15th Triennial World Congress, Barcelona, Spain

DATA MINING IN DRUG AND THERAPY DESIGN

Aleš Belič * **Rihard Karba** * **Iztok Grabnar** ** **Aleš Mrhar** **

* *Tržaška 25, 1000 Ljubljana, Slovenia*
** *Aškerčeva 7, 1000 Ljubljana, Slovenia*

Abstract: Drug development and therapy design are typical areas where relatively large databases are often encountered. The processes of absorption, distribution, metabolism, and elimination are very complex and patterns are hard to find. Therefore, methodology, called data mining, is often needed to extract every possible information from the data. Detailed analysis of the databases can shorten drug development time, reduce costs, and provide more efficient dosage regimens. The aim of this article is to present and discuss problems of data analysis in pharmacokinetic studies which are illustrated with the example of a bioequivalence study database analysis. Suitable graphical representation of data is a simple tool in data mining, however, combined with modelling and simulation can become very effective. *Copyright ©2002 IFAC*

Keywords: Data processing, Databases, Modelling, Simulation, Identification, Pharmacokinetic data, Fuzzy modelling, Genetic algorithms.

1. INTRODUCTION

Drug development and therapy design are typical areas where relatively large databases are often encountered. In drug design, samples from thousands of people and animals can be collected in a single study. In therapy design numbers are lower but still can reach a few hundreds. The processes of absorption, distribution, metabolism, elimination, and their effects that are of main interest in drug and therapy design, are very complex and patterns in data are hard to find. At the same time the costs involved in pharmaceutical studies are very high, ethical as well as financial, and their tendency is rising. Therefore, methodology, called data mining (Gibas *et al.*, 2001), is often needed to extract every possible information from the data. Data mining is a common label for procedures of searching for complex patterns in large databases. Their goal is to transform the data into transparent information. The aim of this article is to present and discuss problems of data analysis in pharmacokinetic studies which are illustrated by dynamic analysis of pharmacokinetic data, collected in bioequivalence study. Bioequivalence studies are important part of generic drug design. Generic drug is a drug with already known substance and original formulation of carrier. Bioequivalence studies are designed to show equivalence in pharmacokinetic properties of original and generic drug. Pharmacokinetics researches processes of absorption, distribution, metabolism, and elimination of drugs and represents unavoidable phase in the process of drug development and therapy design. From measured concentrations of drugs in body fluids, pharmacokinetics extracts information on fate of the drug in species' body. Large "in vitro" and "in vivo" studies are the source of pharmacokinetic studies. Analysis with modelling and simulation offers possibility of substantial rationalisation of costs in problematic in general, expensive, and time consuming measurements as well as advantages in the processes of drug formulations development and dosing regimen design. Due to very rigorous requirements, pharmacokinetic models are becoming ever more complex, including structural complexity as well as time-variability and non-linearity. Artificial intelligence and expert knowledge inclusion are also becoming more and more important in pharmacokin-

etical modelling and simulation. Verified and validated models can be used, according to the modelling aim, for identification of mechanisms important to drug activities in organisms and for improvements of drug formulations, such as specific dissolution profiles, etc.. Due to the models' predictive power and ability to predict unmeasurable quantities, they enable simplified design of general and individual therapies.

2. DATA MINING IN PHARMACOKINETICS

The first step in analysis of pharmacokinetic properties of the drug in drug and therapy design is to represent the data graphically. Usually time vs. concentration plots are used. Next, a model, describing the dynamics of the drug in a body, is often composed. The model, from data mining point of view, is a transformation from measured data space into parameter space. This transformation reduces the dimension of space, making the database more transparent. If measured profiles of drug dynamics suggest relatively simple, mostly linear dynamics, and if modelling aim does not include study of mechanisms, so called population kinetics is near-optimal approach. Methods of population kinetics use parametric and structural identification of a model, providing values of parameters as well as their distributions as results (Jelliffe *et al.*, 2000). However, they are limited by complexity of model structure and non-linearities (Schumitzky, 1991) and are therefore mainly used for therapy design, where good predictions are necessary and the models must therefore be simpler as in drug design.

Complex transport systems of drugs, precise measurement methods, and complex interaction between drug and organism are implying the use of ever more complex models. The problem of large non-linear models is that the reduction of initial problem space dimension is not sufficient to produce transparent information on the system. Therefore, data mining methods must be used again, first to validate the model, and then to analyse the new database, consisting of model parameters and time courses of drug concentration in different measurable and unmeasurable regions of an organism. Model validation that is integral part of parametric identification methods, being one of their strongest points, must be performed separately from parameter estimation for large models. Since model parameters are mostly estimated by optimisation methods, there is no guarantee that optimal set of parameter values can be found nor that it exists, especially, since the dimension of parameter space is high ($>$15). The structure of the model must comply with approximate structure of organism. Then, patterns must be found.

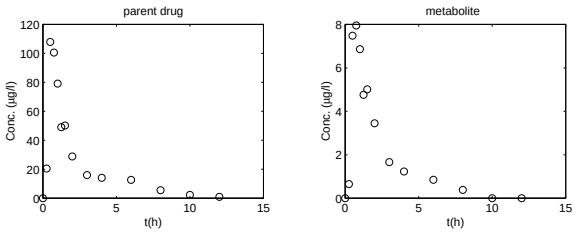

Fig. 1. Concentration vs. time plot.

2.1 Graphical representations in data mining

As mentioned above, database should be mapped into space where research of hypotheses, given by the aim of the work, is as simplified as possible. Different types of graphical representations offer the simplest possibilities for data mining in pharmacokinetic studies. As an example, the bioequivalence study will be presented, to illustrate necessity of data mining in large and complex databases.

3. MODELLING

The aim of the study was to characterise two similar drug formulations in fed and fasted conditions. Three studies were performed with slightly different tablet formulations. There were 144 measured profiles, available for analysis. As mentioned above, measured data is first represented graphically (Fig. 1) to get some information on system dynamics. Next, a model (Cellier, 1991; Matko *et al.*, 1992; Godfrey, 1983; Wagner, 1993) is composed from literature description of organism physiology and data analysis (Fig. 2). Model was fitted to the measured data and structurally changed to meet the specifications and remain as simple as possible. The absorption from gastro-intestinal tract was modelled with time-variable fuzzy sub-model since it is unpredictably variable, as substance is being transported by peristaltics. The presystemic metabolism of the drug in stomach was also modelled with time-variable fuzzy sub-model, since pH levels, governing the metabolism of the drug, can also erraticly change with time. To estimate the model parameters curve fitting procedure was used, where parallel genetic algorithm was varying the parameter values to simultaneously obtain best possible fit of model output and measured data of parent drug and metabolite plasma profiles. Model was able to mimic the real plasma profiles (Fig. 3).

3.1 Model Validation

Model was composed according to known physiology of an organism and was able to fit every measured profile equally well. Parameter estimation procedure was repeated 20 times for each individual with similar visual quality of fit (Fig. 3) and acceptable parameter value deviations (Fig. 4). No further attempts to validate the model were made at that time.

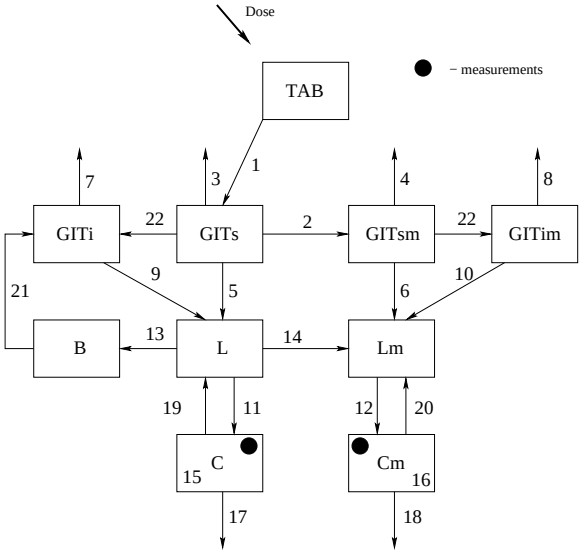

Fig. 2. Model of drug dynamics in human body. Numbers in figure denote consequent number of parameter, representing physical flow rates except for numbers 15 and 16, denoting apparent volumes of distribution for drug and metabolite, respectively, labels ending with m describe metabolite dynamics, the rest describe parent drug dynamics: TAB - tablet compartment, GITs - stomach comapartment, GITi - intestines compartment, B - bile bladder compartment, L - liver compartment, C - central compartment.

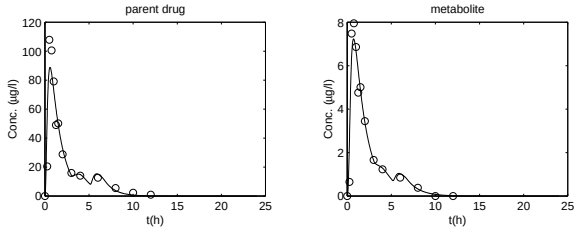

Fig. 3. Model responses (line) vs. measured data (circles) (concentration vs. time plot).

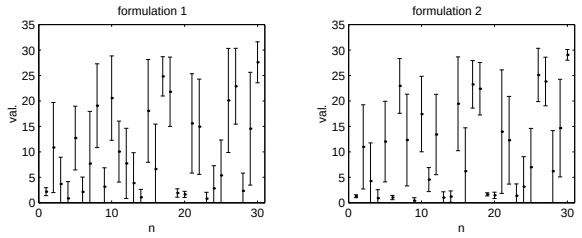

Fig. 4. Parameter values for one person after 20 repeated estimation procedures, on x-axis consequent number of parameter is represented and on y-axis the value of parameter with corresponding deviation is shown.

4. MODEL ANALYSIS

To obtain information from data that model simulations produced, several approaches are possible. When studying population data, some statistical evaluation of model parameters, quantities in compartments, and cumulative quantities in compartments is necessary to extract information that is relevant to the tested population. Means, medians, standard deviations, etc., however, do not always do the job, since the population may not be homogenous but consists of many subpopulations. Therefore, prior to statistical analysis of modelling and simulation results, some sort of clustering should be performed in order to detect subpopulations. In case of bioequivalence studies, detection of clusters in model data space is of main importance, since clusters are primary indications for bio-non-equivalence. Many clusters may indicate systemic differences between drug formulations, thus implying bio-non-equivalence of the two formulations. The simplest approach to clustering is to represent the data graphically and to find clusters visually. If modelling and simulation has reduced dimension of data space sufficiently, detecting the clusters visually is the most effective method.

4.1 Analysis of parameters

Analysis of model parameter values distributions provides information on importance of the parameter to the modelled dynamics and on systemic differences between subjects and their testing conditions (fasted – fed, formulation 1(R) – formulation 2(T), ...). High deviation of parameter results from low influence of the parameter on model dynamics, therefore, the amount of information that such parameter carries is small and vice versa. However, since dealing with large non-linear, time-varying models, they are very likely unidentifiable, therefore, parameter values are not always a reliable measure of comparison. Only when parameter values are grouped in distinct clusters, the parameter may be used for comparison purposes. Since two drug formulations are compared in the study, two dimensional plots are chosen. Each axis carries parameter values for one drug formulation, therefore, each person is represented with one data point. Parameter values, grouped near the main diagonal of the plot, suggest that drug formulations are equal regarding the observed parameter. In Fig. 5, parameter values for tablet dissolution are represented. There it can be seen that dissolution of the tablet is reduced in fed studies and that there are no significant differences between the two formulations, since data points, except for some outliers, are distributed symmetrically around main diagonal. The fuzzy submodel for drug pre-systemic metabolism shows certain differences between fed and fasted conditions, as well as between drug formulations (Fig. 6). In Fig. 6 it can be seen that in study 3, at around 1 h after administration, in fasted conditions, formulation 2 metabolises faster, however, in fed conditions formulation 1 metabolises faster.

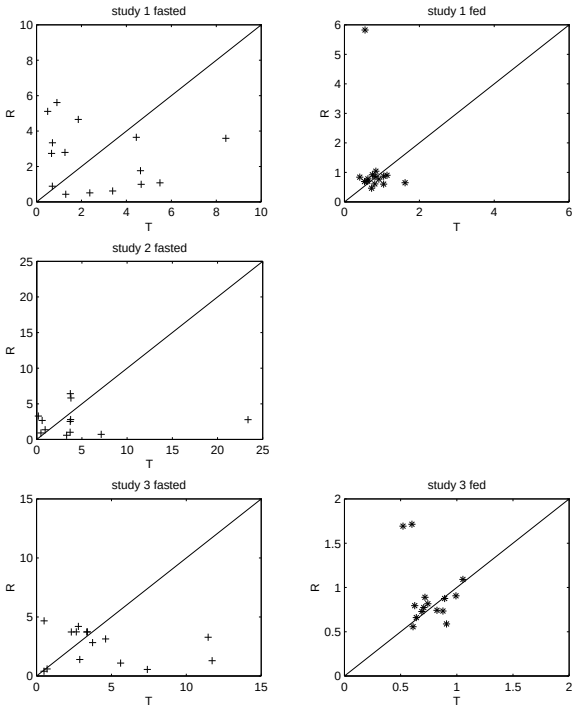

Fig. 5. Tablet dissolution in different studies, on x-axis values of parameters for formulation 2(T) and on y-axis values of parameters for formulation 1(R) are presented, crosses represent fasted study and asterisks represent fed study, each point represents one subject.

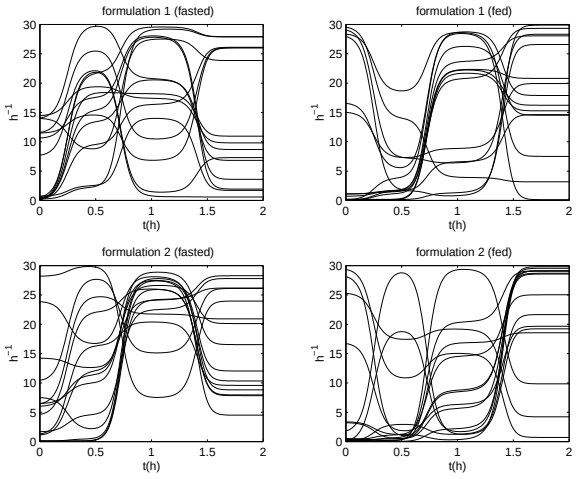

Fig. 6. Time course of fuzzy sub-model in study 3 (15 subjects) for velocity of metabolite generation in stomach.

4.2 Analysis of quantities

Quantities in compartments are more reliable information than parameter values in cases of unidentifiable models, since unidentifiability implies that similar time courses of quantities in compartments can be achieved by different parameter values. In Fig. 7 cumulative quantities of metabolite generated in stomach are shown. It can be seen that in spite of differences in fuzzy sub-model, cumulative profiles for fasted condition do not differ significantly. However, the

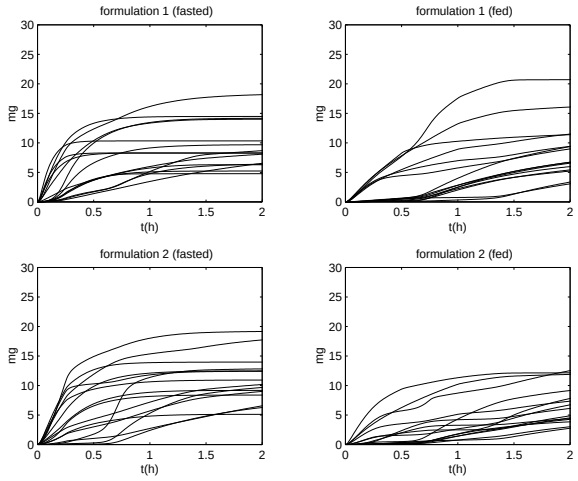

Fig. 7. Cumulative profiles of metabolite generated in stomach for study 3 (15 subjects).

difference is obvious in fed conditions, where higher levels of metabolite were produced for formulation 1 in study 3. Similar time courses of 15 subjects are not very transparent. Cumulative profiles, where final value is often the most important, can also be represented in similar way as parameters in Fig. 5. In Fig. 8 it can be seen that for study 1, there are two distinct clusters, indicating differences between fed and fasted conditions, however, the most important difference for bioequivalence is that the two clusters are not placed symmetrically around main diagonal, indicating differences between the two drug formulations. For study 2, no significant differences can be found. For study 3, one cluster, mostly symmetrically distributed around main diagonal can be found, however, bellow diagonal mostly data points for fed conditions are found and above diagonal mostly data points for fasted conditions are found. The structure of the cluster thus suggests difference regarding to fasted-fed conditions as well as differences between the two drug formulations, however, the differences are not as significant as for study 1.

5. CONCLUSION

Graphical representations of data is a simple tool for data mining in drug and therapy design. However, it may carry a lot of transparent information, when plots are carefully chosen. Compartment based modelling and simulation approach, when complex models are used, often still produces large quantities of new data and the information within is often not transparent. Therefore, a graphical representation can be helpful again. Parameter values are suitable to represent, however, unidentifiability of the model reduces the reliability of the information they carry. Time courses of quantities in compartments are often too complex to analyse, since different subjects can produce a lot of different time profiles. Final values of cumulative profiles carry a very compressed information on system dynamics that can be transparently graphically

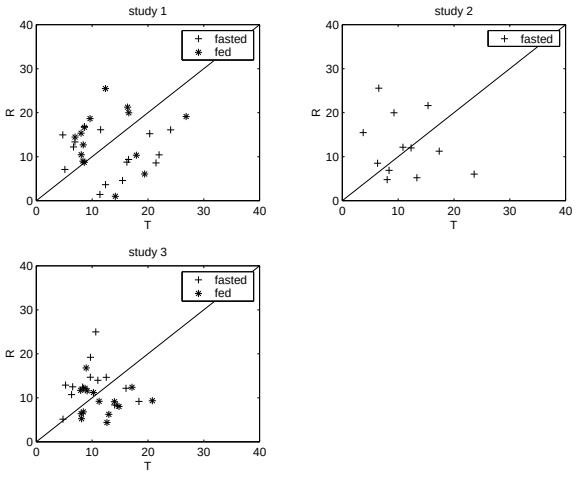

Fig. 8. Cumulative quantities of metabolite generated in stomach, on x-axis values for formulation 2(T) and on y-axis values for formulation 1(R) are presented, crosses represent fasted study and asterisks represent fed study.

represented. Clustering of final values of cumulative quantities carries the information on equivalence of formulations under different conditions and may serve as a starting point for necessary drug formulation redesign. Though perhaps not completely in line with other, more formalistic, data mining techniques the presented approach can serve as a starting point for further work on pharmacokinetic data analysis.

REFERENCES

Cellier, F. E. (1991). *Continuous System Modelling*. Springer-Verlag. New York.

Gibas, C., P. Jambeck and J. Fenton (2001). *Developing Bioinformatics Computer Skills*. O'Reilly & Associates.

Godfrey, K. (1983). *Compartmental Models and Their Application*. Academic Press. London.

Jelliffe, R., A. Schumitzky and M. Van Guilder (2000). Population pharmacokinetics/pharmacodynamics modeling: Parametric and nonparametric methods. *Therapeutic Drug Monitoring* **22**, 354–365.

Matko, D., R. Karba and B. Zupančič (1992). *Simulation and Modelling of Continuous Systems: A Case Study Approach*. Prentice Hall. New York.

Schumitzky, A. (1991). Nonparametric em algorithms for estimating prior distributions. *Applied Mathematics and Computation* **45**, 143–157.

Wagner, J. G. (1993). *Pharmacokinetics for the Pharmaceutical Scientist*. Technomic Publishing inc.. Lancaster.