PhUSE 2010

Paper AD03

Exploratory Data Analysis in Pharmacogenomics using SAS JMP and Windows WF workflows

R.Huisden OCS Consulting, the Netherlands

Abstract

SAS JMP is appropriate for Exploratory Data Analysis with basic statistics and graphics. To use SAS JMP for the analysis of remote information, a workflow system has to automate both the SAS JMP tasks and the remote data capture tasks. As part of a pharmacogenomics analysis of Warfarin, the combination of SAS JMP and Windows WF workflows was studied. Based on a number of possible candidate genes, information on their ethnic distribution is captured from the Hapmap site, loaded into JMP and analyzed with a Chi-square test that can be used to restrict the amount of relevant information. The configuration is seen as a starting point for more complex workflows where Tracking and Persistence allow an uninterrupted workflow.

Introduction

The usual steps after finding a positive dose – response relation in pharmaceutical research are dose ranging / assessing dose linearity and defining a therapeutic window (MED (Minimum Effective Dose), MTD (Maximum Tolerated Dose)). However, with certain drugs there is a large inter-individual variability in the therapeutic dose. Algorithms can be developed to account for this inter-individual variability, based on demographic and clinical factors such as age, weight, height, gender, race, underlying diseases and medications.
For coumarine-like drugs like Warfarin, where the appropriate dose can vary by a factor of 10 among patients, these demographic and clinical factors only partially explain the variability in therapeutic dose. The response of the anti-clotting agents like Warfarin can be measured with the Prothrombine time, the clotting time of a blood sample. For comparison between laboratories the INR, the international normalized ratio of the measured time response is used. This Prothrombine time has to be closely monitored during treatment as the consequences of taking an incorrect dose can be dramatic.
Recently a number of genetic markers have been identified that are related to this dose – response variability. The enzymes expressed and the pathways of drug action and metabolism are identified. The clinical algorithms can now be extended with these pharmacogenetic factors. A number of these algorithms have been proposed. Some of these algorithms are based on uniform subpopulations. Others are developed for large international datasets and try to come to a universal dosing algorithm. An algorithm recently published is based on a large publicly available dataset concerning patients from different countries (1). The algorithm combines clinical and pharmacogenetic factors, but is not able to do without e.g. the factor race. Presumably more unknown factors are involved that are distributed differently in different populations, e.g. dietary habits. However, the algorithm that combines clinical and pharmacogenetic factors is significantly better than the clinical model. After splitting in model and validation groups and repeated modelling and testing, the best modelling technique appeared to be a simple Linear Regression. This best modelling technique was tested on the validation cohort for both the clinical and the pharmacogenetic models / factors. The two models are compared using the MAE (Mean Absolute Error) and McNemar's test for a significant difference.

Subgroup analysis

Given this state of knowledge the question remains whether to use one universal dosing algorithm or to use different dosing algorithms for different subpopulations. According to the article (1) specific models for different subgroups (race, clinical, pharma) were developed, but the general model adjusted for this subgroup performed better than these specific models.
A validation objective is then to develop models on the subpopulations defined to determine whether these dosing algorithms give a better dose prediction within the subpopulations than the universal model. The modelling and validation workflow could be extended with a number of additional steps, i.e. defining and modelling model subgroups and comparing subgroup overall model result and subgroup model result. A workflow containing these additional steps should preferably be automated.
Related to this is that the exploration for additional markers from candidate genes to extend the model, should take into account possible confounding population stratification during the association study to prevent spurious marker – dose response associations(2). Some insight in the heterogeneity of the subgroups before the selection of the candidate genes therefore seems appropriate. As part of this modelling and validation of the pharmacogenomics algorithms of Warfarin, the combination of SAS JMP and Windows Workflow Foundations was studied. Based on a number of possible candidate genes (3), information on their ethnic distribution was captured from the Hapmap site. With 30 genes on 12 chromosomes for 11 ethnic groups and approximately 5

SNP's (Single Nucleotide Polymorphism) per gene, there would be about 150 distributions to analyze. With an optional statistical test this number of distributions to analyze and to model can be restricted. The experience with this simple workflow can then be used to build more complex modelling or model comparison workflows.

JMP / WWF application

JMP is appropriate for EDA (Exploratory Data Analysis) and basic modelling. JMP script is available for programming repetitive tasks. In the above mentioned workflow, data retrieval from websites and JMP analysis are combined. So the workflow should automate both the FTP and the JMP steps. By using workflow programming one can visualize the steps and monitor and record progress with Tracking and Persistence and add Fault handling.
Windows WF is designed for (large) workflows like ASP shopping cart applications or document management applications. Here the build of a basic workflow for Hapmap frequencies (See Figure 1) was studied. This workflow consists of a UI with 6 - 7 steps. For this development phase workflow a State Machine workflow was chosen to enable monitoring of each step and to control the behaviour of the workflow once it is running.
The first State starts the workflow. In the second step a decision has to be made whether to 'FTP' the Hapmap files or to start JMP. This branching enables the use and direction of the workflow application on already downloaded Hapmap files and so to skip the FTP after a previous download. In the next step the actual FTP / Unzip takes place. A UI for file, subgroup and chromosome selection is presented. The DotNet System.Net.FTPWebRequest component is used for FTP. The open source Ionic component is used for unzipping the Hapmap files. A single chromosome for multiple subgroups downloads within minutes. In the first versions of the development phase workflow application no exception handler was necessary. In later versions error handling using available structures was incorporated. The workflow continues with the start of a JMP instance, comparable to the start of Excel of Word. After a UI to select the gene and the chromosome with start and end points, the first JMP script is started. It loads the (text) files and does a number of data steps, combining the subgroup files. The workflow application then presents a UI to test and restrict the Categorical Analysis and starts the second JMP script. The Pearson Chi-square test is used to select only markers with significantly differing relative frequencies. This step also renames, selects and stores the Graph and Test results. In the next step one has to decide on the next gene loop or closing JMP and stopping the workflow. This redirection of the workflow demonstrates a looping structure and jumps back to run the first JMP script. The final step closes the JMP windows and stops the workflow.

Lessons learned

Although the graphical approach seems appealing, there is an initial investment in time and effort required. At first sight this makes Windows WF workflows less attractive for small, at hoc workflows typically build and frequently changed during EDA. A DotNet programming experience is required which is not a Lego-like activity. In the above workflow there is additional JMP script programming experience required. The creation of workflows after an initial learning period will probably be faster and easier. Nevertheless it is a clear advantage to be able to monitor progress and to record the steps of the JMP analysis during the development phase. The results after each step can be inspected using a State Machine workflow. Initially there is no need for (elaborate) fault handling. In due course the workflow can be redesigned to a more uninterrupted (sequential) workflow.
Currently Windows WF is subject to version changes. In the DotNet Framework 4 WF there is not yet a State Machine workflow. There are however constructions for backward compatibility (4).

Future developments

The current application has an elementary workflow Tracking. A more elaborate Tracking Service and a Persistence Service are available and can be helpful to record the progress and to find errors (5). E.g. one could continue modelling after additional laboratory analysis (additional markers) on blood samples taken from the trial population is added. In that way the workflow can be extended with additional model factors and additional model testing. Modelling and model comparison are different from EDA. These are more or less routine activities while EDA is an investigative process. The extension of models with additional factors can be seen as a combination of these processes. The exploratory part can presumably consist of smaller ad hoc workflows.
These smaller workflows should then be compared with e.g. SAS Genomics workflows (6). These JMP workflows are however confined to the JMP environment and sequential in nature.
Another workflow system for genetic exploration is the Taverna system (7). This is an Open Source application that enables the combination of different website services. It uses Beanshell scripts to glue the steps together and requires some (Java) programming experience. With an Rshell workflow step statistical analysis can be introduced. Here R programming experience is required. This compares to the JMP environment where an extension to the SAS system is possible when more sophisticated statistical analysis is desired. The UI of this Taverna system is rather restricted compared to the rich possibilities of a Visual Studio Windows applications.

There are however many existing workflows that can serve as a basis for further exploration and in silico experiments. To include these rich sources in the current Windows WF application, one could make use of a command line version of this Taverna application. It could then be included as a custom activity in a Windows WF workflow.
One possible example of such an inclusion can be taken from the Warfarin dose modelling history (8). As an anti-clotting agent one has to model the High and the Low responders to Warfarin. The drug is also used as a rodenticide where drug resistance is a known phenomenon. This drug resistance led to the identification of the significance of the VKORC1 gene and its markers. The across species VKORC1 gene marker exploratory analysis is seen as a role model for future association studies.
Extensive fault handling was not programmed in this small application but should be added to larger workflows with multiple services to reconnect or exit the workflow if e.g. the connection fails. Existing (open source) workflows will require a validation of the services before use. Subdividing these workflows in smaller parts and incorporating these parts in custom Windows WF activities was mentioned above. With the available fault handling in these custom workflow activities debugging time will probably be reduced which in turn will enable the reuse of elaborate workflows.

Literature

1. The International Warfarin Pharmacogenetics Consortium. Estimation of the Warfarin Dose with Clinical and Pharmacogenetic Data. N Engl J Med 2009;360(8):753-764.
2. Turner SD, Crawford DC, Ritchie MD. Methods for optimizing statistical analyses in pharmacogenomics. Expert Rev Clin Pharmacol 2009;2(5):559-570.
3. Wadelius M, Pirmohamed M. Pharmacogenetics of warfarin: current status and future challenges. Pharmacogenomics J 2006;7(2):99-111.
4. Chappell D, The Workflow Way. Microsoft. 2009; 1-27.
5. Kitta, T. Professional Windows Workflow Foundation. Wiley Publishing, Inc. 2007.
6. SAS JMP Genomics, www.JMP.com
7. Taverna, http:\\taverna.sourceforge.net
8. Crawford DC, Ritchie MD, Rieder MJ. Identifying the genotype behind the phenotype: a role model found in VKORC1 and its association with warfarin dosing. Pharmacogenomics 2007;8(5):487-496.

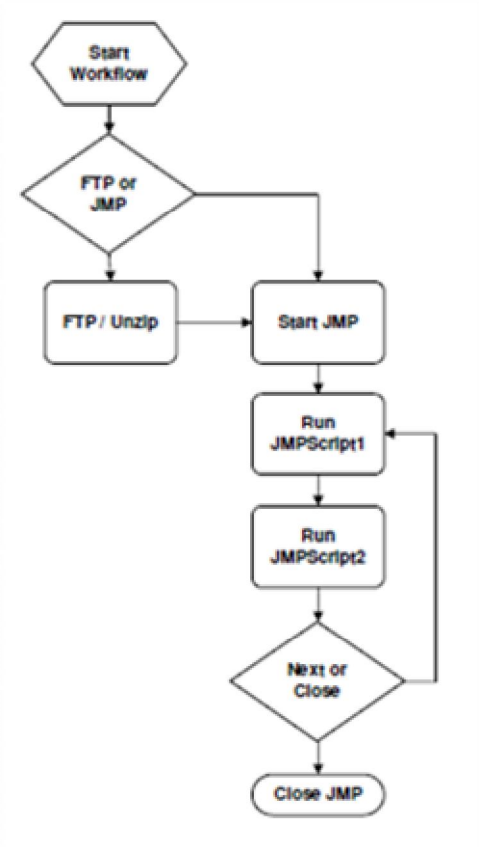
Contact Information

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

Robert Huisden
OCS Consulting
Ruwekampweg 2g
5222 AT 's-Hertogenbosch
The Netherlands

+31 (0)73 526 6000
sasquestions@ocs-consulting.com
www.ocs-consulting.com/nl

Figure 1


Start Workflow
FTP or JMP
FTP / Unzip
Start JMP
Run JMPScript1
Run JMPScript2
Next or Close
Close JMP