Toxicology in Vitro 27 (2013) 1402–1409

Contents lists available at SciVerse ScienceDirect

Toxicology in Vitro

journal homepage: www.elsevier.com/locate/toxinvit

Data management in large-scale collaborative toxicity studies: How to file experimental data for automated statistical analysis

Sven Stanzel *, Marc Weimer, Annette Kopp-Schneider

Department of Biostatistics, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany

ARTICLE INFO

Article history:
Available online 20 December 2012

Keywords:
Concentration-response analysis
Data extraction
Data management
In vitro study
Large-scale toxicological project
REACH

ABSTRACT

High-throughput screening approaches are carried out for the toxicity assessment of a large number of chemical compounds. In such large-scale *in vitro* toxicity studies several hundred or thousand concentration-response experiments are conducted. The automated evaluation of concentration-response data using statistical analysis scripts saves time and yields more consistent results in comparison to data analysis performed by the use of menu-driven statistical software. Automated statistical analysis requires that concentration-response data are available in a standardised data format across all compounds. To obtain consistent data formats, a standardised data management workflow must be established, including guidelines for data storage, data handling and data extraction.

In this paper two procedures for data management within large-scale toxicological projects are proposed. Both procedures are based on Microsoft Excel files as the researcher's primary data format and use a computer programme to automate the handling of data files. The first procedure assumes that data collection has not yet started whereas the second procedure can be used when data files already exist. Successful implementation of the two approaches into the European project ACuteTox is illustrated.

1. Introduction

Safety assessment of chemical compounds is a key issue in predictive toxicology. A major initiative in this light was the formulation of Regulation EC 1907/2006 on Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), which became law in 2007 (Anon, 2007). It stipulates that by 2018 the toxicological properties of all existing and new substances sold in the EU in annual quantities of more than one ton (at least 30,000 compounds) must be determined. Conducting animal studies for toxicity testing of all these compounds would be time-consuming and would require a very large number of animals to be tested, and thus economically impractical (Kinsner-Ovaskainen et al., 2009a). Thus, REACH has necessitated the development, acceptance and use of alternative, non-animal test methods. Practical implementation of such *in vitro* test methods can reduce the number of animals used for toxicity testing (Anon, 2005; Kinsner-Ovaskainen et al., 2009a). In this situation, large-scale *in vitro* toxicity studies are carried out for high-throughput screening of a large number of compounds. Several hundred or thousand concentration-response experiments are conducted to examine a set of biologically relevant *in vitro* endpoints (e.g., differential gene expression, change in metabolite levels).

One of the main goals of large-scale *in vitro* toxicity studies is to carry out one common, combined, standardised and automated statistical analysis of the concentration-response data collected across all compounds. Important steps in this statistical data evaluation comprise (a) data preprocessing (e.g., data normalisation, background correction, outlier detection); (b) concentration-response curve fitting; (c) parameter and confidence interval estimation (e.g., estimation of an EC_{50} value with its respective 95% confidence interval); (d) parameter summary across all experiments conducted for the same compound × endpoint combination; and (e) statistical significance testing. Various choices exist for each of these steps of statistical concentration-response analysis. These choices must be evaluated systematically beforehand to determine the best option in each case.

In general, two types of software approaches must be differentiated with respect to statistical data analysis: a scripting approach and a menu-driven software approach.

Scripting approaches inevitably require that raw data are available in the same data format across all compounds. Specifically, the raw data of all experiments must be organised such that every row of the data matrix reflects a single concentration-response data pair. A scripting approach uses custom statistical analysis scripts, written, for example, in the software environments SAS (SAS Institute Inc., Cary, NC, USA) or R (<http://www.R-project.org>), to evaluate the data. Each analysis script consists of a sequence of instructions which are applied automatically to all data collected

* Corresponding author. Tel.: +49 6221 42 2135; fax: +49 6221 42 2397.
E-mail address: s.stanzel@dkfz.de (S. Stanzel).

0887-2333/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.tiv.2012.12.009

in the same data format. In the case of concentration-response analysis such instructions include, for example, data normalisation, concentration-response curve fitting and EC_{50} estimation. The following computer code shows an exemplary R script that can be used for concentration-response analysis:

```
library(drc)
data ← read.table(file = "./Mydata.csv")
model.data ←
as.data.frame(cbind(data$Response,data$Concentration))
model ← drm(model.data, fct = LL2.4())
ED(model, 50, level = 0.95)
plot(model, type = "all", log = "x")
```

Line 1 of this R code loads the R package "drc", which is typically used for dose–response curve modeling. Line 2 imports the data file "Mydata" into the R software, and assigns it to the R object "data". Every row of this data set contains a single concentration-response data pair. If the original data file is a Microsoft Excel file, this file must be converted to "comma separated values (csv)" file format before data import. Line 3 creates a data frame that contains only the concentration-response data of the original data file (column 1: response values; column 2: concentration levels). In line 4 the four-parameter log–logistic model is fitted to the concentration-response data to estimate the concentration-response curve. Alternative concentration-response models can be specified by replacing "LL2.4()" by other options. Line 5 estimates the EC_{50} and its 95% confidence interval. If, for example, the EC_{10} should be estimated instead of the EC_{50}, the number "50" in this command line simply has to be replaced by "10". Line 6 finally creates a plot of the concentration-response data, together with the fitted concentration-response curve. The option "x" for "log" causes the concentration levels to be plotted on log10 scale (on the x-axis of this graph), resulting in a more informative visualisation of the data.

Menu-driven statistical software, e.g., GraphPadPrism (GraphPad Software, La Jolla, CA, USA), on the other hand, allows the user to carry out statistical analyses by making selections from a list of options in a series of onscreen menus. The GraphPadPrism software is used frequently by toxicologists for the statistical evaluation of *in vitro* toxicology data, for example, to fit concentration-response curves or to estimate characteristic values such as the EC_{50}.

GraphPadPrism is a user-friendly statistical software system and thus well suited for data evaluation within toxicity studies in which only one or a few concentration-response experiments are conducted. For toxicologists, GraphPadPrism or alternative menu-driven statistical software environments are much easier to learn than statistical scripting languages such as R or SAS. However, in large-scale toxicological projects, typically several hundred or thousand concentration-response experiments must be evaluated jointly. In this situation, statistical scripting languages provide greater flexibility, are less time-consuming and yield more consistent statistical analysis results as compared to menu-driven statistical software. Specifically, in our experience, the need to modify the analysis strategy during the course of the statistical analysis occurs quite often. In this case, modifications can be carried out by adjusting the statistical analysis scripts. Executing the modified scripts will create an update of the analysis results in a short time, in smaller projects often within a few seconds. This is less error-prone and less time-consuming compared to the use of menu-driven statistical software, which requires re-opening every data file and manually changing the analysis by adjusting the sequence of selections from the available onscreen menus. This difference in terms of time, flexibility and consistency between the scripting approach and the menu-driven software approach is especially relevant for large-scale toxicity studies.

In vitro concentration-response experiments are often conducted by use of 24- or 96-well plate assays. Companies that are involved in the analysis of these types of assays usually have special scanners or high-throughput reading machines available in their laboratories. Hardware is connected to personal computers and, by application of user-friendly software, allows for conversion of the produced raw data into many different data formats and thus for automated data export into most of the commonly used data bases or spreadsheets, e.g., Microsoft Access or Microsoft Excel. This hardware and software architecture ensures that raw data can be saved in the same standardised format for all concentration-response experiments conducted for a single *in vitro* assay. Nevertheless, in large-scale toxicological projects, various types of *in vitro* assays are performed and raw data are often stored in heterogeneous data formats. This violates the essential requirement of one standardised data format across all compounds. Thus, automated concentration-response analysis using statistical analysis scripts cannot be applied directly to such large-scale *in vitro* data sets.

2. Recommendations for data management

In view of the size of the concentration-response data sets to be evaluated in large-scale toxicological projects, a standardised data management workflow is the only feasible approach to ensure consistent data usable for statistical analysis. The exact structure of this workflow depends on the type of the toxicological project. In general, the workflow should include guidelines for data storage, data handling and automated data extraction from various inconsistent Microsoft Excel data formats into one universal data format. To ensure an efficient data management workflow, it is recommended that standardised data formats are used for data storage. The important goal of standardising data storage, data base configuration, data handling and data extraction can be achieved by various strategies. We propose the data management strategy illustrated in Fig. 1.

2.1. Automated data template generation for data management at project start

Thorough data handling is central to all subsequent steps, including statistical analysis and interpretation of results, and thus for the success of a toxicological project. Therefore, it is advisable to determine the complete data management process, including recommendations for data storage, data base configuration, data handling and data extraction, as early as possible in the planning phase of the project, before data acquisition begins. It is also beneficial if a data manager or a central data management group is appointed before the project begins. Moreover, several alternative data management solutions should be considered and discussed in the planning phase of a toxicological project to determine the most suitable strategy.

It is further recommended that the data manager generates data templates for data entry during the course of the project. A separate data template should be created for every concentration-response experiment. Generation of data templates should be performed automatically and in close cooperation with the investigators producing the project data. Data templates must be created before data acquisition begins and must be characterised by a standardised data format. Specifically, in that part of a data template that shows the concentration-response data, one row should correspond to a single concentration-response data pair. Typically, Microsoft Excel files are used as data templates.