PhUSE 2008

Paper CD02

CDISC Implementation Strategies: Exploit your data assets while still giving the FDA what they want

Dave Smith, SAS Institute, Marlow, UK

ABSTRACT

With the unparalleled opportunity for standardisation offered by CDISC and the introduction of the FDA JANUS initiative many pharmaceutical companies are re-evaluating their clinical data processing strategies.
Companies are trying to drive improvements in the efficiency and quality of their process, and are also looking for opportunities to better exploit their data assets.
In this paper we will describe the main options for structuring clinical data during the implementation of CDISC standards, examine the strengths and weaknesses of each one, and recommend a strategy that will give the greatest potential benefit.

INTRODUCTION

The production of submission data for regulatory review in SDTM structures is rapidly becoming the norm, with very few companies attempting to hold on to their legacy standards beyond the immediate term. This means that they will inevitably need to make (or have made) changes in their data processes. The planned introduction of the FDA's JANUS warehouse has made many companies consider creating their own clinical data warehouse structures, but the data modelling implications need careful consideration. This paper will examine the data structure options and the implications of the adoption of each one.

WHY CHANGE? POTENTIAL BENEFITS THAT THE NEW WORLD WILL BRING

In this section we will examine the benefits of introducing both CDISC standards and a common clinical data warehouse.

PROCESS IMPROVEMENT

One of the main arguments in favour of standardisation is that it should increase the throughput of studies, with more re-use of code, and far easier pooling of data.
There will be an initial overhead while companies get used to the implementation of the CDISC standards and may lose some of the benefits they have accrued through the use of their in-house standards in the short term. The advantages of a cross-industry standard should be rapidly apparent as systems and procedures are implemented around this standard.

DATA MINING

Data Mining is a process of exploring and modelling large amounts of data for business benefit. Through tools such as SAS® Enterprise Miner™ companies can quickly identify previously unidentified patterns and connections in their data leading to insights about their business. Examples might include the identification of line extensions or the propensity for patients from particular investigators to withdraw early from studies.

For data mining to be effective it is generally best to create a large single data table that contains as much information as possible; this is not the same as the JANUS Warehouse data structure conceived and developed by Norman Stockbridge at the FDA, according to the presentation given by Wayne Kubick at the NIH BECON/BISTIC Symposium, entitled Data Management: The Clinical Research and Regulatory Perspective. There are, however, benefits to starting with a structure such as SDTM before preparing data for mining, as the combination of data across trials becomes much easier and the loading of data into a data mining mart can be automated.

IN-STREAM DATA REVIEW

Opportunities that standards bring to allow continuous review of data and the potential impact on data quality
With data in a standard format it becomes much easier to provide clinical experts with standard reports and visualisations to allow them to review data in-stream, with the resulting improvements in patient safety.
The one cautionary note to make is that the provision of in-stream data gives a much higher profile to data quality; although decisions based upon uncleaned data may be not be different from the cleaned version in most cases the risks are still identifiable and therefore data quality should be carefully managed, ideally in real-time or as close to real-time as is practical.

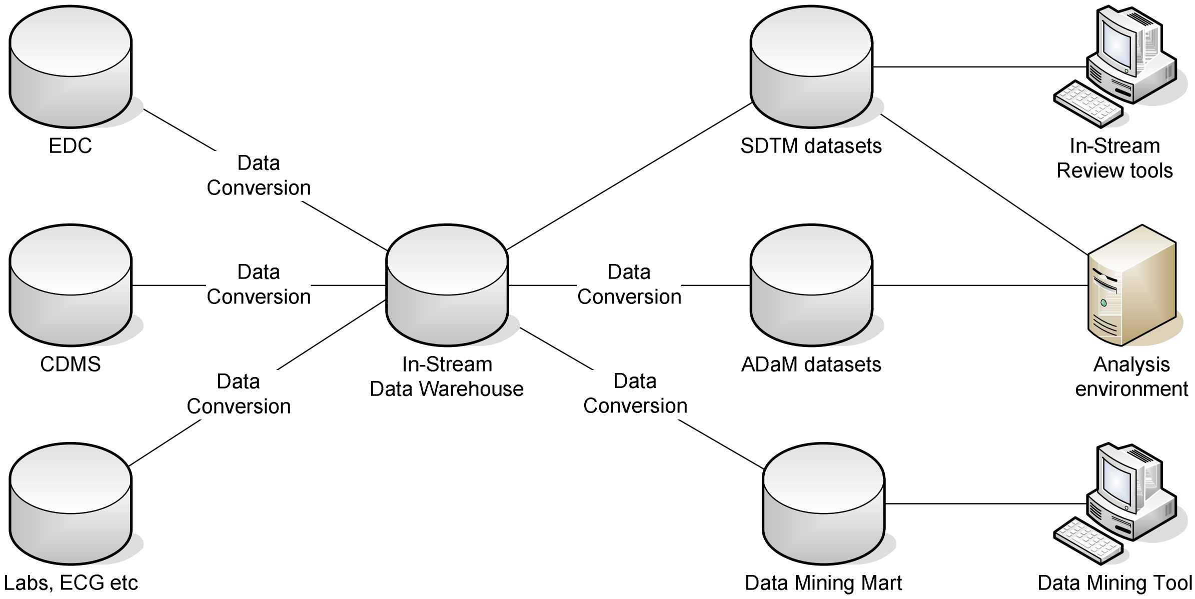
SIGNAL DETECTION

Signal detection algorithms are used to identify drug safety problems as soon as possible by identifying adverse events that occur more often than would be expected through chance. There are four commonly used algorithms (Proportional Reporting Ratios, Relative Odds Ratios, Bayesian Confidence Propagation Neural Networks and Multi-Gamma Poisson Shrinkers) and each has its proponents. All of these require that the data are combined to allow the algorithms to have sufficient power, and standard structures in the clinical databases make this much easier. It is likely that the study data would be combined with data from spontaneous adverse event reporting systems and so this is less likely to come from a JANUS warehouse.

DESCRIPTION OF THE MAIN OPTIONS

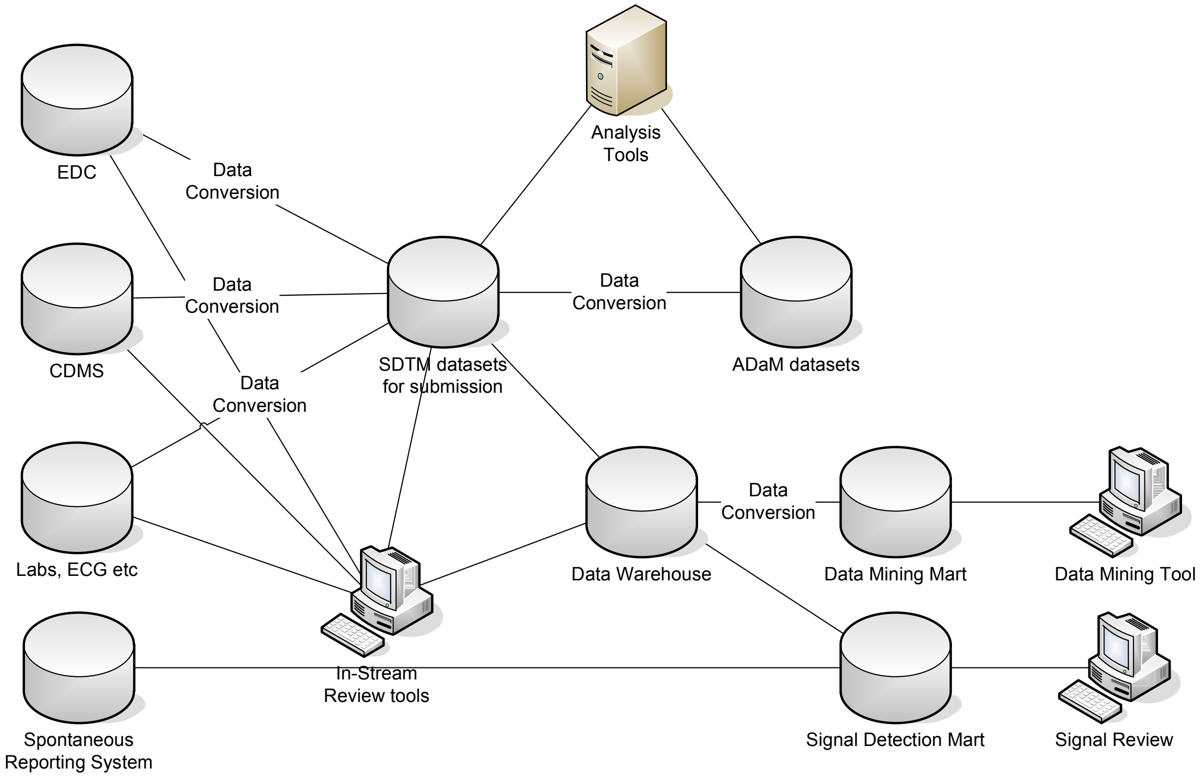
In this section we will describe the main options for structuring data through the clinical development process and on to exploitation structures.

OPTION 1 – THE IN-STREAM DATA WAREHOUSE



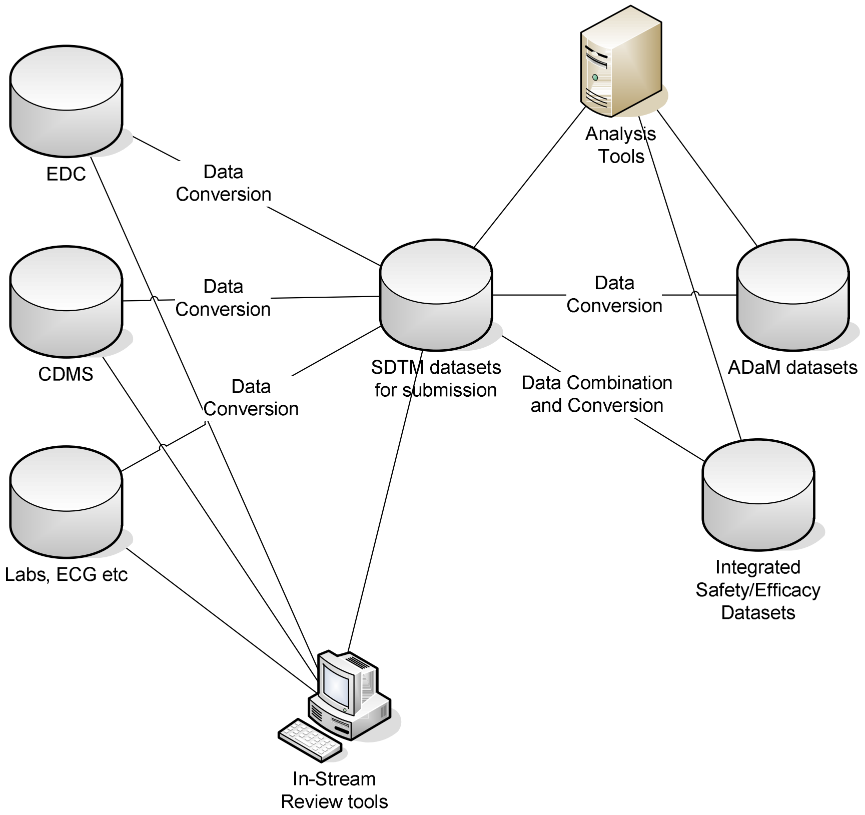
In Option 1, all incoming clinical data is converted to a single structure immediately and all further processes are fed from there. It is typical that the Spontaneous Reporting Systems are not loaded into this warehouse, as this is outside the JANUS design. The in-stream data warehouse is already in an SDTM structure and therefore only needs extraction to produce submission-ready datasets for single studies and integrated summaries. Conversion is still required for the creation of ADaM datasets. In-stream review tools are typically those that require SDTM structured data. A data mining mart would require data conversion as the JANUS data structures are not suitable for data mining, which tends to require a large single table structure.

OPTION 2 – THE OFF-LINE DATA WAREHOUSE



In Option 2, separate process streams are maintained for each study and a separate structure is built for combined analysis across all trials. The combined data warehouse would not need to comply with the JANUS structure (although it could do so) and there would be further marts for data mining across projects and for adverse event signal detection. In-stream review tools are typically applied earlier in the process, and are therefore usually not of the kind that requires SDTM structured data.

OPTION 3 – THE FULLY FEDERATED MODEL



In Option 3, separate process streams are maintained for each study and data is only combined for specific purposes, such as integrated safety and efficacy summaries. In-stream data review tools are again of the kind that can be applied without SDTM conversion. This option represents the traditional non-warehoused approach, and in many pharmaceutical organisations represents the current status.

SWOT ANALYSIS

In this section we will look at the strengths, weaknesses, opportunities and threats from each option

OPTION 1 – STRENGTHS

The main strength of option 1 is that the data has been standardised early in the process and therefore downstream of the JANUS warehouse the opportunities for automation are considerable. To an extent this approach also reflects the planned efforts of the FDA with its JANUS program, allowing a similar insight. Integrated summaries of safety and efficacy are made easier, as is the exploration of the data across compounds and therapeutic areas for cross-business insight.

OPTION 1 – WEAKNESSES

The main weakness of option 1 is a loss of traceability. Regulators have traditionally been able to trace the provenance of each study's data through its lifecycle with clarity and one-to-one matches. When the data are moved into a relatively data complex model this transparency is much harder to express to a reviewer. The other issue is that the warehouse is being used for direct clinical decision making, and therefore needs to be validated. Maintaining the validated state has a time overhead, but also reduces the flexibility of the solution to react to new opportunities. Signal detection algorithms would only be applied to the data in the JANUS warehouse, and would not be combined with data with the spontaneous reporting system.

OPTION 1 – OPPORTUNITIES

The main opportunity with option 1 is the ability to produce an overview of the entire research program, both current and historical, on a like-for-like basis. This could be invaluable in business process management and forecasting.

OPTION 1 – THREATS

There is some latency in loading the data into the JANUS warehouse format, giving a delay before in-stream data can be reviewed, which would certainly be unsuitable for some early phase studies (e.g. dose finding phase 1 studies). This latency could lead to a safety issue being missed that an early review of lab data might have identified.

OPTION 2 – STRENGTHS

The main strength of option 2 is traceability. It is clear for every analysis that is performed which dataset is contributing to which analysis, and with appropriate tools (such as SAS Clinical Data Integration) the provenance of each column of data can be clearly identified. Tools for in-stream data review can operate from non-SDTM structured data, so a wider choice of tools is available and the latency associated with conversion to SDTM is not present.

OPTION 2 – WEAKNESSES

The main weakness of option 2 is that the production of summaries across trials would not be as easy as in a properly constructed data model.

OPTION 2 – OPPORTUNITIES

The main opportunity with option 2 is the ability to provide high performance analytical marts to derive maximum intelligence from a company's clinical data. The data warehouse would not be constrained to a JANUS structure (although this could be chosen) which allows a potentially higher performing data model to be chosen. If no direct clinical decision making is required from the data warehouse then this would not necessarily need to be a validated data warehouse, which would reduce the time and cost to create this layer.

OPTION 2 – THREATS

If the summary warehouse is not built to a JANUS structure there may be a view of the data constructed by the FDA that is not exactly replicated; this is unlikely to occur as similarly constructed summaries are available in the warehouse layer.

OPTION 3 – STRENGTHS

The main strength of option 3 is that it reflects current practice in most organisations and therefore reflects the option to do nothing, with no costs associated. Traceability can still be maintained if an appropriate tool is used to manage the metadata through the creation of analysis datasets.

OPTION 3 – WEAKNESSES

The main weakness of option 3 is that it doesn't really exploit the clinical data assets of the company. There is also less opportunity to improve process efficiency through the introduction of data standards.

OPTION 3 – OPPORTUNITIES

There are no incremental opportunities associated with option 3.

OPTION 3 – THREATS

The main threat associated with option 3 is that a regulator will identify an issue within the submitted clinical data that has not been identified by the company. There is also the threat of a missed opportunity to better exploit the clinical data which might either save costs or generate incremental revenue.

PRACTICALITIES

CONSIDERATIONS OVER CDISC VERSIONS; SDTM +/-

One of the issues that companies are faced with is that CDISC is an evolving standard. Although SDTM 3.1.1 has been a relatively stable option for some time now, 3.1.2 is on the near horizon and will provide incremental benefit, particularly with the introduction of the PK domains. There are many other standards that are evolving, particularly ADaM, and define.xml (CRT-DDS) which will need to be supported, plus eventually SEND, LAB and CDASH.

The management of standards over the life of a long term clinical project is therefore a key issue for companies. The main considerations should be ensuring that it is clear which standard has been employed at each stage of the development process and management of the conversion between versions of standards where appropriate. Ideally a library of standards needs to be maintained in a flexible metadata environment. The metadata standard used should also be associated with each item in the analysis datasets, ensuring that there is adequate clarity throughout the process. This is particularly important for service providers such as Contract Research Organisations, which may have to manage many sponsor company standards as well as different CDISC standards.

These standards are being introduced into an environment where companies have already made great benefits from introducing in-house standards, and are understandably reluctant to lose those benefits. Some companies have chosen to standardise upon a variant of the SDTM standard (either SDTM-, where some variables are not included, or SDTM+, where additional company standard variables are present). This allows the company to maintain some of their historical standardisation benefits while making the eventual generation of "pure" SDTM for submission relatively straightforward. The disadvantage of this approach is that collaboration with development partners is less easy and standard tools that require pure SDTM will not necessarily operate correctly.

The implications for the decision between options 1 to 3 are that unless a "pure" SDTM approach is taken, option 1 is unlikely to be viable, and may be unsuitable for companies with large existing investments in standardisation. Option 1 could lead to an overhead in terms of conversion to a later version of a particular standard towards the end of a clinical study. Option 3 is likely to make the most of existing standardisation and might therefore lead to inertia against change to options 1 or 2.

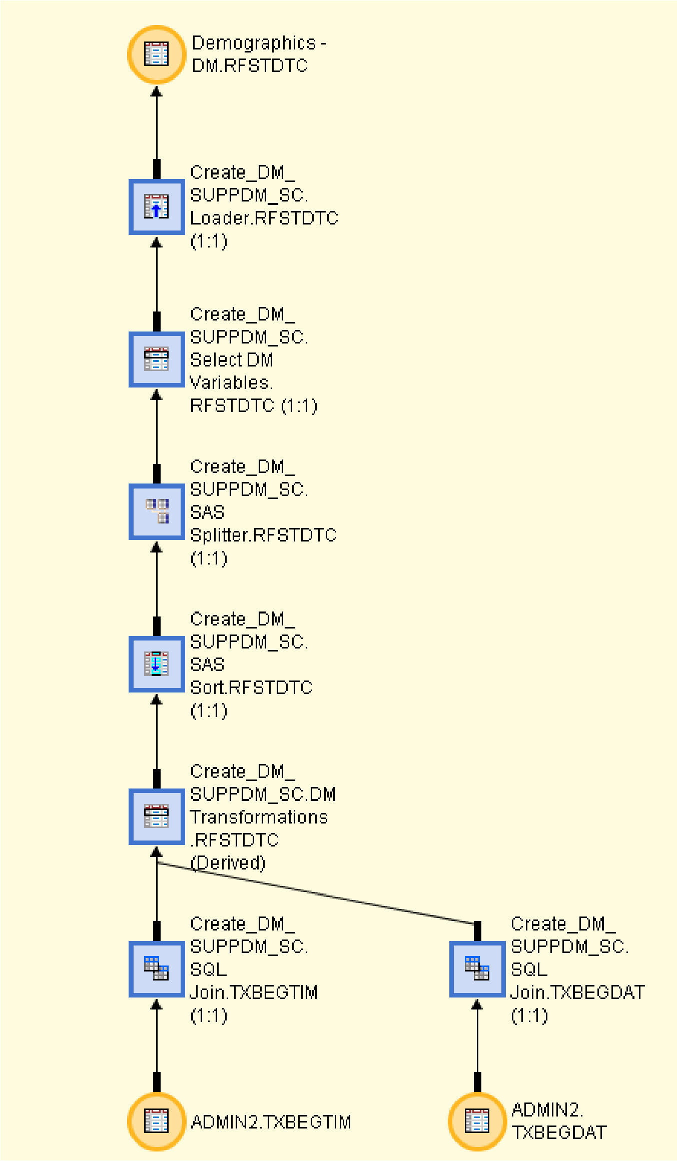
VALIDATION

The principle of validating each component required for clinical decision making is key in this environment. In option 1 the in-stream data warehouse is part of the flow of validated data and therefore must be validated to the highest standards. In option 2 the data warehouse is outside the main flow of validated clinical data and may not need to be validated – this will depend whether or not the data are used for clinical decision making. There may be many uses of the data, such as modelling the likelihood of withdrawals from a study that will lead to business decisions being made (e.g. number and profile of investigators being recruited) as opposed to clinical decisions. This is likely to lead to a lower level of validation and control being required. In option 3 summary datasets would be validated as in previous good practice.

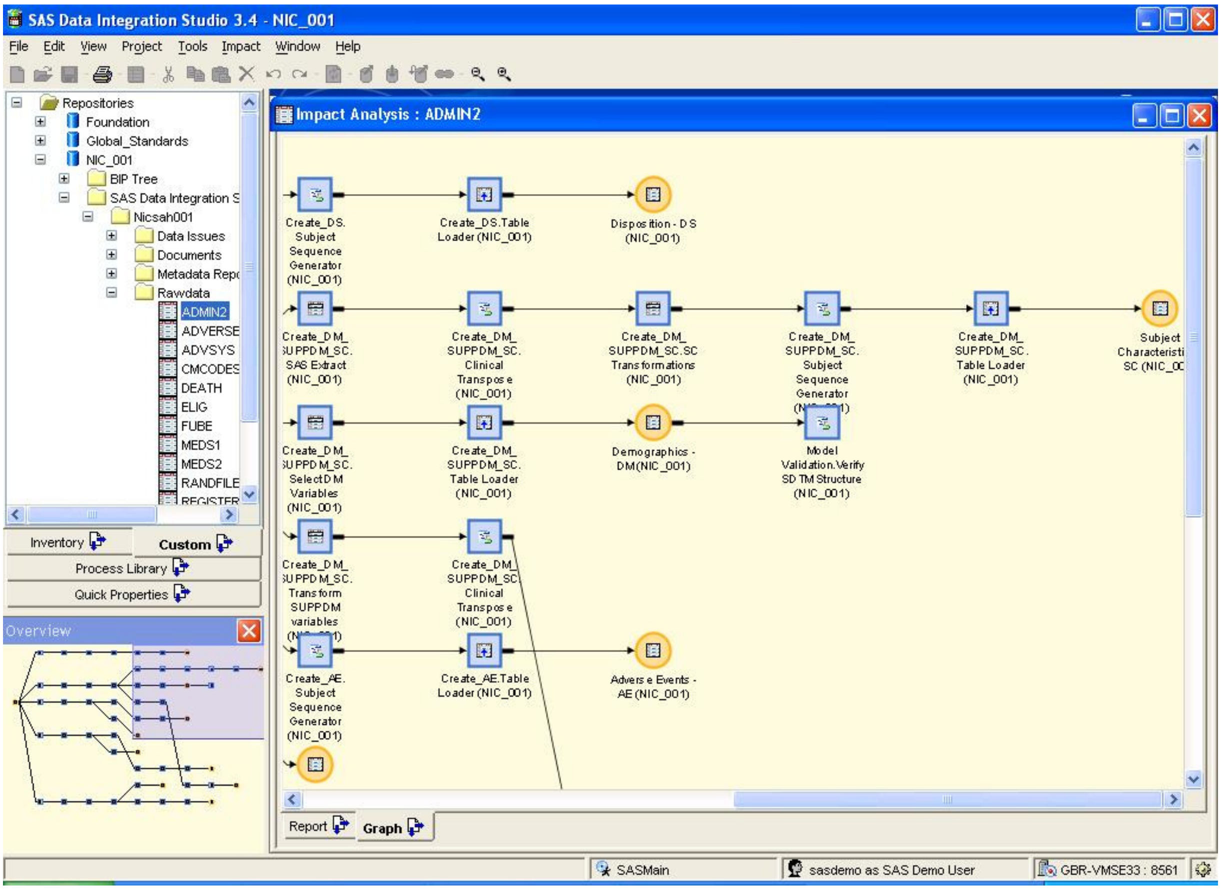
TRACEABILITY

This paper has mentioned traceability several times; the importance of maintaining a clear path between source and target data throughout each stage of the clinical reporting process is clear; it ensures that the regulators are able to fully understand the analysis presented at submission, it makes questions from regulators much easier to answer, and it allows the management of any changes to the input data structures and output analysis datasets with a clear understanding of the resulting re-validation requirements.

In options 2 and 3 the traceability at the dataset level is relatively straightforward, although it would be greatly enhanced by the use of a suitable metadata management tool such as SAS Clinical Data Integration (pictured below). In option 1 the conversion to a JANUS data model at the first stage breaks that traceability immediately.


Demographics - DM.RFSTDTC
Create_DM_ SUPPDM_SC. Loader.RFSTDTC (1:1)
Create_DM_ SUPPDM_SC. Select DM Variables. RFSTDTC (1:1)
Create_DM_ SUPPDM_SC. SAS Splitter.RFSTDTC (1:1)
Create_DM_ SUPPDM_SC. SAS Sort.RFSTDTC (1:1)
Create_DM_ SUPPDM_SC.DM Transformations .RFSTDTC (Derived)
Create_DM_ SUPPDM_SC. SQL Join.TXBEGTIM (1:1)
Create_DM_ SUPPDM_SC. SQL Join.TXBEGDAT (1:1)
ADMIN2.TXBEGTIM
ADMIN2. TXBEGDAT

SAS Clinical Data Integration: Column Level Impact Analysis. In this figure the Clinical Data Integration Studio solution is tracing the creation of the RFSTDTC column in the DM domain from the TXBEGTIM and TXBEGDAT columns from the ADMIN2 dataset. The step in which the columns are combined is indicated by the label (Derived) as opposed to (1:1).


SAS Data Integration Studio 3.4 - NIC_001
File Edit View Project Tools Impact Window Help
Repositories
Foundation
Global_Standards
NIC_001
BIP Tree
SAS Data Integration S
Nicsah001
Data Issues
Documents
Metadata Repo
Rawdata
ADMIN2
ADVERSE
ADVSYS
CMCODES
DEATH
ELIG
FUBE
MEDS1
MEDS2
RANDFILE
REGISTER
Inventory
Custom
Process Library
Quick Properties
Overview
Impact Analysis : ADMIN2
Create_DS. Subject Sequence Generator (NIC_001)
Create_DS.Table Loader (NIC_001)
Disposition - DS (NIC_001)
Create_DM_ SUPPDM_SC. SAS Extract (NIC_001)
Create_DM_ SUPPDM_SC. Clinical Transpose (NIC_001)
Create_DM_ SUPPDM_SC.SC Transformations (NIC_001)
Create_DM_ SUPPDM_SC. Subject Sequence Generator (NIC_001)
Create_DM_ SUPPDM_SC. Table Loader (NIC_001)
Subject Characteristi SC (NIC_00
Create_DM_ SUPPDM_SC. SelectDM Variables (NIC_001)
Create_DM_ SUPPDM_SC. Table Loader (NIC_001)
Demographics - DM(NIC_001)
Model Validation.Verify SDTM Structure (NIC_001)
Create_DM_ SUPPDM_SC. Transform SUPPDM variables (NIC_001)
Create_DM_ SUPPDM_SC. Clinical Transpose (NIC_001)
Create_AE. Subject Sequence Generator (NIC_001)
Create_AE.Table Loader (NIC_001)
Adverse Events - AE (NIC_001)
Report
Graph
SASMain
sasdemo as SAS Demo User
GBR-VMSE33 : 8561

SAS Clinical Data Integration: Table Level Impact Analysis. In this figure all the uses of a source table are displayed. The overview on the bottom left allows the detailed view to be navigated.

WHAT DOES THE FDA REALLY WANT?

The ADaM Pilot results presented by Cathy Barrows at the CDISC European Interchange in Copenhagen in 2008 were very clear that the FDA want traceability between source and target data, particularly as far as the ADaM structures are concerned. The FDA reviewers were also very keen to understand where each variable is created and how; this was particularly brought out in questions over the creation of the LOCF imputation, and this was one of the main issues raised with the initial version of the submission, along with issues over the structure of the define.xml. The main conclusion slide included the line “Maintaining transparency is key”, and this is perhaps the main message as far as CDISC implementation is concerned.

CONCLUSION

Based upon the analysis above the least attractive option for many organisations is to select option 3, which is essentially to do nothing. This leaves organisations open to risks that regulators will identify shortcomings in their data or issues with their products that better exploitation of their data assets might have prevented.

Option 1 has theoretical attractions and may allow the greatest degree of standardisation early in the process; however the lack of traceability through the in-stream data warehouse layer makes this an unattractive option. The lack of flexibility makes this particularly unsuitable for clinical research organisations.

Option 2 seems to offer the best overall strategy, allowing the most exploitation of combined data whilst still maintaining the transparency needed by regulators.

REFERENCES

The CDISC SDTM / ADaM Pilot Project - Results and Learnings. Cathy Barrows, Associate Director, Statistics & Programming, GlaxoSmithKline. Presented at the CDISC European Interchange 2008.

Data Management: The Clinical Research and Regulatory Perspective. Wayne Kubick, Vice-President, Lincoln Technologies. Presented at the NIH BECON/BISTIC Symposium

CONTACT INFORMATION

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

Dave Smith
SAS Institute
Wittington House
Henley Road
Medmenham
Marlow
Bucks
SL7 2EB

Work Phone: 01628 404379
Fax: 01628 490550
Email: david.smith@suk.sas.com
Web: http://www.sas.com