the DM expert should have a sound knowledge of statistics (and the corresponding SAS<sup>®</sup> procedures) and how to use them, not how to misuse them. At the same time, he/she be able to apply non-usual methods like decision trees, Neural Networks and he should be open towards new concepts like genetic algorithms.

### **VALIDATING/TESTING THE CONCLUSIONS**

When conducting a DM project, we should keep ourselves from drawing false conclusion. For this reason, we split our database in two *distinct* parts. The first part, the *model* part is used for constructing the model. This model will then be applied on the second part of the data, the *test* part. These data will be used to validate the correctness of the conclusions of the *model* part.

If for testing the same hypothesis, one likes to evaluate different techniques (e.g. logistic regression, decision trees), one could even split the database in three distinct parts, the latter being used to get an idea of the model giving the best prediction results.

### **CASE STUDY**

Here a case study will be given on data collected from the FDA website with Quarterly Adverse Events data.

#### **BUSINESS PROBLEM**

*Some Adverse Events are more frequent than others*

#### **SIGNAL GENERATION**

Which AE should we be aware of for each suspect drug?

Beware that a drug is called *suspect* as soon as one single AE has been reported for the specific drug.

#### **SIGNAL EVALUATION**

Does the AE only come from the fact one took the drug, or are 'circumstances' influencing this possibility?

#### **TECHNICAL SOLUTION**

##### **SIGNAL EVALUATION**

To calculate the *evaluation* of a specific drug, a frequency table is made with as factors

- The drugs
- The possible AE's (grouped by primary term)

These frequencies are then modelled using a log linear model on the *model* dataset. Doing so, a Confidence Interval can be constructed around the estimate. If the count coming from the *test* dataset (having approximately the same number of patients) for the same drug\*AE interaction falls into the confidence interval, we conclude that there is a relation between the specific drug and the AE. On top of this, as we have such confidence intervals for each quarter, we can make 'overall' confidence intervals for the model and test parts by using Bayesian techniques.

##### **SIGNAL EVALUATION**

With the list of possibly related AE's, the Medical Doctor responsible for the safety of the drug can choose which AE's he'd like to see further evaluated. For doing so, all the cases for the specific AE receive a flag indicating the suspect AE being present or not. These flags can than be modelled using Logistic regression or decision trees when trying to find explanatory variables like age, sex, weight, concomitant drug use...

#### **BUSINESS BENEFITS**

Thanks to the applied methodology, the workload for the MD could easily be diminished: important combinations from the frequency table of Drug/AE interactions can be highlighted, on top of that, indications can be found in which direction the MD should look when trying to explain the AE's.

### **CONCLUSION**

In this paper, it was shown that DM could easily applied in a pharmaceutical environment.

### **CONTACT INFORMATION**

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

Franky De Cooman  
Business & Decision

Omwentelingsstraat 8  
B-1000 Brussel  
Belgium  
Work Phone: +32/476 26 10 25  
Fax: +32/2 510 05 41  
Email: [fdecooman@businessdecision.com](mailto:fdecooman@businessdecision.com)