Automate Data Integration Processes for Pharmaceutical Data Warehouse

Sandy Lei, Johnson & Johnson Pharmaceutical Research and Development, L.L.C, Titusville, NJ
Kwang-Shi Shu, Johnson & Johnson Pharmaceutical Research and Development, L.L.C, Titusville, NJ

ABSTRACT

SAS® Data Integration (DI) Studio is a JAVA® GUI tool for processing inbound data into a data warehouse. For an environment that has established processes and rules for data manipulation, DI studio provides a visual mechanism to display extracted data attributes, transformation mappings, and loading processes. This is the ideal case and the preferred way to use the tool.

Applying the DI studio to pharmaceutical clinical data sometimes encounters the following issues: 1) the length, label, or type of a variable differs from study to study within the same compound, 2) a study may have many datasets and each dataset may contain many variables. Applying a manual process via DI Studio is time-consuming and error-prone. Often the same process may need to be repeated for different studies.

To resolve these issues, an automated Data Integration for data extraction, transformation and loading process has been developed using combination JAVA plug-ins, user-written codes and SAS macros to interface with SAS Metadata server.

INTRODUCTION

Clinical trials are usually conducted on a per study basis. Most of the studies require the consistency on the case report form (CRF) across all studies for the same compound, while others allow variations. Based on various needs, statistically or programmatically, different variables may be introduced into the analysis data sets for different studies. As a result, combining studies to form a uniform clinical data set for integrated summary of safety (ISS) or integrated summary of efficacy (ISE) is often a challenging task for either metadata analysis and/or e-submission in the pharmaceutical industry. This paper addresses the process of loading and creating uniform clinical data sets into a Data Warehouse. To ease the work, third-party tools, such as SAS Metadata Server and DI Studio, are used.

Business Intelligence is a set of software tools and applications that enable business users and analysts to interact with their company data in an easy, efficient and effective manner. SAS Business Intelligence includes the following: 1) a set of client applications designed for a specific type of business or analyst, 2) SAS server processes designed to provide specific type of services for the client applications, 3) a centralized metadata management facility.

As one of the client side tools, SAS Data Integration Studio enables a data warehouse specialist to do the following: 1) extract data from operational data stores, 2) transform data into a desired form, and 3) load data into a data warehouse.

On the server side, SAS implements SAS Open Metadata Architecture, which is a general-purpose metadata management facility that provides common metadata services to SAS applications. The architecture includes an application metadata model, a repository metadata model, an application programming interface, and a metadata server. There are several implementations of Application Program Interface (API), such as JAVA, C++, SAS, and all of them use XML as the transport language.
The pharmaceutical industry today is facing ever-increasing pressure to bring new drugs to the market that are safe as well as cost effective but with shorter timelines and less resources. It has become critical that a well-designed Clinical Data Warehouse be able to help scientists and researchers to tap into for potential innovative ideas and solutions to address these new business and regulatory challenges.

As the pharmaceutical arm of Johnson & Johnson, Johnson & Johnson Pharmaceutical Research and Development, L.L.C (J & J PRD) strives to manage and mitigate risks, manage intellectual assets, leverage intellectual assets to improve science and exploratory analysis, create robust submissions, increase the ability to do cross-study and cross-product analysis, and address the FDA’s new proactive risk assessment guidelines. All these business requirements lead us to work on building a clinical data warehouse that has a uniform repository for loading, organizing, combining and interrelating clinical information from multiple sources for use in risk management, has search and access capabilities, can perform various analyses, report and output capabilities and meet compliance.

GOAL OF THE PROCESS

Clinical information is organized around compounds. Each compound may have many studies at various phases, such as PK, Phase I, II, III or IV, across many years. The label, length and even type of a variable may differ among these studies. To create a uniform data repository, one has to ask what is the uniform data set? What kind of data model is the uniform data set going to use? For studies starting from 2006, J&JPRD has rolled out the Clinical Data Interchange (CDISC) Study Data Tabulation Model (SDTM) based data model; any study starting earlier than that is based on J&J PRD’s proprietary Analysis Data Model (ADM). Converting data from ADM model to SDTM model will be dealt later.

The ultimate goal of the process is to create harmonized analysis data sets (HADS) across studies for the same compound. The reality is that there are many variations among studies and once a study is completed and submitted to FDA, the data cannot be changed. Therefore we propose a two-stage approach: harmonize the study data based on the data model, and then create integrated data sets by resolving the remaining inconsistencies in a systematic approach.

THE AUTOMATED PROCESS

The overall automated process is presented in the following process map.
The automated process starts by creating and loading compound metadata information; that includes all data domains and attributes of all variables, such as comments, length, type, formats, code list, and origin, into the Metadata server. A compound metadata can be developed as expanding a selected existing standard model such as SDTM, ADaM, or any in-house models.

The following figures are snapshots of the DI studio that show the extended attributes of the compound metadata. Figure 1 contains the circled symbols for keys, notes, and the extended attributes of each variable. Figure 2 shows the comments from the data specifications via the DI studio note text field.

Figure 1: Compound Metadata

Figure 2: Algorithm defined in the compound metadata notes
The next step is to load study datasets into the Metadata server, extract the attributes from study datasets, and compare the attributes from study verse compound to identify any deviation from the data model. In order to capture the transformation process information, an excel workbook is created which contains various comparison reports such as missing compound-required variables, variables in study but not in compound, mismatched attributes, etc. Some of the reports, such as finding data sets containing null key variables, take a long time when running sequentially. They are ideally for parallel programming due to logical independency between data sets and can be programmed in parallel. The Excel workbook that captures these comparison reports will be used to document any necessary transformations to create harmonized data.

The following are the examples of three comparison reports created in this step: missing compound-required variables, in study and not in compound, and mismatch study vs. compound. Among them, missing compound-required variables, is shown in the Figure 3 below, will be used as example for transformation and loading process.

Two methodologies were explored in the SAS DI studio for data transformation: 1) interactive approach through the DI mapping process, 2) create user-written codes as background support for the transformations. We believe that using user-written code approach is more efficient when the in-bound databases require similar or identical transformations.

Figure 4 below shows that there are three variables renamed through interactive process via the DI studio. Besides renaming, Proc SQL statements can be used to perform data manipulation for the variables. The same process will be repeated for all other studies in the same compound since names of these variables were changed in the data model after studies have been completed.
Figure 5 and Figure 6 demonstrated the user-written code with job template approach. The actual codes are SAS data step statements with additional macro variables. It can be stored as a template and shared by studies. This approach reduces repetitive work by sharing the same codes. The job template contains information such as where the actual user-written codes located.
After applying transformation and loading, the exact same comparison reports will be created again to make sure there is no deviation from the data model. The transformation and loading study process can be iterated several times until all concerns are addressed and harmonized study data sets are created.

A compound level transformation spreadsheet document can be dynamically updated by adding any new transformations when the new study transformations are created.

With the reality that the clinical database usually contains large amount of variables, structures, and code-lists within a compound, and the time and resource constrains, we define a harmonized individual study database as one that meet all data model specifications with the following exceptions across studies: 1) code-decode inconsistency for CRF pre-printed variables, 2) length inconsistency for the variables, 3) type inconsistency for variables that are not relevant to analysis.

An automated data pooling tool will be developed to resolve the length and type inconsistencies, such as applying the largest length and convert numerical variables to character ones. Resolving code-decode inconsistency is complicated since it usually is based on the study statistical analysis plan. For example, same treatment group codes in different studies usually have different decodes. As the result, treatment group code-decode relationship has to be redefined when creating the ISS/ISE databases. In general, it is a case-by-case approach.

Pooling all harmonized study databases with the data-pooling tool described above, a uniform integrated compound database and its corresponding metadata will be created. The conjunction of the integrated database and the metadata database could be sufficient to address regulatory authorities’ requests and to support further clinical researches.

CONCLUSION

For many years, to have a uniform clinical data set across studies has been a big challenge for pharmaceutical companies. We proposed and implemented a two-stage automated approach to achieve this goal. These processes can be audited by examining the transformation workbook that is also used as the
basis for automating the transformation. In the future, we will extend the automated work by periodically checking the deviation of ongoing studies from the data model to ensure consistency across studies, which can reduce the work involved to resolve inconsistencies during the integration stage. With the single source of data, the consistent data format and the metadata integrity, searching and accessing across studies can be conducted systematically, and the time and the efforts spent will be reduced significantly.

REFERENCES

2. SAS Using SAS ETL Studio to Integrate Your Data, SAS Institute Inc., Cary, NC USA 2005
5. Introduction to the SAS Business Intelligence Client Tools, SAS Institute Inc., Cary, NC USA 2004
7. CDISC Study Data Tabulation Model: SDTM V1.1 & SDTM IG V3.1.1, 2005

ACKNOWLEDGEMENTS

We would like to thank our managers for their support, encouragement as well as continuously challenging us to enhance the automated clinical data warehousing processes.

TRADEMARKS

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. JAVA is a trademark of SUN Microsystems Inc.

CONTACT INFORMATION

Sandy Lei
Johnson & Johnson PRD
1125 Trenton-Harbourton Road
Titusville, NJ 08560
Work Phone: (609) 730-3320
Email: slei@prdus.jnj.com

Kwang-Shi Shu
Johnson & Johnson PRD
1125 Trenton-Harbourton Road
Titusville, NJ 08560
Work Phone: (609) 730-4633
Email: kshu@prdus.jnj.com